Enhancing Data-Driven Decision Support with Multi-Perspective Solutions

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Yaqiong Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Gediminas Adomavicius (Advisor)

August, 2020

Acknowledgements

There are many people that have earned my gratitude for their contribution through my Ph.D study.

First and foremost, I would like to express my most sincere gratitude to my advisor, Dr. Gediminas Adomavicius. He is a great mentor. Through my Ph.D study he has provided me with continuous support and guidance that helped me to grow as a researcher. All the time working with him is full of critical thinking, enlightening conversations and inspiration. This thesis is not possible without him. His deep passion, scientific attitude and impressive erudition in research have, and always will, constantly inspire me to pursue my academic career.

My sincere gratitude also goes to the rest of my committee members, Dr. Alok Gupta, Dr. Joseph Konstan and Dr. Jason Chan, for their constructive comments. The feedback and insights which they generously and patiently provided make it possible for this thesis to be as good as it is today.

I would like to thank all faculty members and doctoral students in the Department of Information and Decision Sciences, who have graciously helped me during the past few years. All the kind and inspiring words I heard from them greatly encouraged me to go further in this journey.

Last but not least, I would like to thank my families. Without their unconditional support, encouragement and love, I could not have completed this thesis and my degree.

Abstract

As digital systems become ubiquitous, providing all-around support for decision makers has become a significant part of contemporary information systems. To this end, numerous data-driven analytics techniques have been widely adopted by various platforms to facilitate decision making in a wide variety of application domains, e.g., product choice, employee recruitment, and medical diagnosis. The appropriate application of various data-driven methodologies for decision support in complex real-world contexts is crucial to gain benefits and to avoid unexpected consequences and, thus, the ability take into account multiple perspectives for better decision support represents an important challenge. In order to provide insights into this question, this thesis focuses on investigating some of the problems existing in decision support applications and attempts to provide various solutions and empirical evidence of the effectiveness of these solutions. Specifically, my thesis proposes to provide more nuanced decision support in different application domains by balancing different aspects of decision support models or by providing complementary sources of information for decision makers, e.g., balancing *accuracy* and *long-tailness* to address *popularity bias* in recommender systems; using *individual prediction reliability* to complement outcome prediction to support decision making in highly risk-sensitive domains like medical diagnosis or financial markets; providing complementary channels to fulfill online consumption decision support in the retailing industry. Solutions and findings provided by my thesis advance the understanding of decision support problems in multifaceted contexts, and have practical implications for information systems that adopt data-driven methods.

Contents

Acknowledgements **i**

Abstract **ii**

List of Tables **vi**

List of Figures **ix**

1 Introduction **1**

1.1 Research Background and Motivation 1

1.2 Overview of Three Essays 6

2 Essay 1: Efficient and Flexible Recommendation Using Cosine Patterns **11**

2.1 Introduction and Motivation 11

2.2 Background and Related Work 17

2.2.1 Recommender Systems and the Long-Tail Challenge 17

2.2.2 Association Analysis and Its Application to Recommender Systems 19

2.2.3 Cosine Patterns and Their Properties 23

2.3 Cosine Pattern-Based Recommendation 26

2.3.1 Cosine Patterns for Recommendation 26
2.3.2 Basic Recommendation Scheme 28
2.3.3 Cosine-Pattern Tree Traversal Approach 32
2.3.4 Parallelizing Cosine Pattern-based Recommendation 37
2.4 Experimental Results 41
2.4.1 Experimental Setup 41
2.4.2 Recommendation Accuracy Performance 43
2.4.3 Long-Tail Recommendation Performance 45
2.4.4 Additional Experiments 47
2.4.5 Scalability Demo: CORE for Hashtag Recommendation 54
2.5 Conclusions 56

3 Essay 2: Improving Reliability Estimation for Individual Numeric Predictions 60

3.1 Introduction and Motivation 60
3.2 Related Work 65
3.3 Machine Learning Approach to Individual Prediction Reliability Estimation 72
3.3.1 Individual Prediction Reliability Indicator: General Overview . 72
3.3.2 Estimating and Evaluating the Proposed Reliability Indicator: ML-Based Framework 76
3.4 Experiments 79
3.4.1 Experimental Setup 79
3.4.2 Performance Comparison Based on Correlation 82
3.4.3 Performance Comparison Based on RMSE 85
3.5 Conclusions 87

4 Essay 3: The Role of Physical Stores in the Digital Age **93**

4.1 Introduction . . . 93

4.2 Related Work and Theoretical Background . . . 98

4.2.1 Multi-channel Retailing . . . 98

4.2.2 Roles of Physical Stores in Chinese Markets . . . 101

4.3 Empirical Context and Data Description . . . 107

4.3.1 Empirical Context . . . 107

4.3.2 Sales Tract Definition . . . 107

4.3.3 Descriptive Statistics . . . 109

4.3.4 Natural Experiment Involving Offline Store Expansion . . . 110

4.4 Econometric Model and Identification Strategies . . . 112

4.4.1 Difference-in-Differences . . . 112

4.4.2 Propensity Score Matching . . . 114

4.4.3 Individual and Product level Analysis . . . 116

4.5 Empirical Results . . . 119

4.5.1 Impact of Stores on Online Sales . . . 119

4.5.2 Robustness Checks . . . 123

4.5.3 Product level Analysis . . . 126

4.6 Mechanisms and Heterogeneous Effects . . . 126

4.6.1 Conspicuous and Experiential Roles of Stores . . . 126

4.6.2 Impact in Different Territories . . . 131

4.6.3 Impact on Customer Types . . . 132

4.7 Implications and Future Research . . . 135

5 Concluding Remarks **140**

References **143**

List of Tables

2.1 Impact of null-consumptions on pattern/rule evaluation measures. 22

2.2 Association rules and cosine patterns discovered from the MovieLens dataset. . 29

2.3 Descriptive statistics of top-150 rules/patterns. 29

2.4 Summary statistics of datasets. 41

2.5 Average popularity of recommended items. 46

2.6 Percentage of niche items recommended. 46

2.7 Descriptions of adjusted data sets. 49

2.8 Straightforward vs. CP-tree based recommendation efficiency. 56

3.1 Common Notations for Describing Reliability Estimation Methods 71

3.2 Description of Baseline Reliability Estimation Methods 71

3.3 Overview of Data Sets Used in Computational Experiments 81

3.4 Predictive Accuracy (RMSE) of Different Outcome Prediction Models.
(Average performance based on 30 runs; best performance on each data set is shown in bold.) 82

3.5 Comparison of Reliability Estimation Performance (Correlation Coefficient)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better) 82

3.6 Performance of Machine-Learning-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.) . . . 83
3.7 Performance of Heuristic-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.) . . . 90
3.8 Reliability Estimation Performance of Machine-Learning-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.) . . . 91
3.9 Comparison of Reliability Estimation Performance (RMSE)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better) . . . 92
3.10 Reliability Estimation Performance of Heuristic-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.) . . . 92
4.1 Descriptive Statistics . . . 111
4.2 Illustrative Summary Statistics-Mean(SD) . . . 116
4.3 Impact of Store Opening on Total Sales (Volume) . . . 120
4.4 Robustness Checks. . . . 125
4.5 Effect of Product Showcase on Online Sales (Volume) . . . 127
4.6 Effect of Offline Showcase on Online Sales (Volume) . . . 127
4.7 Impact of Stores on Online Sales (Split by Product Involvement) . . . 130
4.8 Effect of Stores on Online Sales (Fine-grained Categorization) . . . 131
4.9 Effect of Stores Opened in Different Types of Locations . . . 133
4.10 Impact of Store Opening on Online Sales (Split by Customer Group) . . . 134

4.11 Customer Maintenance in Different Types of Locations 136

List of Figures

2.1 An example for CP-tree construction and traversal. 33

2.2 Illustration of parallelization based on CP-tree partitioning. 38

2.3 CORE vs. baseline algorithms on accuracy (test-set ratio = 30%). 44

2.4 CORE vs. baseline algorithms on accuracy under different K. 44

2.5 CORE vs. baseline algorithms on accuracy (test-set ratio = 10%). 45

2.6 CORE vs. baseline algorithms on accuracy (test-set ratio = 50%). 45

2.7 CORE vs. baseline algorithms on accuracy and long-tail recommendation. . . 47

2.8 CORE vs. WRMF on different distributions of item popularity. 48

2.9 Two dimensional (*Precision-AvgPop*) performance comparison. 50

2.10 Two dimensional (*Precision-NicheRatio*) performance comparison. 51

2.11 Performance of hybrid CORE and WRMF approaches on MovieLens data. . . 54

3.1 Synthetic Data Example for Prediction Reliability Issue.
(X/Y axis: input/outcome variable. Dots: data points. Solid line: predictive model based on linear regression.) 62

3.2 IPR Representation Based on 95% Confidence / Prediction Interval
(X/Y axis: input/outcome variable. Dots: data points. Solid line: model based on different techniques.) 68

3.3 Pointwise Prediction Error Estimation of Linear Regression Model from Fig. 3.1 (X axis: input variable. Y axis: absolute prediction error. Black dots: actual abs. prediction error. Blue dots: estimated abs. prediction error.) 75
3.4 Two-Stage Machine-Learning-Based Framework for Reliability Estimation . . . 79
4.1 Clusters Discovered in Nanjing. 109
4.2 Distribution of Consumer Base of Different Clusters in Nanjing. 110
4.3 New Store Opening Over Time. 112
4.4 Coefficients of the Weekly Interactions Before and After Offline Expansion. . . 121
4.5 Discovered Product Clusters. 129

Chapter 1

Introduction

1.1 Research Background and Motivation

The past few decades have witnessed the evolution of numerous data-driven analytics techniques and their applications in business, economics, and various other fields. These techniques have been shown to greatly improve decision quality and create value for firms and customers. For example, companies adopting "data-driven decision making" have been shown to achieve productivity gains that were 5-6% higher than other factors could explain (Brynjolfsson et al. 2011a). In the healthcare industry, associations and patterns discovered from data can help healthcare providers and other stakeholders develop more informed diagnoses and treatment, contributing to higher quality care at lower costs, i.e., better overall outcomes (Raghupathi and Raghupathi 2014). On Spotify, more than 40% of the users continuously listen to personalized playlists generated by the platform (Buskirk 2016). The emergence of myriad applications also increases the need and reliance on data-driven technologies to address decision-making problems in more complex contexts, such as product recommendation, medical diagnosis, personal finance, retailer channel management, where multiple perspectives are often necessary

and challenges exist due to inadequate or noisy data, and multiple, often conflicting, objectives.

Many real-world decision making problems are multifaceted in nature. Decision making tools and methods which fail to take into account multiple important factors can lead to unintended consequences or suboptimality. One example is the *popularity bias* of most existing recommendation algorithms (Hosanagar et al. 2013) which emphasize popular items (i.e., items with more ratings) over "long tail" items. Part of the reason for such side effects is that most automated decision making models are trained on single or limited measures (e.g., accuracy), leaving out other factors like diversity and, thus, causing less favorable outcome to already disadvantaged group (Barocas and Selbst 2016). Another example is MHC[1] binding prediction in biological and pharmaceutical research where experimental validation should proceed based on not only the prediction, but also the *confidence* in the prediction in order to confirm more good binders with fewer costly experiments (Briesemeister et al. 2012). This thesis is motivated by these phenomena and attempts to address some of these issues by balancing several goals in predictive model development or integrating perspectives from complementary sources of information to facilitate decision making. Developing these multi-perspective solutions is critical to provide intelligent decision support as data-driven techniques penetrate all aspects of our daily lives. The thesis uses a multi-essay format and consists of three essays addressing three multi-perspective decision support problems.

Recommender systems are widely used predictive analytics applications that help users to make consumption decisions among huge amount of choices. They have been creating great value for online business, as personalized recommendations can have significant impact on users' purchase decisions (Pathak et al. 2010). Nevertheless, some recent studies have shown that classic recommendation techniques, i.e., collaborative

[1]Major Histocompatibility Complex

filtering-based methods, have popularity bias which directs users' attention to top percentile products (Hosanagar et al. 2013, Abdollahpouri et al. 2019). This is due partly to the fact that most existing methods are accuracy-oriented, downplaying other aspects like recommendation *diversity* or *long-tailness*. However, there is an increasing understanding that long tail recommendations are also valuable, since they better satisfy heterogeneous consumer needs (Brynjolfsson et al. 2011b) and, thus, can lead to more engagement (Baumol and Ide 1956, Kekre and Srinivasan 1990) and drive more consumption. For platforms, niche items or products can have higher profit margin. For example, Netflix could lower its movie licensing costs by recommending users to watch niche movies (Goldstein and Goldstein 2006). Sales of the tremendous number of niche products offered on platforms could also add up to a large share of total sales (Anderson 2006, Hart 2007). An intuitive way to address the suboptimality caused by popularity bias is to take a multi-perspective view, i.e., evaluate recommendations beyond their accuracy, and develop new methods to improve recommendation quality on multiple dimensions. Thus, the first essay of this thesis explores *popularity bias* caused by single-dimensional (i.e., accuracy-oriented) recommendation evaluation in most machine learning-based recommender systems. Specifically, it utilizes additional metrics beyond accuracy to inspect this issue and proposes a cosine pattern-based recommendation technique to tackle the problem.

The complexity of using data-driven decision support also comes from the fact that the automated predictions based on which decisions are made are often imperfect due to noisy, limited data or simplified mathematical or statistical assumptions (Kleinberg et al. 2018). The imperfection of predictive models poses challenges in decision-making facilitation, especially in highly risk-sensitive domains like pharmaceutical research, medical diagnosis, or financial markets (Briesemeister et al. 2012, Tomassetti et al. 2016, Huang

et al. 2018). So far, most predictive models focus primarily on providing *individual* prediction outcomes, while the quality of predictions is commonly evaluated using *aggregate* prediction accuracy metrics. While the overall performance is an important aspect of model evaluation, a more nuanced understanding of the model, e.g., when and how much it works better or worse, can be critical in many real-world applications. Individual prediction reliability, e.g., prediction confidence (Wonnacott 1987), constitutes as an additional and valuable aspect for model evaluation and application. Providing such extra reliability information to complement individual predictions can be a practical way to intelligently apply predictive results. For example, to help a specific customer to make a final decision about investing in a stock portfolio, it would be important to know not only the actual prediction of its 3-year return but also the estimated reliability of such a prediction. Thus, the second essay of this thesis examines the issue of providing individual prediction *reliability* information in addition to outcome predictions for a decision maker. A general machine learning-based framework is proposed to estimate individual prediction reliability, i.e., interpretable measurement of prediction confidence for decision support.

In addition to deploying (online) information systems based on data-driven methods for automated decision making, offering complementary sources for multi-perspective information collection is also imperative to provide effective decision support. This is especially true under the context of product choices where purchase decisions often rely on the inspection of multi-dimensional product features (Bell et al. 2017). While the penetration of digital commerce has shown the incomparable advantages of digital channel in many aspects, for example, breaking geographical barrier in communicating with potential customers at much lower costs, providing richer product information in the forms of product reviews and user ratings to support decision making, offering broader selection of products to improve shopping satisfaction (Brynjolfsson et al. 2013),

etc., given the complexity of purchase decision making, traditional brick-and-mortar stores are also valuable in reducing consumers uncertainties about products by allowing them to touch and feel merchandise. Despite the conspicuous benefits of both outlets, there have been inconsistent strategies adopted by different retailers in practice, i.e., while some are reducing their physical presence (Egan 2016, Gustafson and Reagan 2016), others are increasingly investing in offline assets (Addady 2016, Cadell 2017). Similarly, existing academic studies have also yielded different insights on the economic impact of offline channels. Some studies provide evidence of cannibalization among different channels (Choi and Bell 2011, Shriver and Bollinger 2015) while others find the complementary effect of multiple channels during purchases (Ansari et al. 2008, Balasubramanian et al. 2005). These contradictory findings call for further investigation of the underlying mechanism of the multi-channel effect and deeper understanding of potential strategies to provide comprehensive support that consumers need to make purchase decisions. Thus, the third essay of this thesis investigates the complementarity among different channels in consumer purchase conversion and provides managerial insights on firms channel management for effective consumer decision making.

It is not unreasonable to state that the significance of data-driven methodologies and information systems based on them will continue to dominate the landscape across many application domains. As information systems encompass greater scope of application relevance and responsibility, it is necessary for the platforms or firms to develop decision support methodologies and theories to practically deal with complex and multifaceted nature of decision support. This thesis takes a multi-perspective view toward several important decision support problems encountered in various application domains and attempts to enhance existing data-driven techniques for better decision support.

In summary, all three essays are related to the overarching theme of the thesis. The first two essays advocate multi-dimensional evaluation of data-driven predictive modeling solutions for the purpose of their intelligent applications and make methodological contributions to advance predictive-analytics-based decision support. The third essay makes a theoretical contribution and provides empirical evidence of the complementary role of multiple information sources in facilitating consumer decision making. Each essay is written in a self-contained manner. The first two essays are in collaboration with my adviser Dr. Gediminas Adomavicius and the third essay is in collaboration with Dr. Jason Chan. To acknowledge their contributions, I use "we" throughout the thesis. An overview of three individual essays is provided in the next section.

1.2 Overview of Three Essays

The first essay aims at addressing the long-standing *popularity bias* of most existing collaborative filtering-based techniques in the recommender systems field by designing a new method that balances recommendation quality from multiple aspects.

As mentioned earlier, one of the main properties of most widely-used recommendation methods, i.e., popularity bias, leads to more recommendations and, thus, exposure of popular items, creating rich-get-richer effects for these popular products and vice versa for unpopular ones. It is well understood that recommending items that are already well-known and bestselling (and that the users are likely to be aware of) arguably might be less valuable than the ability to find something truly relevant and personalized from the "long tail" of the item popularity distribution (Baumol and Ide 1956, Kekre and Srinivasan 1990). For platforms like Netflix, recommending users to rent niche movies could also lower the cost of licensing blockbusters (Goldstein and Goldstein 2006). To alleviate the popularity bias caused by accuracy-oriented evaluation of recommendations, other performance aspects of recommender systems, e.g., diversity

and novelty, have been studied in prior work (Levy and Bosteels 2010, Adomavicius and Kwon 2012, Castells et al. 2015). While improvements in recommendation diversity and novelty may be associated with better long-tail recommendation as well, it's not guaranteed to be so. Long-tail recommendation could also be achieved by improving rating estimation specifically for niche items (Park and Tuzhilin 2008, Zhang and Hurley 2009, Niemann and Wolpers 2013). However, such methods tend to achieve this at the substantial expense of overall recommendation accuracy or by requiring a much richer set of features and extra preprocessing. Meanwhile, as the user population and item catalog in the system grow over time, scalability of long-tail recommendation is also an important concern. Also, the ability to easily adjust the degree of popularity of recommended items is another highly practical and important characteristic that is not present in existing long-tail recommendation approaches.

To address some of these limitations, in this essay we propose the CORE (COsine pattern-based REcommendation) approach to long-tail recommendation. CORE is a pattern-based method, which finds associations, represented by *cosine patterns*, among different items (especially niche items) and then utilizes the discovered associations for the purpose of item recommendation. The scalability of CORE is facilitated by a specialized data structure as well as parallel computing schemes, which is crucially important for real-time recommendation capabilities in large-scale applications. Through comprehensive experimental comparison between the proposed approach and various baseline recommendation algorithms, we demonstrate that our method provides practical benefits in accuracy, flexibility, scalability, in addition to the superior long-tail recommendation performance.

My second essay focuses on providing individual prediction reliability (confidence) as complementary information to outcome predictions for the purposes of comprehensive interpretation and prudent application of predictive models.

The development and improvement of machine learning in the past few decades have spurred a myriad of applications where critical human decisions (e.g., employee recruitment, defendant bail, patient diagnosis, etc.) start relying on model predictions. However, those predictions can be imperfect due to limited or noisy data, simplified computational or probabilistic reasoning. Thus in addition to individual predictions, providing reliability estimation for each single prediction adds a finer-grained evaluation even for presumably well-trained predictive models and gives practitioners more confidence in making decisions based on predictions.

While previous studies have proposed various approaches, e.g., confidence intervals (Wonnacott and Wonnacott 1990), prediction intervals (Khosravi et al. 2010, Shrestha and Solomatine 2006), Gaussian process-based predictive distribution (Rasmussen 2003), and heuristic-based methods (Bosnić and Kononenko 2008a), for reliability estimation, many of them are missing some desirable features, which limits their application and evaluation. First, traditional confidence interval-based and predictive distribution-based reliability representations are highly oriented toward regression models and rely on distributional assumptions about prediction errors. Thus, these approaches are not directly applicable to estimate prediction reliability of many other predictive techniques, such as neural networks, and can fail when homoscedasticity assumption is violated in real-world settings. Second, a number of traditional approaches have the problem of low interpretability. For example, when using "density" of data points as an indicator for prediction reliability (Clark 2009, Bosnić and Kononenko 2008a), we can only interpret that a higher density (e.g., input space with more learning examples) is an indicator of higher prediction reliability, yet a precise magnitude of expected prediction error cannot be inferred, which is often crucial for interpretability and decision-making purposes.

To alleviate those concerns, in this essay, we propose to estimate the reliability of individual predictions of any given numerical outcome prediction model by using machine

learning techniques. We convert the reliability estimation problem to a numeric prediction problem by proposing to use absolute prediction error as an indicator of prediction reliability due to its merits of higher interpretability and easier evaluation. Based on this idea, individual prediction reliability could then be estimated by using machine learning techniques that attempt to directly learn prediction errors obtained from the given model. A complete general-purpose framework is also designed for implementing and testing the proposed approach. Experimental results show that machine learning methods are advantageous in identifying the relationships between prediction reliability and complex input features, and thus can significantly improve prediction reliability estimation as compared to a number of heuristic approaches used in prior work, especially in more complex predictive scenarios.

The third essay examines the complementary role played by retailers' offline channels in consumers' online purchase decision making.

With increasing ecommerce penetration, it is believed that consumers are spending more of their shopping time online and away from physical stores. Recently, there has been a heated debate on whether the traditional offline retailers are still relevant. On one hand, a lot of major retailers in the United States are closing their stores. As an example, Macy's, the iconic retailer that used to be a main stay in America's malls, closed over 15% of its 650 stores (Egan 2016). Nevertheless, some rising retailers like TJ-max and fashion companies like Zara are still expanding their businesses by opening more brick-and-mortar stores (Valladares 2017). More interestingly, online first retailers like Amazon and Alibaba which have no legacy of offline presence before are now keen on investing in offline stores (Cadell 2017). Investigations into cross-channel effect, and more specifically effect of launching brick-and-mortar stores on consumers' online purchases, have also yielded different insights. Substitution can occur as newly opened physical stores compete with online stores for sales, leading to cannibalization (Choi and

Bell 2011), while some other studies find that offline stores can play a complimentary role to online channels as well (Ansari et al. 2008, Avery et al. 2012, Wang and Goldfarb 2017). Thus whether brick-and-mortar stores still matter, and if so, how, with the thriving of online purchase facilities, has become an intriguing question to think about.

In this essay, we attempt to shed light on this question using a quasi-experiment, taking place through a nationwide retailer that expanded its physical presence over time. Through a difference-in-differences-in-differences model applied at the product-level, we provide more direct evidence on the positive effect of the physical store on sales in the online channel. In particular, we explicitly show that the showcase of products in stores, i.e., the conspicuous benefits of physical stores, has a positive relationship with products' online purchases. We also find that online purchases for products showcased in physical stores increase for both high and low involvement products. This set of results suggest that two mechanisms are likely driving the online sales of products displayed in physical stores. An experiential effect helps consumers converge their purchase for high involvement products by providing them additional sources of product information that is not available online, and an exposure effect increases purchases by generating greater top-of-the-mind awareness of low-involvement products. This study dispels concerns on the diminishing role of physical stores and demonstrates how those stores can fulfill a crucial decision support purpose in the digital age.

Chapter 2

Essay 1: Efficient and Flexible Recommendation Using Cosine Patterns

2.1 Introduction and Motivation

The development and adoption of web technologies and services gave rise to so-called "infinite inventory" digital commerce and content delivery platforms (Anderson 2006), such as Amazon, Netflix, and Spotify, which are able to provide many more items than their brick-and-mortar counterparts, and also fueled the development of various business analytics applications for targeting customers more effectively and for providing better services to them. Recommender systems play a unique role in online business, as high-quality personalized recommendations have been shown to have huge impact on users' purchase and consumption decisions (Pathak et al. 2010). For example, 60% of Netflix rentals and 35% of Amazon's sales are attributed to their recommendation systems (Hosanagar et al. 2013); also, more than 40% of users on Spotify continuously

listen to personalized playlists generated by the platform (Buskirk 2016). Over the years, a wide variety of methods, typically based on collaborative filtering techniques, have been proposed to improve the relevance of recommended items (Adomavicius and Tuzhilin 2005, Ricci et al. 2015). Although these methods have ween widely applied, such recommender systems also have been shown to have *popularity bias*, which refers to the tendency of the systems to recommend disproportionately more popular items, i.e., items with more ratings or purchases, to users (Fleder and Hosanagar 2009, Abdollahpouri et al. 2019).

Recommender systems manifest popularity bias due to different reasons. To begin with, personalized recommendations are generated from historical data on users' previous consumptions, which is inherently skewed towards popular items. Collaborative filtering methods generate recommendations to a given user by analyzing consumption of similar users; thus, popular items consumed by a large number of other users naturally get more exposure as part of the modeling process (Fleder and Hosanagar 2009). Also, recommender systems have been focusing largely on predictive-accuracy-oriented performance metrics, and recommending popular items can lead to higher accuracy since they are more likely to be consumed by users. The popularity bias of traditional recommender systems focuses users' demand mainly to top-percentile products or services, and the profit from popular items is definitely important for online businesses. However, there is a growing understanding that recommending comparatively popular items is not necessarily always the most advantageous strategy. For instance, recommending items that are already well-known and bestselling (and that the users are much more likely to be already aware of) arguably might be less valuable than the ability to find something truly relevant and personalized from the "long tail" of the item popularity distribution. Also, when a platform wants to push its "back catalog", adaptation of collaborative filtering algorithms is required to identify relevant but less popular products (Lee and

Hosanagar 2019). Yet, due to the fact that less data is available on these products, user preferences for them are harder to predict and, thus, accurately recommending the long-tail (niche) items remains an important challenge. This constitutes the focus of our study.

The value of long-tail recommendations has been increasingly recognized in different sectors. This can be illustrated from both demand- and supply-sides. On the demand-side, the long-tail (niche) titles can increase consumer surplus and drive consumption (Brynjolfsson et al. 2006). In particular, consumers are known to have a propensity to seek variety over time (Farquhar and Rao 1976, Pessemier 1978), while recommending niche items can encourage users to try products that are outside their awareness (Brynjolfsson et al. 2011b) and, thus, better satisfy customers' heterogeneous needs. In the long run, more user engagement on the platform can be stimulated to increase overall demand (Baumol and Ide 1956, Kekre and Srinivasan 1990). On the supply side, it has been suggested that niche products can be more profitable for companies, e.g., niche movies cost a fraction of blockbusters to make and market (Anderson 2006). Furthermore, for platforms like Netflix, recommending users to view more niche movies and fewer blockbusters is advantageous, as blockbusters tend to have much higher licensing costs (Goldstein and Goldstein 2006). In addition, unlike offline shops, online platforms are not constrained by limited shelf space and can typically carry tremendous amount of niche products. Sales of these niche items could also grow to take up a large share of total sales (Anderson 2006, Hart 2007). As an example, an average Barnes Noble store carries 130,000 titles; yet more than half of Amazon's book sales come from outside its top 130,000 titles (Kresh 2007). Furthermore, generating niche recommendations in online marketplace like Amazon can incentivize the producers of niche products to stay on the platform instead of being crowded out by producers of popular products (Abdollahpouri et al. 2019).

However, many platforms have not taken advantage of niche products' potential due to the fact that discovering long-tail items is not easy (Tan et al. 2017), and popularity bias (which is perpetuated in most existing online recommender systems) is known to exacerbate this problem. As a result, ability to provide effective long-tail recommendations is critical for helping platforms to maintain flexible business strategies.

To alleviate the popularity bias caused by accuracy-oriented evaluation of recommendations, other performance aspects of recommender systems, e.g., diversity and novelty, have been studied in prior work (Castells et al. 2015). Novelty and diversity of recommendations could be enhanced by re-ranking the initial recommendation list (Ziegler et al. 2005, Zhang 2009, Levy and Bosteels 2010, Adomavicius and Kwon 2012) or by optimizing the ranking process using a combined objective of accuracy and diversity (Yin et al. 2012, Shi 2013, Hurley 2013, Su et al. 2013). While improvements in recommendation diversity and novelty may be associated with better long-tail recommendation as well, it's not guaranteed to be so. For example, some studies find that recommender systems can improve individual diversity while still reducing aggregate diversity (Fleder and Hosanagar 2009). Long tail recommendation could also be achieved by improving rating estimation specifically for niche items (Park and Tuzhilin 2008, Zhang and Hurley 2009, Niemann and Wolpers 2013). However, such methods tend to achieve this at the substantial expense of overall recommendation accuracy or by requiring a much richer set of features and extra preprocessing. Meanwhile, as the user population and item catalog in the system grow over time, scalability of long-tail recommendation is also an important concern. Also, the ability to easily adjust (parameterize) the degree of popularity of recommended items is another highly practical and important characteristic that is not present in a number of long-tail recommendation approaches.

Pattern-based, especially association rule-based, recommendation algorithms have attracted some attention since the early days of recommender systems research (Mobasher et al. 2001, Lin et al. 2002). One key reason is the interpretability of their recommendations ("people who bought these items also bought..."). Typically, pattern-based algorithms first build a knowledge base containing patterns (e.g., association rules or itemsets) of typically co-occurring items and then recommend items to users based on this knowledge. Different target items are ranked by some interestingness measures, such as *support* or *confidence*, of the discovered patterns (Zaíane 2002, Kazienko 2009, Ghoshal and Sarkar 2014). Many websites have embedded rule-based approach into their commercial recommender systems, e.g., YouTube used association rules to recommend relevant videos to their users (Davidson et al. 2010). However, the traditional framework for association rule discovery, which is based on confidence and support metrics, has certain limitations that can lead to less accurate recommendations (especially with skewed data distributions). First, confidence as an interestingness measure might fail to filter some spurious associations among items (Brin et al. 1997), and recommendations based on spurious rules would not be useful. Second, support-based (i.e., frequency-based) pruning strategy of pattern mining might be problematic on data with inherently skewed support distributions, i.e., where correlations among niche items are key to long-tail recommendation yet are harder to be discovered due to low rates of niche item occurrence. As a result, traditional pattern-based recommender systems built on the support-confidence framework might not be perfectly suitable for long-tail recommendation.

To address some of these limitations, in this study we propose the CORE (COsine pattern-based REcommendation) approach to long-tail recommendation. CORE is a pattern-based method, which finds associations, represented by *cosine patterns*, among different items (especially niche items) and then utilizes the discovered associations

for the purpose of item recommendation. The scalability of CORE is facilitated by a specialized data structure as well as parallel computing schemes, which is crucially important for real-time recommendation capabilities in large-scale applications.

Specifically, this paper makes the following contributions. First, we overview several pattern-based recommendation methods and discuss their limitations. In contrast, the proposed approach uses an extra measure, i.e., *cosine*, to select more relevant patterns for recommendation. Second, we develop CORE, a new recommendation method based on *cosine patterns*, which is able to substantially improve both long-tail and accuracy performance of pattern-based approaches by limiting the discovery of spurious patterns. The proposed method also supports convenient parameterization of the popularity of recommended items; that is, the dual configuration of *support* and *cosine* thresholds adds more flexibility in generating recommendations of different popularity (or long tail) levels in order to achieve various recommendation goals. Third, to facilitate the scalability of the proposed approach, we further use a special CP-tree (*Cosine-Pattern tree*) structure to accelerate the recommendation process. The CP-tree is easily built and stored, and can be partitioned to parallelize the recommendation process. Based on the proposed data structure, we also design a parallel computing and load balancing framework for recommendation generation to guarantee the scalability of CORE. Fourth, comprehensive experiments on three benchmark data sets illustrate the advantages of CORE to existing pattern-based methods in terms of accuracy and long-tail recommendation as well as the advantages of CORE with respect to a number of classical, widely used collaborative-filtering-based approaches. And, a separate experiment on a larger-scale dataset further emphasizes the applicability and scalability of the proposed approach for practical, real-time recommendation tasks.

2.2 Background and Related Work

2.2.1 Recommender Systems and the Long-Tail Challenge

With the boom of the consumer-oriented content delivery and retail platforms since early 2000s, recommender systems have been progressively developed for various application domains, including movies, music, books, etc., to confront information overload and facilitate personalized information retrieval (Ricci et al. 2015). Over the years, *accuracy* of recommender systems has been the major lens through which their performance is evaluated and compared. For example, Netflix held an open competition (with $1M prize for the winner) for the most accurate recommendation algorithm to predict user ratings for movies (Koren et al. 2009). However, research studies increasingly point out that focusing on accuracy alone in recommender systems can result in sales diversity reduction (Fleder and Hosanagar 2007), since classical collaborative-filtering-based methods tend to disproportionately recommend popular items. According to Hart (2007), taking advantage of the long-tail market is one of the keys towards increasing profits on e-commerce platforms. Thus, in this study, we focus on the *long-tail* perspective of recommender systems with the goal of developing a recommendation method that can achieve better long-tail performance while still being highly competitive in terms of recommendation accuracy.

Previous studies have attempted to address the long-tail challenge in different ways. One general stream of research has focused on taking a broader perspective on recommender systems evaluation, rather than focusing just on accuracy, which gave rise to a number of additional recommendation performance dimensions, such as diversity, novelty, etc. (Castells et al. 2015). In particular, different metrics related to recommendation diversity and novelty have been proposed, e.g., average individual (intra-list) diversity (Zhang and Hurley 2008, Ziegler et al. 2005), aggregate diversity (Fleder and

Hosanagar 2009), serendipity (Murakami et al. 2007, Zhou et al. 2010), unexpectedness (Adamopoulos and Tuzhilin 2015), as well as recommendation algorithms for improving these metrics (Zhang 2009, Adomavicius and Kwon 2012, Adamopoulos and Tuzhilin 2015). Of course, diversity and novelty metrics represent an indirect way to affect long-tail recommendation performance; i.e., although these metrics have some correlation with long-tail recommendation performance (as they typically affect the distribution of recommended items), it is not guaranteed to be the case. For example, some studies show that aggregate diversity of recommendations can decrease even when individual diversity increases (Fleder and Hosanagar 2009, Jannach et al. 2013).

Therefore, some other studies have investigated ways to tackle the long-tail recommendation problem more directly. These studies can be further categorized by whether the long-tail-aware computations are applied as a pre-processing step vs. embedded more directly into the recommendation process. As an example of a pre-processing approach, Park and Tuzhilin (2008) proposed to first split items into head- and tail-groups based on their rating frequency, cluster tail items into different clusters, and then predict user ratings within each cluster. In their follow up study (Park 2013), different head-tail grouping strategies to enhance long-tail recommendation are compared, and the results show that accuracy of the long-tail item recommendations indeed increases through clustering. A similar idea was explored by Zhang and Hurley (2009), where items in each active user's profile are clustered first, and recommendations are generated based on each cluster instead of complete user profiles. In contrast to pre-processing approaches, several other studies propose to embed information about item popularity into recommendation generation process, e.g., discounting popular items when learning to rank. For instance, in order to promote long-tail items in recommendations, probability for a user to consume a certain item (i.e., recommendation score) based on the whole user-item interaction graph could be discounted by the rating frequency of

that item (Yin et al. 2012). Similarly, Shi (2013) proposes a Markovian graph-based recommendation approach, where weights on edges could be tuned to enhance the probability of recommending long-tail items. Other related studies propose to optimize the recommendation list based on different objectives, e.g., increasing accuracy, reducing popularity (Hamedani and Kaedi 2019), or prioritizing niche items based on their usage context (e.g., co-occurrence with other popular items) (Niemann and Wolpers 2013).

Another set of studies propose *hybrid* approaches to improve long-tail recommendation performance. For example, in Alshammari et al. (2017), content-based and collaborative-filtering recommendation methods are used in combination to recommend long-tail and popular items, respectively. Similarly, in Zhang et al. (2012) and Ribeiro et al. (2015), the ensemble of outputs of multiple recommendation algorithms are used to balance accuracy and novelty. Other related studies use side information (i.e., information other than user-item interactions) to direct long-tail recommendation. Examples include adding semantic knowledge extracted from content information to better represent long-tail items (Craw et al. 2015) or explicitly collecting users' preferences for different types of items (Taramigkou et al. 2013).

In this paper, we focus on a new pattern-based recommendation method that tries to avoid some of the limitations of existing long-tail recommendation approaches (such as requiring a much richer set of features, significant pre-processing, or resulting in substantial reductions in accuracy), while exhibiting scalability, flexibility, and explainability benefits.

2.2.2 Association Analysis and Its Application to Recommender Systems

Association mining or frequent pattern discovery (Agrawal et al. 1993, Agrawal and Srikant 1994, Ceglar and Roddick 2006) is known as the task of discovering co-occurrences

between items, which has its roots in the analysis of shopping basket data to better understand consumer purchasing behavior. More formally, given a database of all users' consumption histories $\mathcal{T}$, $C_u \in \mathcal{T}$ is a set of items $\{i_1, i_2, \cdots, i_{|C_u|}\}$ consumed by the user u. A pattern can be represented either by an itemset or a rule. *Itemset* P is simply a set of items, i.e., $P = \{i_1, i_2, \cdots, i_K\}$ that represents some co-occurence relationship among the items. *Rule* $P \rightarrow Q$ is an association between two disjoint itemsets that can be interpreted as an if-then statement, i.e., if P happens, then Q happens as well.

Many different measures have been proposed to evaluate the pattern strength or interestingness (Tan et al. 2002), and *support* is one of the most fundamental, popular measures. The support of itemset P is given by $supp(P) = \sigma(P)/|\mathcal{T}|$, where $\sigma(P)$ is the support count of P defined as the number of users u for whom $P \subseteq C_u$ (i.e., users who consumed all items in P). For example, $supp(P) = 0.3$ means 30% of users have consumed all items that are present in pattern P. Similarly, the support of rule $P \rightarrow Q$ is defined as $supp(P \rightarrow Q) = \sigma(P \cup Q)/|\mathcal{T}|$. In summary, the support measure reflects the prevalence of a pattern in the data. In addition to the use of support as the pattern prevalence metric, rule discovery commonly uses an additional metric, i.e., *confidence*, which is calculated as $conf(P \rightarrow Q) = supp(P \rightarrow Q)/supp(P)$, higher confidence indicating stronger association between P and Q. For example, $conf(P \rightarrow Q) = 0.6$ means that, out of all users who consumed all items in P, 60% of them also consumed all items in Q. Confidence could also be interpreted as conditional probability, i.e., users who consume items in P tend to also consume items in Q with a probability of 60%.

If the support of itemset P satisfies a prespecified minimum support threshold τ_s, i.e., $supp(P) \geq \tau_s$, then P is a *frequent* itemset or pattern. Similarly, if the support and confidence of the rule $P \rightarrow Q$ satisfy pre-specified minimum support and confidence thresholds τ_s and τ_{conf}, i.e., $supp(P \rightarrow Q) \geq \tau_s$ and $conf(P \rightarrow Q) \geq \tau_{conf}$, then $P \rightarrow Q$ will be a discovered association rule. These thresholds can be set by users or domain

experts, and several algorithms, e.g., *Apriori* (Agrawal et al. 1993) and *FP-growth* (Han et al. 2000) have been proposed to discover frequent patterns and association rules efficiently from data.

Association-based approaches have been used in many application domains, including recommender systems, for their high interpretability (Sarwar et al. 2000, Lin et al. 2002, Davidson et al. 2010). For example, in an e-commerce application, the discovered rule *"computer, monitor → keyboard, mouse"* could be used to recommend a keyboard and a mouse to consumers who already have a computer and a monitor in their shopping carts.

However, the traditional association rule discovery framework based on support and confidence has certain limitations that make it less appealing for recommendation and, more specifically, long tail recommendation. In particular, the *confidence* metric often might not reflect a meaningful association among items, in part due to item-popularity-related issues, as illustrated by the classic "coffee-tea" example (Brin et al. 1997, Tan et al. 2006). Consider a scenario where 90% of shopping baskets have coffee, i.e., in general consumers buy coffee 90% of the time. Furthermore, let's assume that, among all baskets containing tea, 75% of them contain coffee as well, i.e., the confidence of rule *tea* → *coffee* is 75%. In other words, even though the confidence of buying coffee given that tea is already in the shopping basket is quite high (i.e., 75%), the two items are actually negatively associated with each other, i.e., buying tea reduces users' probability of buying coffee. Thus, recommending based on such rules is likely to reduce recommendation accuracy.

To deal with the aforementioned limitation, correlation-oriented measures can be used to augment the support-confidence framework for association rules (Tan et al. 2006). Among such measures, *lift* has been a popular choice. Lift of rule $P \rightarrow Q$ is calculated as $lift(P \rightarrow Q) = conf(P \rightarrow Q)/supp(Q) = \frac{supp(P \rightarrow Q)}{supp(P) * supp(Q)}$, which reflects

Table 2.1: Impact of null-consumptions on pattern/rule evaluation measures.

Data Set	A	B	AB	$\overline{A}B$	$A\overline{B}$	$\overline{AB}$	$conf(A \rightarrow B)$	$lift(A \rightarrow B)$	$cos(\{A, B\})$
D_1	11,000	15,000	10,000	5,000	1,000	100,000	0.91	7.03	0.78
D_2	11,000	15,000	10,000	5,000	1,000	100	0.91	0.98	0.78

A: # of users who consumed A; AB: # of users who consumed A and B; $\overline{AB}$: # of users who consumed neither A nor B; B: # of users who consumed B; $\overline{A}B$: # of users who consumed B not A; $A\overline{B}$: # of users who consumed A not B.

the degree to which the occurrence of P "lifts" the occurrence of Q. However, the lift metric is sensitive to *null-consumption* histories, i.e., to the number of users who did not consume *any* items contained in the rule of interest. As a quick illustration, Table 2.1 shows two example data sets about consumption of items A and B, including statistics on number of users who consumed both A and B (denoted as AB), not A but B ($\overline{A}B$), A but not B ($A\overline{B}$), neither A nor B ($\overline{AB}$). The latter value represents the number of null-consumption histories with respect to A and B. The table shows that $lift(A \rightarrow B)$ changes significantly with the size of $\overline{AB}$, even when the consumption statistics of A, B, and AB are identical, while we don't see such problems with other metrics like *confidence* and *cosine*, i.e., $conf(A \rightarrow B)$ and $cosine(\{A, B\})$ remain the same in both cases. In summary, for any general pattern $P \rightarrow Q$, such sensitivity of a key association metric (i.e., lift) to null-consumption histories might not be desirable, as the metric becomes more reflective of the prevalence of the pattern in data (which is already captured by support) than of the association between P and Q. As a result, *cosine* can serve as a more appropriate metric than *lift* to measure the association among items.

Another limitation of the support-confidence framework is that the traditional support-based pruning strategy for pattern mining (based on setting an appropriate support threshold) might be inadequate on data with skewed support distributions (Xiong et al. 2006). Lower thresholds often result in excessive or redundant patterns, as well as patterns that have items with substantially different support levels (i.e, the so-called

cross-support patterns, to be discussed in Section 2.2.3), leading to lower-quality recommendations. Higher thresholds favor highly frequent items and omit less frequent but potentially advantageous itemsets (patterns), which leads to popularity bias.

To address the limitations of confidence and lift metrics, we propose to use *cosine* patterns for recommendation. In what follows, we give some preliminaries on cosine patterns and describe important properties that make them particularly suitable for long tail and scalable recommendation.

2.2.3 Cosine Patterns and Their Properties

In this paper, we adopt *cosine* (Tan et al. 2002) as an interestingness measure to be used simultaneously with support for pattern evaluation and pattern-based recommendation. The *cosine* value of a K-itemset (i.e., an itemset of K items) P is defined as:

$$cos(P) = \frac{supp(P)}{\left(\prod_{k=1}^{K} supp(\{i_k\})\right)^{1/K}},\ K \geq 2, \tag{2.1}$$

which reduces to the traditional *cosine* measure when $K = 2$ (Wu et al. 2012). An important advantage of the cosine metric, as can be seen from the example in Table 2.1, is that the cosine value of a pattern is not affected by the number of null-consumptions.

In traditional association analysis, itemset P is called a *frequent* pattern if $supp(P) \geq \tau_s$, where τ_s is the user-defined minimum support threshold. Cosine pattern discovery takes one more threshold τ_c, i.e., the minimum cosine threshold, for pattern evaluation.

Definition 2.2.1 (Cosine Pattern)**.** Itemset P is a cosine pattern w.r.t. τ_s and τ_c, if $supp(P) \geq \tau_s$ and $cos(P) \geq \tau_c$, where τ_s and τ_c are user-defined thresholds.

As is shown in Equation (1), the cosine value of a pattern is calculated as the support (i.e., overall prevalence) of the pattern normalized by the geometric mean of the support of each single item within the pattern. Intuitively, cosine value reflects

the "cohesiveness" of a pattern. Patterns with higher cosine values contain items with similar popularity, indicating stronger association among items. This is also independent of the overall prevalence of the pattern, i.e., patterns with lower support can have high cosine values as long as the co-occurring items have similar support. Such patterns are extremely useful for long-tail recommendation, as will be illustrated in Section 2.3.1. Another key appeal of the cosine measure lies in its *anti-cross-support* property. As defined by Xiong et al. (2006), P is a *cross-support pattern* (CSP) w.r.t. τ ($0 \leq \tau \leq 1$) if its CSP value $V_{csp}(P) \leq \tau$. Here $V_{csp}(P) = s(i_l)/s(i_h)$, with i_l and i_h representing items with lowest and highest support values in P, respectively. A smaller $V_{csp}(P)$ value indicates a more severe imbalance of item supports in a pattern. By its definition, a CSP is a pattern containing items with significantly different support levels and, thus, more likely representing spurious, less cohesive (and potentially less meaningful) associations among items, as will be discussed in more detail in Section 2.3.1. The minimum *confidence* threshold, which is traditionally used in association rule mining, is not sufficient for filtering out CSPs. It could be argued that CSPs can be filtered by setting a higher τ_s, but this would lead to the omission of rare but interesting (niche) patterns and, thus, to the information loss for recommendation, especially for long tail recommendation. As shown by Wu et al. (2012), for $P = \{i_1, i_2, \cdots, i_K\}$, $cos(P) \leq \sqrt[K]{V_{csp}(P)}$. This implies that a pattern tends to have a lower cosine value as $V_{csp}(P)$ gets smaller; i.e., the patterns with lower cross-support values are less likely to be cosine patterns. Thus, cosine measure has the *anti-cross-support* property.

The anti-cross-support property of the cosine metric allows to control the "cohesiveness" of patterns discovered, making cosine patterns suitable for flexible recommendations. For example, setting large τ_s and moderate τ_c allows to obtain cohesive patterns with highly popular items; in contrast, reducing τ_s while keeping τ_c high allows to obtain the most cohesive patterns with both popular and niche items. Both cases, based

on threshold use, can guard against excessive generation of redundant patterns while providing flexibility to recommend items of desired popularity levels.

Mining cosine patterns based on τ_s and τ_c thresholds is not a trivial task. As is well known, the support measure holds the so-called *anti-monotone property* (AMP) defined as follows.

Definition 2.2.2 (AMP)**.** Itemset interestingness measure M is said to possess **AMP**, if for every P and P' such that $P \subset P'$, we have $M(P) \geq M(P')$.

Since support holds AMP, if itemset P is infrequent (i.e., $supp(P) < \tau_s$), then *all* its supersets are also infrequent (i.e., $\forall X$ such that $X \supset P$, we have $supp(X) \leq supp(P) < \tau_s$) and can be removed from consideration without further calculations. AMP of an interestingness measure is critical for efficient pattern mining (Agrawal and Srikant 1994). While the cosine measure does not hold AMP, it holds the following *Conditional Anti-Monotone Property* (CAMP) (Wu et al. 2012).

Definition 2.2.3 (CAMP)**.** Itemset interestingness measure M is said to possess **CAMP**, if for every P and P' such that (i) $P \subset P'$ and (ii) $\forall\ i \in P$, $\forall\ i' \in P' \setminus P$, $s(\{i\}) \leq s(\{i'\})$, we have $M(P) \geq M(P')$.

Compared with AMP, CAMP provides an extra condition to gain the anti-monotone property. Because cosine holds CAMP, cosine patterns can be mined efficiently. In particular, if P is not a cosine pattern, then any pattern P' that can be created by adding high-support items to P (i.e., items that have higher support than any item in P) will not be a cosine pattern due to CAMP, and thus can be removed from consideration without further calculations. This principle allows to avoid extra computations and lays the foundation of efficient cosine pattern mining.

The *CoPaMi* algorithm was proposed specifically to mine cosine patterns (Wu et al. 2014). It aligns the items in each consumption history in a support-ascending order (to

facilitate CAMP conditions for patterns sharing the same prefix), and then employs the tree-based data structure and depth-first traversal strategy to effectively prune patterns based on both support and cosine. We parallelize *CoPaMi* on the Spark platform and use it to mine cosine patterns in our experiments.

2.3 Cosine Pattern-Based Recommendation

2.3.1 Cosine Patterns for Recommendation

As mentioned earlier, pattern-based recommender systems have attracted substantial attention (Lin et al. 2002, Zaíane 2002, Kazienko 2009), partly for their high interpretability of recommendations. A recent survey (Paraschakis et al. 2015) on more than 30 popular e-commerce platforms also reveals that industries often favor less complex recommendation techniques like association rules mining or nearest neighbor-based collaborative filtering for efficiency and engineering cost concerns. This indicates that improving the recommendation performance of pattern-based (such as itemset- or rule-based) methods is of theoretical and practical importance.

We argue that the prevalence of the cross-support patterns is largely responsible for the lower accuracy and higher popularity bias of traditional rule-based recommender systems. To illustrate this, Table 2.2 shows some representative examples of 2-item movie association rules (10 out of top 150 highest confidence rules) and 2-item cosine patterns (10 out of top 150 highest cosine patterns)[1] discovered from the MovieLens data[2] with $\tau_s = 2\%$, $\tau_{conf} = 76\%$, and $\tau_c = 0.5$.[3]

[1]For comparison, the illustrative sets of 10 cosine patterns and 10 association rules were picked to have similar *support* (and to be representative of the broad range of support values).

[2]Detailed information of this data set could be found in Table 2.4.

[3]Thresholds were set to have similar number (between 150 and 200) of patterns discovered in both cases.

From Table 2.2 we can see that there is a large imbalance of support levels between the antecedent and consequent in each example association rule, as indicated by low V_{csp} values. In contrast, the V_{csp} values for cosine patterns are substantially higher. Moreover, in the last two columns of Table 2.2, we present two additional indicators of how related are the two movies that appear in each pattern (or how "cohesive" the pattern is). The first indicator is the correlation coefficient (CorrCoef) of the ratings for two movies, i.e., how similar are the preferences for these two movies among the users who saw both of them. The second indicator is the Jaccard similarity (JSim) of the movie consumptions, i.e., how similar are the sets of users who saw each movie. Both of these indicators consistently show that association rules contain items that are less related to each other (i.e., have substantially lower CorrCoef and JSim) than cosine patterns. This insight is further emphasized by Table 2.3, which provides the aggregate statistics across top-150 association rules and top-150 cosine patterns. In particular, the patterns mined based on the confidence measure (i.e., association rules) contain items that are highly imbalanced in terms of their support and substantially less related to each other than the patterns mined based on the cosine measure.

In summary, recommendations generated from association rules have several limitations. In particular, as shown above, association rules tend to contain items that are less related to each other, which can lead to lower accuracy when deployed for recommendations. As importantly, due to the fact that many discovered association rules have low cross-support values, the movies that end up being recommended using association rules (i.e., movies in the consequent of the rule) are largely high-support items (i.e., popular movies), such as *Star Wars*, *Fargo*, *Back to the Future*, and *Raiders of the Lost Ark*. This perpetuates the so-called popularity bias existing in many collaborative-filtering-based recommender systems, leading to insufficient recommendation of niche movies and long-tail recommendation challenges. As shown in Tables 2.2 and 2.3, cosine

pattern mining can effectively overcome these limitations due to its anti-cross-support objective, which allows for discovery of more cohesive patterns, including patterns with less common (niche) items. Take pattern {*Manon of the Spring, Jean de Florette*} as an example. Although movies in this pattern have been watched only by a relatively small number of users, the cohesiveness of this pattern is quite high: the movies are liked very similarly by users who saw both of them (CorrCoef = 0.76), and the sets of users who saw each movie are quite similar as well (JSim = 0.60). Nevertheless, such a pattern is unlikely to be discovered as an association rule due to its relatively low confidence. The ability to discover connections among items with comparatively smaller audience is a key to addressing the long-tail recommendation challenge and, thus, is one of the motivating factors for the proposed cosine pattern-based method for long-tail recommendation. At the same time, the minimum support and cosine threshold parameters provide the flexibility to fine-tune the proposed method to the desired specifications (e.g., in terms of recommending popular vs. niche items), as will be shown later in the paper.

2.3.2 Basic Recommendation Scheme

Here we introduce the basic scheme of cosine pattern-based recommendation. We assume that the set of all applicable cosine patterns $\mathcal{CP}$ (i.e., patterns with $supp(P) \geq \tau_s$ and $\cos(P) \geq \tau_c$ for some user-specified thresholds τ_s and τ_c) has been mined in advance (e.g., using the standard library *CoPaMi*) and focus on the problem of generating top-K recommendations for each user.

Given the set of all discovered cosine patterns $\mathcal{CP}$ and user u (represented by her consumption history C_u), the recommendation process consists of three main stages: (i) identifying u's target items T_u; (ii) identifying the set of eligible patterns $\mathcal{EP}_{ui}$ (where $\mathcal{EP}_{ui} \subseteq \mathcal{CP}$) for each target item $i \in T_u$; and (iii) calculating recommendation scores

Table 2.2: Association rules and cosine patterns discovered from the MovieLens dataset.

Association Rules							
Antecedent	**Consequent**	**Supp (%)**	**Conf (%)**	**Cosine**	**V_{csp}**	**Corr Coef**	**JSim**
Dead Man (34)	Star Wars (583)	2.3	77.3	0.15	0.06	-0.22	0.05
Only You (39)	The Princess Bride (324)	2.9	77.8	0.24	0.12	0.06	0.10
Giant (51)	Casablanca (243)	3.9	81.1	0.33	0.21	0.21	0.20
Trees Lounge (50)	Fargo (508)	4.1	79.5	0.24	0.10	0.23	0.10
Weekend at Bernie's (60)	Back to the Future (350)	4.6	79.1	0.30	0.31	0.31	0.16
Dumb and Dumber (50)	Back to the Future (350)	4.8	82.2	0.29	0.20	-0.06	0.18
Flirting With Disaster (42)	The Empire Strikes Back (367)	5.5	77.4	0.25	0.10	0.12	0.24
Nell (81)	Raiders of the Lost Ark (420)	5.8	76.4	0.30	0.20	0.32	0.17
Victor/Victoria (77)	Return of the Jedi (507)	5.8	80.0	0.28	0.15	0.09	0.13
Casino (91)	Raiders of the Lost Ark (420)	7.0	78.8	0.34	0.22	0.14	0.20
Cosine Patterns							
Movie 1	**Movie 2**	**Supp (%)**	**Conf (%)**	**Cosine**	**V_{csp}**	**Corr Coef**	**JSim**
Manon of the Spring (58)	Jean de Florette (64)	2.3	37.9	0.51	0.38	0.76	0.60
Three Colors: White (59)	Three Colors: Blue (64)	2.9	45.8	0.87	0.42	0.75	0.58
A Grand Day Out (66)	The Wrong Trousers (118)	3.9	56.1	0.58	0.31	0.66	0.50
Private Benjamin (66)	Home Alone (137)	4.1	58.2	0.52	0.28	0.32	0.31
Bram Stoker's Dracula (120)	Interview with the Vampire (137)	4.8	38.3	0.52	0.34	0.45	0.55
Batman Forever (114)	Batman Returns (142)	4.8	35.4	0.51	0.40	0.35	0.55
Star Trek: Final Frontier (63)	Star Trek: Motion Picture (117)	5.6	45.3	0.55	0.33	0.44	0.52
Die Hard With a Vengeance (151)	Die Hard 2 (166)	5.7	35.8	0.51	0.33	0.75	0.68
Under Siege (124)	Clear and Present Danger (179)	5.8	44.4	0.51	0.31	0.50	0.47
Ghost (170)	Mrs. Doubtfire (192)	7.0	39.4	0.51	0.35	0.45	0.48

In parentheses after each movie title: the number of users who rated the movie;
V_{csp}: support ratio of items within rules/patterns;
CorrCoef: correlation coefficient of movie ratings;
JSim: Jaccard similarity of movie consumptions.

Table 2.3: Descriptive statistics of top-150 rules/patterns.

	Supp (%)	**Conf (%)**	**Cosine**	**V_{csp}**	**CorrCoef**	**JSim**
Association Rules	3.0 (1.08)	80.37 (3.49)	0.25 (0.05)	0.13 (0.05)	0.17 (0.22)	0.11 (0.05)
Cosine Patterns	11.6 (3.70)	42.50 (4.00)	0.53 (0.02)	0.83 (0.14)	0.32 (0.15)	0.55 (0.06)

Standard Deviation in Parentheses.

for each $i \in T_u$ (by aggregating information from $\mathcal{EP}_{ui}$) and ranking all target items according to the scores. We describe each stage below.

Stage 1: For each user u, target item set T_u consists of items not yet consumed by user u, i.e., $T_u = I \setminus C_u$, where I represents the set of all possible items.

Stage 2: For each target item i in T_u, we need to find eligible pattern set $\mathcal{EP}_{ui}$. An *eligible pattern* is defined as follows.

Definition 2.3.1 (ui-Related Eligible Pattern)**.** Let P be a cosine pattern and C_u be user u's consumption history. P is an **eligible pattern** for u w.r.t. i, if $i \in P$, $i \notin C_u$, and $P \setminus \{i\} \subseteq C_u$.

In other words, cosine pattern P is a ui-related eligible pattern if: (1) P contains target item i, and (2) all other items in P (i.e., other than i) have been consumed by user u. Thus, any ui-related eligible pattern represents a cohesive itemset consisting of some items already consumed by u and one new item i, making item i a natural candidate for recommendation. The set of all ui-related eligible patterns is denoted as $\mathcal{EP}_{ui}$.

As a simple illustration, let's assume that we have a recommendation application with seven items, i.e., $I = \{A, B, C, D, E, F, G\}$, where the following three cosine patterns (itemsets) have been mined from users' consumption histories: $\mathcal{CP} = \{\{A, B, C\}, \{A, D, E, G\}, \{A, B, D\}\}$. Also, let's assume that user u's consumption history is $C_u = \{B, C, D, E\}$. Then, user u's target item set is $T_u = I \setminus C_u = \{A, F, G\}$. Considering item A as a target item, we can see that $\mathcal{EP}_{uA} = \{\{A, B, C\}, \{A, B, D\}\}$. In other words, $\{A, B, C\}$ is a uA-related eligible pattern, since all items in it except A have been consumed by u, and so is pattern $\{A, B, D\}$. Note, however, that $\{A, D, E, G\}$ is not an uA-related eligible pattern even though it also contains target item A, because there are multiple (i.e., more than one) target items in it (i.e., A and G).

Stage 3: For user u, recommendation score for target item i is derived from $\mathcal{EP}_{ui}$ by summing the cosine values of all patterns in $\mathcal{EP}_{ui}$, i.e., $score(u, i) = \sum_{P \in \mathcal{EP}_{ui}} cos(P)$. For any given user u, all target items in T_u will be ranked by their recommendation scores, and the recommendation list L_u for user u would be generated by selecting the top-K ranked items.

In other words, the current version of CORE adopts a relatively simple scoring method for target items $i \in T_u$, which adds up the cosine values of all ui-related eligible patterns. There have been several studies investigating ways in which pattern-based recommendation algorithms could combine eligible patterns to provide better recommendations. For instance, Wickramaratna et al. (2009) proposed a Dempster-Shaffer-based

approach to combine rules with conflicting predictions. Ghoshal and Sarkar (2014) proposed to partition eligible rules into groups of rules with disjoint antecedents and same consequent and developed a probability model to select the group that maximizes the likelihood of purchasing target item for recommendation. Lin et al. (2002) adopted heuristics like adding up the supports and confidences of all eligible rules with the same consequent as its recommendation score. Even though there exist different strategies for determining recommendation scores for items based on discovered patterns, theoretical or empirical studies on deriving optimal strategies are rarely seen. In this study, we used the cosine value of a pattern, as it provides a meaningful quantification of the itemset cohesiveness, as discussed earlier. Furthermore, if target item i appears in *multiple* ui-related eligible patterns (i.e., provides a cohesive combination with several consumed itemsets of user u), arguably this provides an even stronger signal of item i's relevance to u; hence, we chose to use a simple aggregation of the cosine values across *all* eligible patterns to empirically show the benefits of using cosine patterns for long tail recommendation. This scoring approach is not only easy to implement and computationally scalable, but also demonstrates excellent recommendation performance (as demonstrated by our experimental evaluation). In-depth analysis of optimal recommendation score aggregation across multiple patterns represents an interesting direction for future work.

We call the above approach *COsine pattern-based REcommendation*, or CORE for short. As will be demonstrated in the evaluation section, by leveraging the anti-cross-support property of cosine patterns, CORE not only exhibits good recommendation accuracy, but is also able to successfully recommend long-tail items. Furthermore, the two threshold parameters for cosine pattern mining (i.e., τ_s and τ_c) provide CORE with flexibility in recommending items across the popularity distribution. For example, in

order to to recommend more popular items we can set a high τ_s and a moderate τ_c, while to recommend more niche items we can set a small τ_s but a high τ_c.

2.3.3 Cosine-Pattern Tree Traversal Approach

The key computational challenge of the proposed cosine-pattern-based recommendation process is finding ui-related eligible patterns for given user u's all target items i. To deal with this, we propose to use a data structure called Cosine-Pattern Tree to boost the eligible pattern discovery process, which leads to CORE+ (an enhanced version of CORE).

Based on the definition, for given user u and target item $i \in T_u$, a simple way to compute $\mathcal{EP}_{ui}$ is to first find all cosine patterns that contain i, i.e., $\mathcal{CP}_i = \{P \in \mathcal{CP} | i \in P\}$, and then keep only those where the remaining items are covered by C_u, i.e., $\mathcal{EP}_{ui} = \{P \in \mathcal{CP}_i | P \setminus \{i\} \subseteq C_u\}$. The most time-consuming aspect of this calculation is determining whether pattern $P \in \mathcal{CP}_i$ satisfies the latter condition. A straightforward way to do this is to examine each item in $P \setminus \{i\}$ to see whether it is contained in C_u. By storing C_u in a hash table, where time complexity to look up any item is $\Theta(1)$, the overall time complexity for checking all items in one pattern is $O(|P|)$, and it would take $O(\sum_{P \in \mathcal{CP}_i} |P|)$ to go through all candidates and identify all ui-related eligible patterns.

Under this strategy, two factors can slow down the recommendation process. First, the target item set, i.e., number of non-consumed items, for each user is typically very large; thus, time required to find candidate and eligible patterns for all items can add up quickly. Second, due to CAMP, a cosine pattern typically contains many similar cosine patterns as subsets, which would entail a great deal of repeated (redundant) matching of highly similar patterns with C_u.

Both of these factors can be addressed by using advantageous data structures and algorithms for storing and retrieving cosine patterns. In particular, to facilitate efficient cosine pattern traversal and reduce redundant matching in eligible pattern detection,

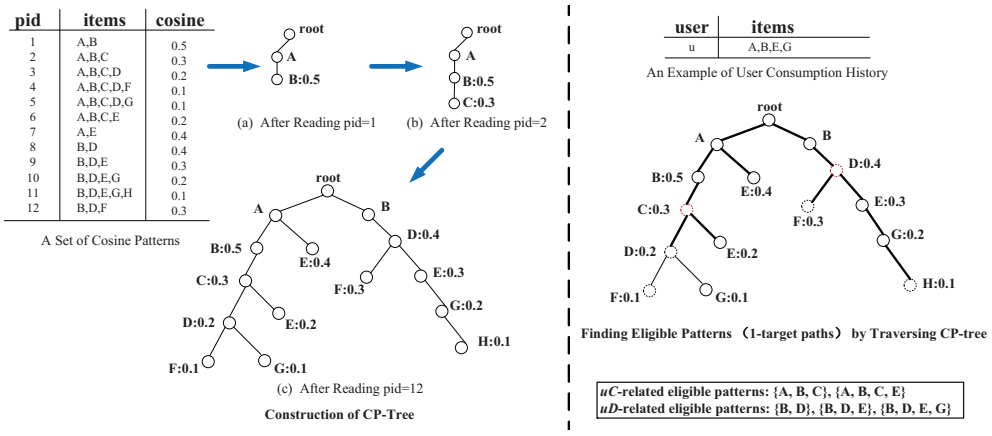


Figure 2.1: An example for CP-tree construction and traversal.

we build upon the notion of the FP-tree (Han et al. 2000) as well as the ideas from the *CoPaMi* cosine pattern mining approach (Wu et al. 2014) to use *Cosine-Pattern Tree* (CP-tree). While FP-tree is proposed for compact storage of user consumption and efficient discovery of frequent patterns, CP-tree in our study is used for compact representation of discovered cosine patterns. Based on CP-tree, an intelligent, depth-first search strategy is designed to find all eligible patterns efficiently.

Generally, CP-tree is constructed in a similar fashion as FP-tree, i.e., it is a fully-connected, hierarchical tree structure, where each node represents an item. Each path from the root node to any other node in the CP-tree represents a cosine pattern (itemset) containing all the items on that path. Thus, cosine patterns that share a common prefix will share (a part of) the same path.

More formally, CP-tree consists of two types of nodes: (i) one special node denoting the start of all paths (i.e., all patterns stored in the tree) and labeled as *root*, and (ii) possibly multiple regular nodes denoting items, each labeled with a corresponding item name. Each node has three fields, storing its relevant information: `item` (i.e., the label of this node), `cosine` (i.e., the cosine value of the pattern that ends with this node), and `childlist` (i.e., the list of children nodes that are connected to this node).

Before tree construction, all items in each cosine pattern P are sorted in a support-ascending order, which guarantees that every prefix of P is also a cosine pattern, due to CAMP. E.g., if $supp(A) \leq supp(B) \leq supp(C)$ and $\{A, B, C\}$ is a cosine pattern, then $\{A, B\}$ is guaranteed to be a cosine pattern, due to CAMP. An illustration of CP-tree construction process is shown on the left side of Fig. 2.1, where all items are assumed to be sorted in the support-ascending order. Initially, the CP-tree contains only the default root node. After reading the first cosine pattern $\{A, B\}$, the nodes labeled as A and B are created, and path $root \rightarrow A \rightarrow B$ is then added, and the value of $cos(\{A, B\}) = 0.5$ is saved in the `cosine` field of node B, as shown in Fig. 2.1a. The second cosine pattern $\{A, B, C\}$ shares a common prefix $\{A, B\}$ with the first pattern, and therefore only one new node marked as C is added to the end of path $root \rightarrow A \rightarrow B$ with the corresponding $cos(\{A, B, C\}) = 0.3$ value in Fig. 2.1b. This process continues until every cosine pattern has been mapped onto one of the paths in CP-tree, which leads to the final CP-tree in Fig. 2.1c. Because of this specific prefix-based tree construction as well as the CAMP property of the cosine measure, all distinct cosine patterns are represented by all the paths from the root node to *every* node that is below the first two levels of CP-tree (since a cosine pattern must have at least two items).

Thus, any arbitrary path $[P] = root \rightarrow i_1 \rightarrow \cdots \rightarrow i_p$ in CP-tree represents cosine pattern $P = \{i_1, \cdots, i_p\}$, when $|P| \geq 2$. With respect to specific user u, any given path $[P]$ (and its corresponding cosine pattern P) can be classified into three different categories – 0-target path, 1-target path, and multi-target path – depending on how many target items the path contains. In particular, $[P]$ is a *0-target path* if $|P \setminus C_u| = 0$ or, equivalently, $P \subseteq C_u$, i.e., the path does not contain any target items for user u. Alternatively, $[P]$ is a *1-target path* if $|P \setminus C_u| = 1$. Finally, $[P]$ is a *multi-target path* if $|P \setminus C_u| \geq 2$, i.e., if it contains multiple target items for user u. Note that only 1-target

paths represent *ui*-related eligible cosine patterns that can be used for recommendation, as they contain exactly one target item.

The above categorization suggests an intelligent and highly efficient computational strategy that allows, for any user u, to find all relevant target items and calculate their recommendation scores with a *single traversal through CP-tree*. Intuitively, the main idea is to traverse CP-tree, visiting nodes in a depth-first manner. Each visited node represents path $[P]$ (i.e., path from the root to this node), which can be one of the following: (i) 0-target path, in which case no recommendation decisions need to be made, and the depth-first search continues to the children of this node; (ii) 1-target path, in which case the recommendation score for target item i (that is contained in $[P]$) is increased by $cos(P)$, and the depth-first search continues to the children of this node; or (iii) multi-target path, in which case no recommendation decisions need to be made, and the depth-first search is stopped along this path, as all subsequent extensions of path $[P]$ will continue to be multi-target paths. In summary, the proposed approach does not have to check all possible target items, but rather finds them (and calculates their recommendation scores) organically in one shot by efficiently browsing cosine patterns and matching them with each user's consumption history.

Algorithm 1 provides more detailed overview of the implementation of the proposed search-and-scoring routine. The main function `SearchCP` (Lines 1-23) employs a depth-first search on CP-tree. Besides variable *cur* to indicate the current node, variable *target* is introduced to indicate the target item contained in the current eligible pattern. Traversal along any given path in CP-tree terminates as soon as the path becomes a multi-target path (Lines 13-14), which avoids unnecessary search of longer patterns. Last-in-first-out stack S is used to execute the depth-first traversal, and variable *target* corresponding to target item contained in the current path is pushed into (or popped out from) the stack simultaneously with each node (Lines 6, 8, 12, 17, 22). Lines 11

Algorithm 1 Eligible-pattern searching and target-item scoring from CP-tree.

Input: $root$, C_u ▷ $root$ is the root node of CP-tree, C_u is the consumption history of user u
Output: $score$ ▷ list of recommendation scores of all items for user u

```
1: procedure SearchCP(root, C_u)
2:     S := ∅;                                              ▷ S: a last-in-first-out stack
3:     for i ∈ I do                                         ▷ I: the set of all items
4:         score(i) := 0;                                   ▷ initialization of recommendation score for each item
5:     target := null                                       ▷ detected target item in the beginning is null
6:     PushChildNodes(S, root, target);
7:     while S ≠ ∅ do                                       ▷ continue traversing until the stack is empty
8:         (cur, target) := S.POP;                          ▷ pop out the top tuple (current node, target item) in S
9:         if target is not null then                       ▷ check whether a target item is already detected
10:            if cur.item ∈ C_u then                       ▷ check whether current item is in user's consumption history
11:                score(target) := score(target) + cur.cosine;   ▷ update score of the target item
12:                PushChildNodes(S, cur, target);          ▷ push node's children and target item into the stack
13:            else
14:                continue                                 ▷ traversal of a path stops when a second target item is detected
15:        else                                             ▷ in case no target item has been detected so far
16:            if cur.item ∈ C_u then
17:                PushChildNodes(S, cur, target);          ▷ keep traversing when no target item is detected
18:            else
19:                target := cur.item;                      ▷ update the value of taget when a target item is detected
20:                if cur is not child of root then         ▷ update score when the target item is beyond the first level
21:                    score(target) := score(target) + cur.cosine;
22:                PushChildNodes(S, cur, target);          ▷ push node's children and target item into the stack
23: end procedure
24: procedure PushChildNodes(S, cur, target)
25:     for node ∈ cur.childlist do
26:         S.Push(node, target);                           ▷ push current node's children and detected target item into the stack
27: end procedure
```

and 21 update the recommendation scores of target items whenever the traversed path is a 1-target path.

The right side of Fig. 2.1 illustrates the use of Alg. 1 for user u with specific consumption history C_u. Note that the cosine values of patterns are shown next to the end nodes of their corresponding paths, the dashed nodes are target items (i.e., items that are not in C_u), and thick lines represent actual traversals performed by Alg. 1. Consider traversal along $root \rightarrow A \rightarrow B \rightarrow C$. Because C is the first item not contained in C_u, $root \rightarrow A \rightarrow B \rightarrow C$ becomes a 1-target path with C as its target item. This path can be extended further either with D or with E. In the case of former, traversal would stop at node D, as it is the second item not contained in C_u, making the path a multi-target path at that point; going deeper would not detect any new eligible patterns.

This avoids redundant checks of successive nodes D, F, and G along the same path. In contrast, $root \rightarrow A \rightarrow B \rightarrow C \rightarrow E$ would be identified as another 1-target path with C as its target item. As another example, after traversal of path $root \rightarrow B \rightarrow D$, D is identified as another target item, and three 1-target paths with D as target item would be discovered: $root \rightarrow B \rightarrow D$, $root \rightarrow B \rightarrow D \rightarrow E$, and $root \rightarrow B \rightarrow D \rightarrow E \rightarrow G$. Eventually, in this example, five eligible patterns that contribute to recommending C and D are identified with a single traversal of CP-tree.

CORE+ recommendation approach is highly computationally efficient. Space complexity of CP-tree is $O(|\mathcal{CP}|)$, since the total number of nodes in the tree equals the total number of cosine patterns plus a small fixed number (i.e., the root node and its immediate children). Meanwhile, given any user consumption history C_u, time complexity of finding all eligible patterns using Alg. 1 is $O(|\mathcal{CP}|)$. To elaborate, there are $O(|\mathcal{CP}|)$ nodes in the CP-tree and, in the extreme case, all of them may have to be examined, with constant amount of time needed per node; e.g., to check whether an item is contained in C_u is $\Theta(1)$ using a hash table. Therefore, with CP-tree and Alg. 1, the time complexity of recommendation given C_u reduces substantially from $O(|T_u| \sum_{P \in \mathcal{CP}} |P|)$ to $O(|\mathcal{CP}|)$, where $|P|$ is the number of items in pattern P. This implies that the upper bound of the running time of CORE+ is simply proportional to the number of cosine patterns, regardless of their size or the number of potential target items.

2.3.4 Parallelizing Cosine Pattern-based Recommendation

In this section, we describe a distributed framework designed for CP-tree based recommendation, which leads to CORE++, a parallelized version of CORE+ for online recommendation.

Intuitively, *user partitioning*, i.e., distributing active users to different servers on which a complete CP-tree is stored in advance for recommendation, is a straightforward

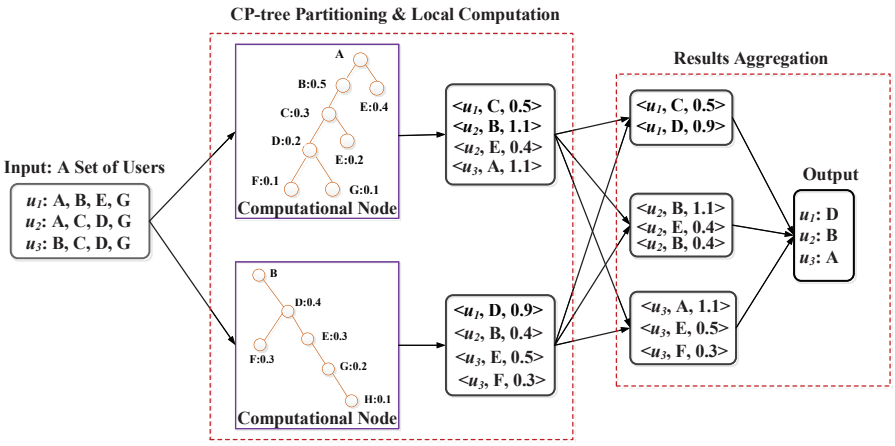


Figure 2.2: Illustration of parallelization based on CP-tree partitioning.

way to improve scalability. For example, according to the widely-used strategy of proxy cache servers (Gama et al. 2001), network traffic could naturally be reduced by replicating static content (CP-tree in our case). However, when the CP-tree is large, i.e., a huge amount of cosine patterns is discovered from consumption or purchase data of a gigantic online retailer like Amazon, recommendation generation for a single user can still take a considerable amount of time.

To address this challenge, we propose to use a parallelization framework based on *CP-tree partitioning* to further speed up recommendation. Specifically, by taking a complete CP-tree as a *forest* with a null root node, each subtree rooted at the second level can then be allocated to one computational node, and thus the entire CP-tree can be stored separately on Z available computational nodes. The parallelized CP-tree based recommendation (i.e., CORE++) could then be done in three stages: (i) broadcast user u's consumption history C_u to Z computational nodes; (ii) run Alg. 1 for u on Z nodes in parallel; (iii) aggregate local scores for each target item to obtain the final recommendation score. Fig. 2.2 illustrates the three stages using the example from Fig. 2.1, where the initial CP-tree in Fig. 2.1c is decomposed into two subtrees by

removing the root node. Each subtree is then distributed to a separate computational node, along with all users' consumption histories. Recommendation score for each target item based on each subtree would be generated by running Alg. 1 simultaneously on the two computational nodes. The final recommendation score for each target item can then be obtained by aggregating scores from each node (e.g., recommendation score of B for u_2 based on two subtrees is aggregated as $1.1 + 0.4 = 1.5$). In Fig. 2.2, for each user, we pick one target item with the highest score for recommendation.

In the above parallelization scheme, the time to generate recommendations for each user hinges on the slowest computational node. Thus, it is desirable to partition the CP-tree in such a way that workload on all nodes is evenly distributed. This could be formally defined as a *load-balanced partitioning* (LBP) problem as follows.

Definition 2.3.2 (LBP Problem)**.** Let $\mathcal{S} = \{s_1, s_2, \cdots, s_q\}$ be the set of subtrees and $\alpha_i \geq 0$, $i \in \{1, \ldots, q\}$, be a load indicator of the i-th subtree. Denote the set of subtrees distributed to z-th computational node as $\mathcal{S}(z)$, $z \in \{1, \ldots, Z\}$. Then, the load of the z-th node is $L_z = \sum_{i:s_i \in \mathcal{S}(z)} \alpha_i$, and the maximum load across all nodes is $L = \max_z L_z$. So, the LBP problem is to find an assignment $\mathcal{S}(z)$, $z \in \{1, \ldots, Z\}$, such that L is minimized.

Based on Def. 2.3.2, minimizing the overall load L essentially means finding Z disjoint subsets of $\mathcal{S}$ on which computation loads L_z ($z \in \{1, \ldots, Z\}$) are close. This, however, is an NP-complete problem (Cormen et al. 2001). Thus, we adopt the *longest processing time* (LPT) strategy which has been proved to be a 4/3-approximation (Graham 1969) for load balancing. The LPT strategy for CP-tree partitioning is applied by sorting all subtrees in descending order based on their load indicators α_i, and then sequentially allocating each tree to the node that currently has the lowest estimated computational load. This process continues until all subtrees have been assigned. The performance of LBP mainly depends on how well the load indicator values (α_i) can be

estimated, and below we propose two simple heuristic approaches for estimating actual computational load.

Intuitively, a subtree of CP-tree rooted with a higher support item is likely to represent patterns that are more commonly (frequently) found in a user population. Thus, such a subtree has a greater chance to be traversed for eligible patterns and, therefore, can entail more intensive calculations. Thus, our first heuristic approach is a *support-based* load indicator, where we use the support value of the subtree root item as the subtree load indicator. Alternatively, the size of a subtree is also related to traversal cost, as bigger subtrees contain more patterns. In other words, the eligible pattern discovery process may result in heavier computation load due to having to check potentially more patterns. Therefore, our second heuristic approach is a *pattern-based* load indicator, where we adopt the subtree size (i.e., number of patterns in a subtree) as the subtree load indicator.

Note that neither of the two indicators dominates the other theoretically; which one works better is context-dependent. The support-based load indicator is likely to be more suitable for relatively sparse data. In such cases, on average, C_u would contain few items, so most subtrees would have limited eligible patterns for recommending certain target items to u. This indicates that the computational cost would mainly come from multiple traversals of frequent subtrees, which would be captured by the support-based load indicator. In contrast, the pattern-based indicator is likely to perform better on denser data. In this case, on average, C_u would contain more items, and most of the subtrees are likely to be traversed. Thus, the computational cost may depend much more on the size of subtrees (the number of patterns to be checked), which would be estimated by the pattern-based load indicator. We compare the two indicators empirically in Section 2.4.5.

Table 2.4: Summary statistics of datasets.

	Rating	#Users	#Items	#Ratings	Sparsity(%)	Avg.#Ratings per Item	% of Ratings from Top 20% Items	% of Ratings from Top 50% Items	Gini Coefficient
`Book-Crossing`	[1,10]	1834	2172	41,337	98.96	13	40%	65%	0.37
`Last.fm`	[1,5]	635	4100	14,175	99.46	5	53%	70%	0.57
`MovieLens`	[1,5]	943	1682	100,000	93.70	59	65%	93%	0.63

Sparsity=100(1−#Ratings/(#Users×#Items)), i.e., percentage of unknown ratings.

2.4 Experimental Results

In this section, we conduct computational experiments to evaluate CORE. We first compare CORE to two types of baselines, i.e., the pattern-based and the classic CF-based (Collaborative Filtering) methods. We then analyze the advantages of CORE for long-tail recommendation followed by a discussion on the flexibility of CORE in recommending items of different popularity levels.

2.4.1 Experimental Setup

Data. CORE is tested using three publicly available datasets[4] that are widely used in recommender systems research: `Book-Crossing`, `Last.fm`, and `MovieLens`. For `Book-Crossing` and `Last.fm`, we sample the data to include users with at least ten ratings to avoid extreme sparsity. Table 2.4 provides the summary statistics of the three data sets, including the information about the distributional inequality in item consumption. E.g., if we consider the most popular 20% of items in each dataset, in `MovieLens` such items receive over 65% of all ratings, while the numbers are significantly lower for the other two datasets, i.e., 40% and 53%, indicating a heavier-tailed nature of `Book-Crossing` and `Last.fm` datasets. The Gini Coefficient values further highlight the differences among datasets in terms of item rating frequency distributions.

[4]The datasets can be found at `http://grouplens.org/datasets/hetrec-2011/`.

Performance metrics. Even though the main motivation behind the proposed approach is addressing the long-tail recommendation challenges, recommendation accuracy is always an important performance dimension. Precision (or precision-in-top-k) is commonly used to evaluate the accuracy of top-K recommendation lists and is calculated as follows. For each user u,

$$Precision_u = \#hits_u/K, \tag{2.2}$$

where $\#hits_u$ is the number of items from user u's recommendation list that are also in u's test set, and K is the length of the recommendation list (by default, we use $K = 10$ in our experiments).

To evaluate the long-tail performance of recommendation algorithms, we use two metrics: (i) the average popularity (*AvgPop*) of items in the top-K recommendation list, where each item's popularity is reflected by the number of ratings it has (Yin et al. 2012, Niemann and Wolpers 2013), and (ii) the ratio of niche items (*NicheRatio*) in the top-K recommendation list. In a given dataset, an item is defined to be a *niche* item if it has fewer ratings than the average number of ratings per item in the data set. More formally, for each user u,

$$AvgPop_u = (\sum_{n=1}^{K} \#ItemRatings_n)_u/K, \qquad NicheRatio_u = \#NicheItems_u/K. \tag{2.3}$$

The overall performance for each metric is obtained by averaging its values over all users.

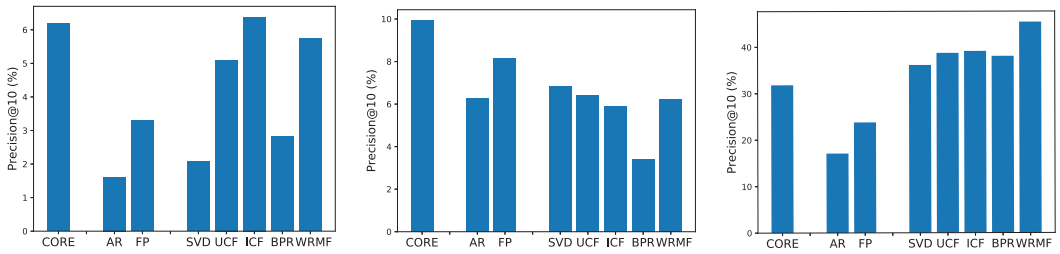
Baselines. We compare CORE with two types of baselines, i.e., pattern-based and collaborative filtering methods. The former includes the association rule-based method (AR) (Lin et al. 2002) and the frequent pattern-based method (FP) (Nakagawa and

Mobasher 2003). The collaborative filtering category includes five widely adopted methods: UCF (User-based Collaborative Filtering) (Resnick et al. 1994), ICF (Item-based Collaborative Filtering) (Sarwar et al. 2001), SVD (Funk 2006), WRMF (Weighted Regularized Matrix Factorization) (Hu et al. 2008), and BPR (Bayesian Personalized Ranking) (Rendle et al. 2009). In particular, AR, FP, BPR, WRMF and CORE are designed specifically for implicit feedback data (i.e., 0/1 data that reflects only whether a user consumed or purchased an item, and not the user's explicit preference rating for that item). For unified comparison with different baselines, we first convert explicit ratings in the three data sets to consumption (i.e., 0/1) data, based on the absence/presence of user rating. We then adopt the standard split-validation method commonly used in recommender systems research to evaluate different methods; that is, we randomly select 70% percent of the consumed items of each user for model training purposes, and use the remaining 30% of items for performance evaluation. Hyperparameters of each algorithm – e.g., the neighborhood size in UCF and ICF, the number of latent factors in SVD, WRMF, and BPR, the support and cosine thresholds for CORE, etc. – were carefully tuned using standard predictive modeling practices for best accuracy performance.

2.4.2 Recommendation Accuracy Performance

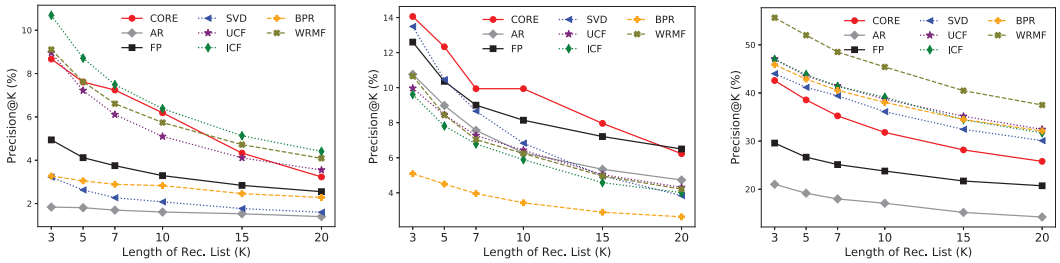
Fig. 2.3 compares CORE to different baselines in terms of precision-in-top-10 (30% of data is used for testing). In terms of accuracy comparisons with other pattern-based approaches, the results show that the precision of CORE is consistently higher than that of AR and FP across all data sets. In terms of accuracy comparisons with collaborative filtering methods, Fig. 2.3 shows that CORE is highly competitive on `Book-Crossing` (only very slightly behind the best baseline ICF) and `Last.fm` (best performance among all methods). On `MovieLens`, collaborative filtering baselines show

performance advantages over CORE, and we will take a closer look at this later in the paper.



(a) `Book-Crossing` (b) `Last.fm` (c) `MovieLens`

Figure 2.3: CORE vs. baseline algorithms on accuracy (test-set ratio = 30%).

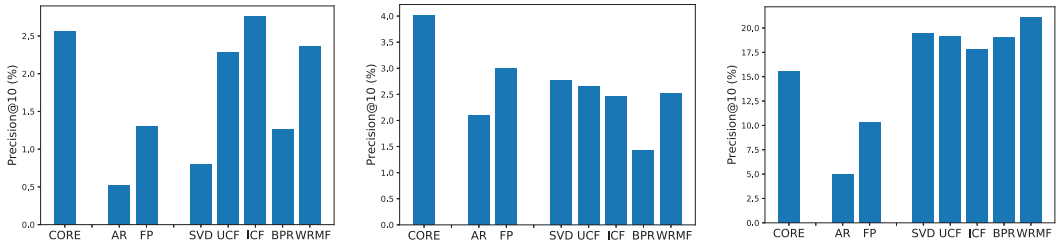


(a) `Book-Crossing` (b) `Last.fm` (c) `MovieLens`

Figure 2.4: CORE vs. baseline algorithms on accuracy under different K.

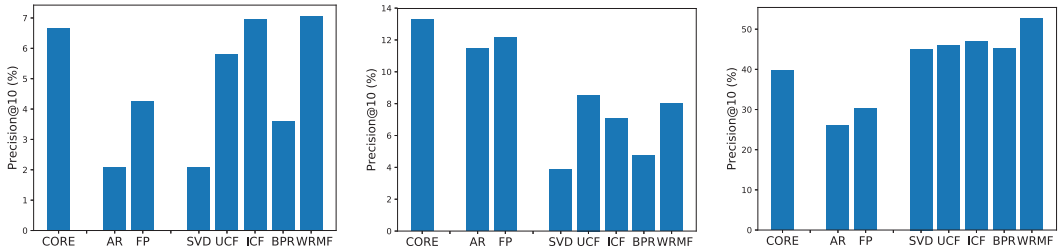
To demonstrate the robustness of the CORE performance under different settings, we provide accuracy comparisons for different lengths of the top-K recommendation list and for different splits of training and test data. The results are shown in Figs. 2.4-2.6. Specifically, Fig. 2.4 shows that the relative accuracy performance of different methods remains consistent when different number of recommendations is provided. I.e., CORE remains competitive on `Book-Crossing` and demonstrates superior performance on `Last.fm`; on `MovieLens`, CORE is outperformed by collaborative filtering baselines, but performs better than pattern-based methods. Relative accuracy performance also remains consistent for different training-test data splits, as illustrated by Figs. 2.5 and 2.6, which show the precision comparisons among different methods when

90% and 50% of the data is used for model learning (the remaining 10% and 50% are used for model evaluation), respectively.



(a) `Book-Crossing` (b) `Last.fm` (c) `MovieLens`

Figure 2.5: CORE vs. baseline algorithms on accuracy (test-set ratio = 10%).



(a) `Book-Crossing` (b) `Last.fm` (c) `MovieLens`

Figure 2.6: CORE vs. baseline algorithms on accuracy (test-set ratio = 50%).

Overall, accuracy comparisons in Figs. 2.3-2.6 demonstrate highly promising performance of CORE; i.e., it dominates pattern-based methods across all data sets and is highly competitive with the widely used CF-based approaches on two heavier-tailed data sets, i.e., `Book-Crossing` and `Last.fm`.

2.4.3 Long-Tail Recommendation Performance

The key objective of the proposed approach is the long-tail recommendation. Therefore, in this section, we go beyond recommendation accuracy and focus on the long-tail recommendation performance evaluation. Tables 2.5 and 2.6 show the average popularity (AvgPop), i.e., average number of ratings, and ratio of niche items (NicheRatio) of

Table 2.5: Average popularity of recommended items.

Data Set	AvgFreq	CORE	AR	FP	UCF	ICF	SVD	WRMF	BPR
`Book-Crossing`	13	43	76	101	72	45	69	65	92
`Last.fm`	5	31	89	107	61	52	34	48	96
`MovieLens`	59	135	318	383	333	338	397	301	275

AvgFreq: average number of ratings per item in a data set.

Table 2.6: Percentage of niche items recommended.

Data Set	AvgFreq	CORE(%)	AR(%)	FP(%)	UCF(%)	ICF(%)	SVD(%)	WRMF(%)	BPR(%)
`Book-Crossing`	13	13.50	0.05	0.04	2.10	1.38	1.81	3.20	0.40
`Last.fm`	5	16.90	1.72	0.02	2.16	4.44	5.77	9.30	0.20
`MovieLens`	59	25.35	0.00	0.00	0.02	0.20	0.00	0.11	1.41

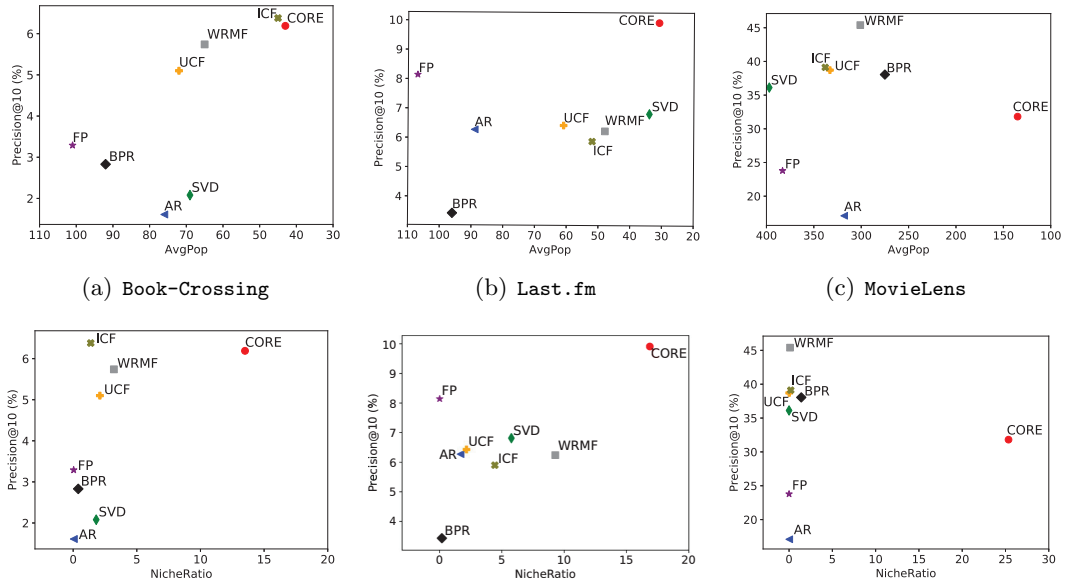
AvgFreq: average number of ratings per item in a data set.

top-10 items recommended by the same exact model configurations evaluated in Section 2.4.2. Again, the same 30% of data was used for testing.

The results highlight the significant advantages of CORE over baseline algorithms for long-tail recommendation. In particular, baseline algorithms tend to recommend popular items, as indicated by the average popularity metric and by the fact that, in the vast majority of settings, less than 2% of recommendations by baseline algorithms are niche items. These results are consistent with the findings of previous studies (Fleder and Hosanagar 2009) that most existing recommender systems have *popularity bias*, creating the rich-get-richer effect for popular items. In contrast, CORE is successfully able to provide recommendations of items with lower popularity and recommendations containing substantially higher percentage of niche items across all datasets.

Fig. 2.7 reiterates the results from Sections 2.4.2 and 2.4.3 by comparing the overall performance of different recommendation techniques in a two-dimensional (i.e., accuracy vs. long-tail recommendation) space. Specifically, Figs. 2.7a-2.7c show each method's position in the *Precision-AvgPop* performance space and, similarly, Figs. 2.7d-2.7f show *Precision-NicheRatio* comparison. Methods appearing in the top-right corner in these figures demonstrate better performance on both accuracy and long-tail recommendation. As shown in Fig. 2.7, for `Book-Crossing` and `Last.fm`, CORE is not only the advantageous choice in terms of the long-tail recommendation performance, but it is also an

excellent overall choice based on *both* performance dimensions. For `MovieLens`, CORE demonstrates dramatic improvements in long-tail performance at the expense of some accuracy reduction with respect to CF baselines (but still significantly outperforming pattern-based baselines).



(a) `Book-Crossing` (b) `Last.fm` (c) `MovieLens`

(d) `Book-Crossing` (e) `Last.fm` (f) `MovieLens`

Figure 2.7: CORE vs. baseline algorithms on accuracy and long-tail recommendation.

2.4.4 Additional Experiments

In this section, we discuss a number of additional important characteristics of the CORE approach.

Preference for Heavier-Tailed Datasets. From the main results discussed so far, we see that traditional pattern-based and collaborative filtering methods tend to recommend items with higher popularity, while CORE is able to recommend substantially more long-tail items across all data sets. In terms of accuracy performance, on more skewed datasets (such as `MovieLens`, where small percentage of popular items are responsible for high percentage of ratings/consumptions, as could be seen in Table 2.4),

traditional (popularity-oriented) collaborative filtering techniques tend to have some inherent advantage. However, this accuracy advantage of traditional techniques disappears on heavier-tailed data sets (such as `Book-Crossing` and `Last.fm`, with larger percentage of ratings dispersed to niche items), where CORE exhibits highly competitive accuracy performance.

We provide additional support for this finding with the following experimental evaluation, where we compare CORE with WRMF (i.e., one of the collaborative filtering baselines that demonstrates consistently good performance) on datasets with varying degrees of skewness.

In particular, we take `MovieLens` data, where WRMF consistently demonstrates superior accuracy performance over other methods (including CORE), manipulate its rating distribution to achieve different levels of skewness, and investigate the changes in relative performance of WRMF vs. CORE. Specifically, we first rank all items in `MovieLens` by their support (i.e., the number of ratings), and then remove items ranking in top 10%, 20%, 30%, 40%, and 50%, respectively, to generate five new data sets with heavier-tailed distribution. Descriptive statistics and distributions of these new datasets are shown in Table 2.7, from which we can see that, the more popular items are being removed, the closer the rating distribution is to the one in `Last.fm`, one of our heavier-tailed datasets.

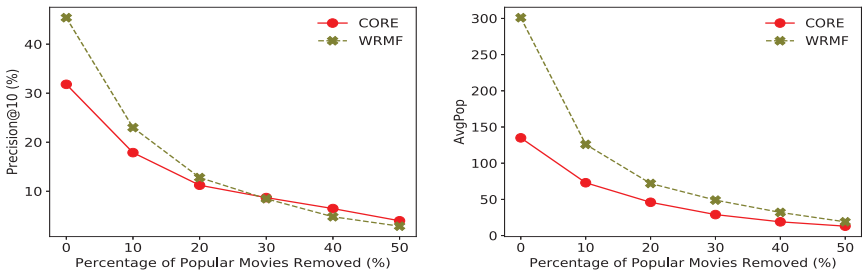


Figure 2.8: CORE vs. WRMF on different distributions of item popularity.

Table 2.7: Descriptions of adjusted data sets.

	Rating	#Users	#Items	#Ratings	Sparsity(%)	Avg.#Ratings per item	% of Ratings from Top 20% Items	% of Ratings from Top 50% Items	Gini Coefficient
`MovieLens`	[1,5]	943	1682	100,000	93.70	59	65%	93%	0.63
`MovieLens-10%`	[1,5]	943	1512	56,960	96.00	38	57%	90%	0.57
`MovieLens-20%`	[1,5]	939	1344	35,181	97.21	26	54%	89%	0.54
`MovieLens-30%`	[1,5]	919	1175	21768	97.98	19	52%	87%	0.52
`MovieLens-40%`	[1,5]	866	1004	12,866	98.52	13	49%	84%	0.50
`MovieLens-50%`	[1,5]	756	838	7,218	98.86	9	48%	83%	0.46
`Last.fm`	[1,5]	635	4100	14,175	99.46	5	53%	70%	0.57

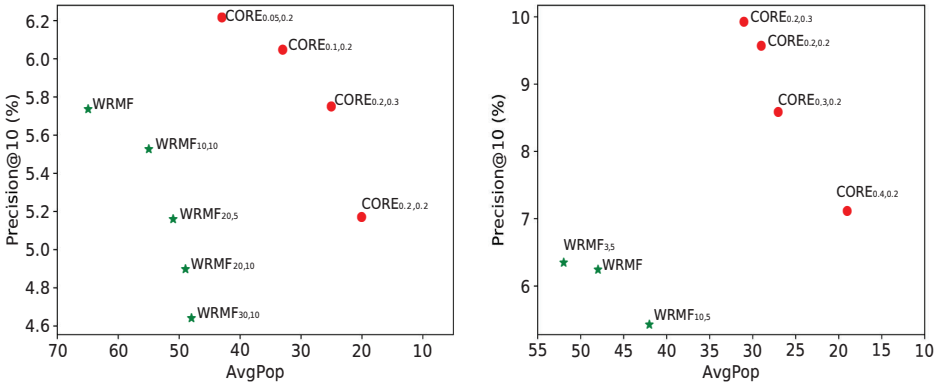
Fig. 2.8 shows the precision (accuracy) and AvgPop (long-tail recommendation performance) of WRMF vs. CORE on the original `MovieLens` data and five new datasets generated from it. In particular, as the rating distribution of `MovieLens` moves more toward niche items, the accuracy gap between WRMF and CORE gradually narrows. Eventually, e.g., in settings where 40% and 50% popular movies are removed, CORE actually starts to outperform WRMF in terms of accuracy. Meanwhile, the long-tail recommendation performance of CORE remains better (i.e., lower AvgPop of recommended items) than WRMF's, albeit by a smaller margin, as there are much fewer popular items for the WRMF's popularity bias to manifest itself strongly.

In summary, this provides additional evidence that, aside from demonstrating superior long-tail recommendation performance across a wide variety of datasets, on heavier-tailed datasets CORE demonstrates highly competitive performance in terms of accuracy as well.

Flexible Recommendation. An important characteristic of the proposed CORE approach is that it allows to fine-tune the popularity level of recommended items in a flexible manner. In particular, as was mentioned in Section 2.3.2, the types of items that appear in the discovered cosine patterns can be easily tuned by setting cosine and/or support thresholds accordingly.

Figs. 2.9 and 2.10 show the performance of different variations of CORE$_{cos,supp}$, i.e., CORE under different cosine and support thresholds. For a given level of support, the popularity of recommended items tends to go down as the cosine threshold

goes up. For `Book-Crossing` data (Fig. 2.9a), this can be seen from the fact that $\text{AvgPop}(\text{CORE}_{0.05,0.2}) \geq \text{AvgPop}(\text{CORE}_{0.1,0.2}) \geq \text{AvgPop}(\text{CORE}_{0.2,0.2})$. For `Last.fm` (Fig. 2.9b): $\text{AvgPop}(\text{CORE}_{0.2,0.2}) \geq \text{AvgPop}(\text{CORE}_{0.3,0.2}) \geq \text{AvgPop}(\text{CORE}_{0.4,0.2})$. Correspondingly, the ratio of niche recommendations tends to go up with the cosine threshold as well, as shown in Fig. 2.10. The intuition is that a higher cosine threshold filters out more cross-support patterns containing high-frequency (i.e., popular) items. Alternatively, for a given level of cosine, the popularity of recommended items tends to go up as the support threshold goes up. E.g., in Fig. 2.9a we can see $\text{AvgPop}(\text{CORE}_{0.2,0.2}) \leq \text{AvgPop}(\text{CORE}_{0.2,0.3})$ for `Book-Crossing`; in Fig. 2.9b we have $\text{AvgPop}(\text{CORE}_{0.2,0.2}) \leq \text{AvgPop}(\text{CORE}_{0.2,0.3})$ for `Last.fm`. The ratio of recommended niche items also correspondingly goes down. With higher support threshold, only items in comparatively more frequent patterns, i.e., more frequently consumed items, would be recommended. Not surprisingly, fine-tuning CORE for even better long-tail recom-

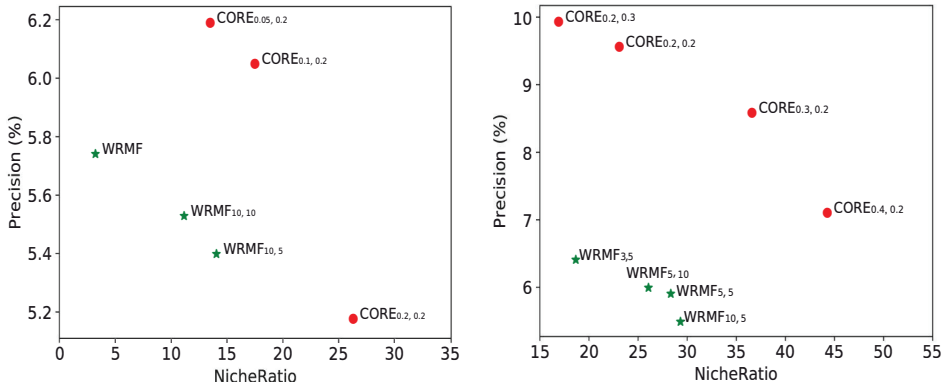


(a) `Book-Crossing` (b) `Last.fm`

Figure 2.9: Two dimensional (*Precision-AvgPop*) performance comparison.

mendation performance may come at the expense of some recommendation accuracy, as Figs. 2.9 and 2.10 show. Therefore, in real-world applications, cosine and support thresholds should be set by the domain experts keeping in mind the specific application

requirements, such as the desired mix of popular and niche item recommendations, various levels of recommendation popularity for users with different popularity preferences, and the trade-offs between accuracy and long-tail performance.



(a) `Book-Crossing` (b) `Last.fm`

Figure 2.10: Two dimensional (*Precision-NicheRatio*) performance comparison.

Comparisons with the Long-Tail-Oriented Baseline. Our main experiments demonstrated the benefits of CORE as compared to multiple popular, general-purpose baseline algorithms. The benefits are especially prominent on the heavier-tailed datasets, where CORE shows advantages not only in long-tail performance, but in accuracy performance as well. Here we conduct an additional experiment to show that CORE's performance advantages on heavier-tailed data remain even when compared to a *specialized* long-tail-oriented baseline. Specifically, we compare CORE with the long-tail recommendation strategy proposed by Park and Tuzhilin (2008). We chose this approach due to its adaptability to different existing recommendation techniques and its flexibility (somewhat similar to CORE's) for parameterizing the long-tail recommendation performance.

As with our main experiments, we hold out 30% of each users' ratings as the ground truth for evaluation purposes. Again, we chose to use the WRMF baseline in conjunction with the long-tail recommendation strategy due to WRMF's consistently good

performance across different datasets. To adapt WRMF for long-tail recommendation, all items in the training data (i.e., 70% of each user's ratings) are first pre-processed by partitioning them into head (H) and tail (T) groups using different rating frequency thresholds α (depending on the overall item frequency in a given dataset). Items with the number of ratings greater than α would be in group H, the rest of the items in group T. Those in group T are further clustered into β clusters as proposed in Park and Tuzhilin (2008). The above process guarantees a more balanced item rating distribution within each group and, thus, alleviates the concern of recommendation bias towards highly popular items. In particular, we set $\alpha \in \{10, 20, 30\}$ for `Book-Crossing`, $\alpha \in \{3, 5, 10\}$ for `Last.fm`, and $\beta \in \{5, 10\}$ for both datasets. Based on the pre-processing, recommendations are first generated using WRMF within H and each of β groups within T and then aggregated to form the final top-K recommendation list. Figs. 2.9 and 2.10 show the comparison between different variations of $\text{CORE}_{cos,supp}$ (i.e., CORE under different cosine and support thresholds, as discussed earlier) and $\text{WRMF}_{\alpha,\beta}$ (i.e., WRMF under different settings of pre-processing).

Note that, although we run a number of different CORE and WRMF variations, for clarity of visualization in Figs. 2.9 and 2.10 we display only the performance "frontiers" both for CORE and WRMF. In other words, we do not display CORE and WRMF variations where the recommendation performance is dominated by some other variations in both dimensions, i.e., *both* in recommendation accuracy and long-tail performance. First, the results verify the effectiveness of the approach proposed by Park and Tuzhilin (2008) for boosting long-tail recommendation of WRMF. Second (and, again, not surprisingly), better long-tail recommendation performance often comes at the expense of recommendation accuracy, which applies for both CORE and long-tail-oriented WRMF and CORE. And, finally and importantly, CORE still provides better performance in

terms of both accuracy and long-tail recommendation, as the performance frontier of CORE dominates the one of the long-tail-oriented WRMF.

Benefits of Hybridization with CORE. As discussed earlier, CORE provides clear performance advantages on heavier-tailed data (both in long-tail and accuracy performance dimensions). However, on other kinds of datasets, such as `MovieLens`, no method strongly dominates others on both dimensions. For example, as was shown in Figs. 2.7c and 2.7f, while CORE still exhibits superior long-tail recommendation performance, it is WRMF that demonstrates best accuracy (although underperforming significantly in terms of long-tail recommendation). In such cases, it is possible to obtain advantages on both dimensions by developing a *hybrid* (or ensemble) recommender system.

As an example, we can hybridize WRMF and CORE in different ways by taking top-i recommended items from CORE and top-$(10-i)$ recommended items from WRMF and merging them into the final top-10 list. For any given $i \in \{0, 1, \ldots, 10\}$, we denote the resulting hybrid method H_i. The two-dimensional performance comparison of original WRMF (i.e., H_0), CORE (i.e., H_{10}), and all hybrid methods (i.e., H_1, H_2, ..., H_9) is shown in Fig. 2.11. For example, consider performance of H_5. The results indicate that, from a simple hybridization strategy of taking top-5 items from WRMF and top-5 items from CORE, H_5 is able to get nearly half of the long-tail performance benefits of CORE (over what original WRMF showed) as well as the vast majority of accuracy benefits of WRMF (over what original CORE showed). This further highlights the practical applicability and value of CORE in achieving different recommendation goals, e.g., providing accurate (popular) recommendations for mainstream users and long-tail recommendations for variety-seeking, idiosyncratic, or contrarian users.

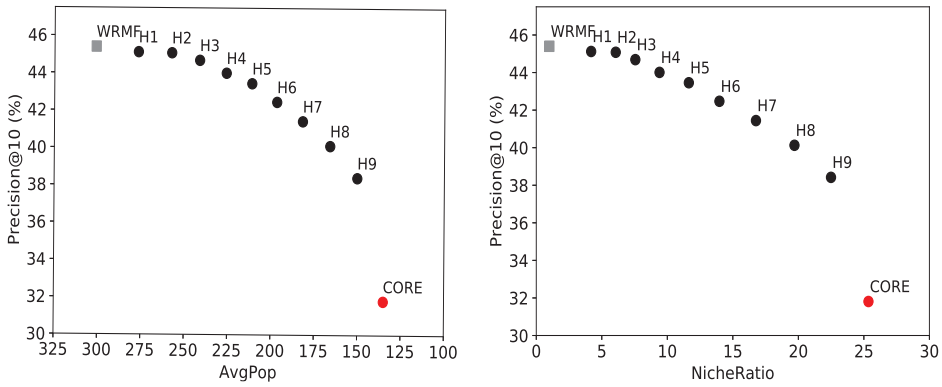


Figure 2.11: Performance of hybrid CORE and WRMF approaches on MovieLens data.

2.4.5 Scalability Demo: CORE for Hashtag Recommendation

In this section, we demonstrate scalability of CORE by applying it on the large-scale application of hashtag recommendation on a social media platform.

Social media platforms like Twitter or Instagram provide opportunities for instant information sharing and diffusion. However, the ease of posting encourages and facilitates rapid content generation, which can lead to information overload for the users. To deal with this issue, some platforms encourage their users to create and cite *hashtags*, i.e., relevant keywords or phrases (that start with hashtag symbol #) to indicate the themes of their posts. For example, a post with hashtag *#SuperBowl* can be easily associated with other Super-Bowl-related posts for future search, recommendation, or analysis. With the accumulation of huge numbers of hashtags over time, analyzing data of user-hashtag engagement and providing personalized hashtag recommendations in real time has become an important function of social media platforms, which facilitates discovery of relevant content and encourages further engagement.

To build and evaluate CORE for hashtag recommendation, we collected data on user engagement with different hashtags from Sina Weibo, a Twitter-like platform in China. Specifically, we collected all tweets that were posted by users over a period of

one month and extracted all hashtags that were ever used. This resulted in a dataset containing 172,981,649 observations (user-hashtag interactions) from 1,629,504 users on 46,281 different hashtags (with extreme sparsity of 99.77%) and over half billion tweets posted by those users. For each user, 70% of all the hashtags used were randomly chosen as the training set and the remaining 30% as the test set.

We had two different implementations of CORE for this recommendation task: (i) we applied CORE directly on the entire user population, and (ii) we partitioned the users into several more homogeneous sub-populations and applied CORE separately on each sub-population. The latter approach is a popular practice in large-scale recommender systems that can often result in better recommendations due to the use of data from more relevant user population (i.e., population with similar interests, tastes, and behavior). Clustering of users was done by first applying LDA (Blei et al. 2003), a well-known topic modeling technique, to all tweets to discover a set of different topics users discussed. Each user would then be represented as a probability distribution over discovered topics (i.e., their preferences to different topics). Based on users' topic preferences, K-means clustering was used to segment users into 10 different clusters, which we denote as C0, C1, ..., C9. The number of clusters was determined according to a commonly used evaluation procedure, i.e., elbow criterion with the sum-of-squared-errors metric. The size of each cluster (i.e., number of users) is shown in Table 2.8.

Our experiments showed a substantial accuracy gain of using CORE on partitioned user population. Specifically, applying CORE on the entire user population resulted in precision-in-top-10 of 31.8%, whereas the performance increased to 37.5% in the partitioned population setting. Therefore, we focus our discussion on the latter implementation in the remainder of this section.

We summarize the computational efficiency comparison of the straightforward (CORE), CP-tree-based (CORE+), and parallel CP-tree-based (CORE++) implementations of

Table 2.8: Straightforward vs. CP-tree based recommendation efficiency.

	Total	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
Size	1,629,504	477,811	170,369	166,566	112,961	23,539	31,388	28,300	61,494	376,222	180,854
T_{CORE} (s)	2,069,134	872,005	183,828	136,917	313,354	74,713	69,430	143,906	79,389	82,016	113,576
T_{CORE+} (s)	21,049	5,439	2,758	3,500	1,219	366	229	894	270	1,036	5,338
T_{CORE++} (s)	2390	590	304	393	124	40	27	103	36	188	585
T_{CORE}/T_{CORE+}	98	160	67	39	257	204	303	161	294	79	21
pattern-based LBP T_{CORE+}/T_{CORE++}	8.8	9.2	9.1	8.9	9.8	9.2	8.4	8.7	7.5	5.5	9.1
support-based LBP T_{CORE+}/T_{CORE++}	8.4	8.3	8.9	8.7	8.7	8.4	8.3	9.3	6.8	6.2	8.5

CORE for hashtag recommendation in Table 2.8. Specifically, T_{CORE}, T_{CORE+}, and T_{CORE++} denote the total amount of time (in seconds) that was needed to generate recommendations. The table also displays the *speedup* ratio of using CP-tree (i.e., T_{CORE}/T_{CORE+}) as an indicator of efficiency gain. While time consumed varies with the cluster size and the number of patterns discovered within each cluster, CORE+ is nearly 100 times faster on average. More specifically, on average, it takes about $1270ms$ to generate recommendations for a single user using the straightforward implementation, while it takes only about $13ms$ using CORE+. Finally, the last two rows of Table 2.8 demonstrate the additional efficiency gain due to parallelization, by comparing runtimes of CORE+ and CORE++, where CORE++ is implemented as parallel CORE+ with 12 computational nodes. In summary, CORE++ is 8-9 times faster than CORE+, and it takes only about $1.5ms$ on average for CORE++ to generate recommendations for a single user. Note that, for CORE++, two different load-balanced partitioning (LBP) strategies (as proposed in Section 2.3.4) were compared, and pattern-based load indicator yielded slightly better speedup than the support-based one. These scalability illustrations further emphasize the practicality and real-time capabilities of CORE.

2.5 Conclusions

With the increasing adoption and growth of digital content provision, recommender systems have become an indispensable component of various online platforms, as they

help users to find relevant products or services from a vast array of choices more efficiently via personalized recommendations. Such systems have been reported to have significant impact on product sales and users' consumption behaviors. In particular, many traditional recommendation approaches (e.g., collaborative filtering approaches) exhibit substantial *popularity bias*, i.e., recommendations tend to direct users' attention largely towards popular products. At the same time, it is widely acknowledged that long-tail recommendations can also be valuable, both for consumers and providers, as they better satisfy heterogeneous consumer needs and can lead to increases in demand, engagement, and loyalty. For certain business models (e.g., for streaming service platforms), recommending more niche content to users could also reduce content licensing costs and, thus, lead to higher provider surplus.

However, due to the highly skewed distribution of consumed items and the fact that user preferences for more idiosyncratic and less popular items are harder to predict, accurate recommendation of long-tail items remains a significant challenge. In this paper, we propose CORE, a cosine-pattern-based technique, for effective long-tail recommendation. The proposed approach has two key components. First, it uses a special type of item co-occurrence patterns, called cosine patterns, that are mined from consumers' consumption histories and, as discussed in the paper, turn out to be highly advantageous for recommendation purposes, especially for long-tail, niche items. Second, it generates personalized recommendations by matching each user's consumption history against the discovered patterns. To ensure scalability of the proposed approach, we design a CP-tree structure for efficient recommendation generation (CORE+) and can further employ a parallel recommendation framework (CORE++) to facilitate real-time recommendation.

In our experimental studies, we observe that cosine patterns indeed demonstrate the advantages of discovering more cohesive relationships among items, including niche

items. The proposed cosine-pattern-based approach (CORE) is tested on three public datasets from different application domains, and we compare it to two types of baseline algorithms – pattern-based and collaborative-filtering-based – in terms of accuracy and long-tail recommendation performance. The results show that CORE dominates all baselines in terms of long-tail recommendation, while being highly competitive in terms of recommendation accuracy. In particular, on heavier-tailed datasets, CORE consistently demonstrates accuracy performance that on par with the highest-performing baselines. On other datasets, in addition to its superior long-tail performance, CORE offers straightforward "hybridization" opportunities for combining it with traditional top-accuracy baselines to achieve combined benefits in both accuracy and long-tail performance. Finally, in addition to its high explainability, which is common to most pattern-based recommendation approaches, CORE demonstrates high flexibility, which provides the system designers with the ability to fine-tune the system (i.e., using different thresholds for support and cosine metrics) towards the desired popularity of recommended items, as well as high scalability, which enables to facilitate real-time recommendations for a given user (i.e., in a matter of milliseconds) in large-scale recommendation applications.

This study also provides several directions for future research. One such direction would be to further our understanding regarding the impact of dataset characteristics on the performance of cosine-pattern-based recommendation method. As shown in the paper, skewness in item popularity and consumption has an impact on the cosine patterns that are discovered and, hence, on the performance of the proposed algorithm. Developing a deeper mathematical understanding of the role that specific dataset characteristics play in the cosine-pattern-based recommendation performance would allow to further improve the effectiveness of pattern-based recommendation algorithms. Another promising research direction would be to move beyond pattern-based methods and

to use advantages of the cosine metric more directly in the recommendation generation process, e.g., perhaps as part of a rating-prediction or learning-to-rank approach based on supervised machine learning methodologies. More specifically, the current approach uses the cosine metric to learn item associations (patterns) first, and then generates recommendations based on patterns. Bypassing the intermediate step of pattern generation and designing more direct methods of using cosine information may lead to additional performance benefits. And, finally, conducting user studies to obtain further insights on users' interactions with (and acceptance of) cosine-based recommender systems constitutes another interesting direction for future work.

Chapter 3

Essay 2: Improving Reliability Estimation for Individual Numeric Predictions

3.1 Introduction and Motivation

Many critical decisions in real world rely on predictions, e.g., investors forecast returns, doctors diagnose diseases, producers predict sales. Facilitated by continuous improvements in data processing and storage technologies, this has spurred development and improvement of machine learning and, more generally, predictive modeling techniques. However, these automated predictions are often imperfect because they are made from noisy, limited data or using simplified computational or probabilistic reasoning.

For numeric prediction tasks, predictive models focus primarily on providing *individual* prediction outcomes; for example, a diabetes risk estimation model would output the risk score of diabetes for each potential patient. Meanwhile, the quality of predictive models is commonly evaluated using *aggregate* prediction accuracy metrics, such as

mean absolute error or root mean squared error, calculated on some test set of data. The issue of *individual prediction reliability* (IPR), i.e., the magnitude of error or level of uncertainty of any *specific* individual prediction, has not been explored as comprehensively. When applying properly trained models, i.e., models with best possible aggregate accuracy, to real-world data, the ability to provide reliability estimation for any specific prediction is undoubtedly important, especially for the purpose of facilitating decision support. As an example, let's assume that, when estimating the severity of Parkinson's disease for two individual patients using Parkinson's Disease Rating Scale (Tsanas et al. 2009), both patients are predicted to have the same rating score of 123, i.e., the same predicted disease severity. At the same time, the prediction reliability could be highly different for numerous reasons, e.g., because these two patients belong to highly different age groups for which different amounts of data are available. For example, it is possible that the prediction of 123 for a younger patient means that the true disease rating value likely is 123 ± 30 (i.e., between 93 and 153), while the same prediction for an older patient might be much more reliable, i.e., 123 ± 5. The diagnosis reliability information is important for deciding on individualized treatment, yet is not captured by the predicted outcome (i.e., 123) alone. In general, knowledge of prediction reliability provides a more nuanced understanding of predictive model performance and can be critical in many real-world numeric prediction applications, especially in highly risk-sensitive domains like pharmaceutical research, medical diagnosis, or financial markets.

As a simple illustration of the research context, consider the stylized, synthetically generated data [1] in Fig. 3.1, where X axis represents the input variable and Y axis represents the outcome to be predicted. Specifically, the black dots represent data points (x, y), and the solid red line represents the estimated linear regression model

[1] 2000 points were created by generating their x values uniformly at random from $[-2, 2]$, and their corresponding y values were generated using function $y = 2.5x + \epsilon$, ϵ $N(0, \sigma^2)$. In particular, $\sigma = 2$ for $x \in [-2, -0.5] \cup [0.5, 2]$; $\sigma = 10$ for $x \in [-0.5, 0.5]$.

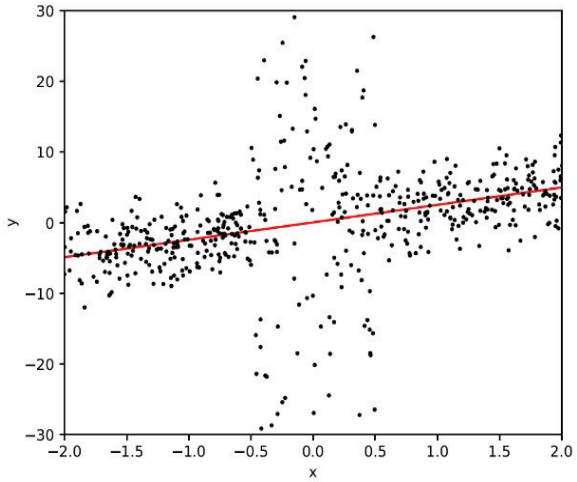


Figure 3.1: Synthetic Data Example for Prediction Reliability Issue.
(X/Y axis: input/outcome variable. Dots: data points. Solid line: predictive model based on linear regression.)

$y = \hat{f}(x)$ that is used for prediction. Although the linear regression model represents the most accurate predictive model for this dataset (as this dataset was generated with this purpose in mind), it is easy to see that the predictions for $x \in [-0.5, 0.5]$ are much less reliable in a given setting, i.e., prediction errors $e = |y - \hat{y}| = |y - f(x)|$ for individual data points in this area are typically much higher than for $x \notin [-0.5, 0.5]$.

In other cases, where the true data generating process cannot be accurately recovered, the prediction errors can result not only from the random noise, which typically leads to the variance of outcome predictions, but also from the misfit of the models which leads to systematic bias of outcome predictions (Domingos 2000, Geman et al. 1992). It is also important to reiterate that individualized prediction reliability estimates are not captured by traditional aggregate accuracy metrics. By design, the goal of individual prediction reliability is not to be another metric that needs to be balanced together (e.g., as part of the machine learning loss function) with the overall model accuracy, but rather to provide diagnostic information to decision makers who use a given outcome prediction model, i.e., providing not only the model's prediction for a

given input, but also the indication of how reliable each specific prediction is expected to be.

It should be mentioned that prediction reliability has been referred to in different ways in previous literature: prediction risk, prediction uncertainty, prediction confidence, etc. We draw on (Bosnić and Kononenko 2008a) to use *prediction reliability* for the sake of terminological consistency. Reliability estimation has been used for two main purposes. One line of research uses estimated reliability as an additional criterion (e.g., in conjunction with accuracy-based metrics) for model evaluation, where models with higher prediction reliability are typically more preferred. Different methods have been proposed for estimating prediction reliability for this purpose, e.g., cross-validation, bootstrapping, Bregman divergence, covariance-based (Efron 2004, Shao 1996). Similar to aggregate accuracy metrics mentioned before, reliability estimated in this type of work is still used as an aggregate model evaluation tool. The other line of research uses prediction reliability for individual prediction explanation or description, which is directly aligned with the focus of this paper. Those studies fall into three finer-grained groups based on the type of outcome to be predicted, i.e., reliability for a single example in classification, probability estimation, or numeric prediction. In this paper, we focus specifically on reliability of numeric prediction models (as will be discussed in the next section), which has been significantly underexplored in research literature, as compared to reliability estimation for other types of outcomes. For example, reliability of probability estimation is often measured by Brier score (Brier 1950) which is calculated as the squared difference between actual outcome (binary or categorical) and predicted probability assigned to that outcome. There have been numerous studies investigating individual classification reliability. For some classifiers, like logistic regression or naïve Bayes (Hand and Yu 2001, Walker and Duncan 1967), the posterior probability of an individual predicted class can be viewed as confidence (reliability) of its prediction.

Most related studies propose more general (model-agnostic) approaches, e.g., transductive reliability estimation (Kukar and Kononenko 2002, Tzikas et al. 2007) drawing on transduction based confidence estimation (Ho and Wechsler 2003, Proedrou et al. 2002, Saunders et al. 1999) or the *typicalness* framework (Melluish et al. 2001, Nouretdinov et al. 2001).

Even though the reliability estimation for numeric prediction models has been significantly underexplored in research literature, it undoubtedly represents an increasingly important issue due to the needs for more fine-grained understanding of predictive model performance, as will be discussed in next section. Therefore, going beyond the evaluation of the overall (i.e., aggregate) accuracy performance of numeric prediction models, in this study we focus on providing a *general-purpose*, data-driven approach to *individual prediction reliability* (**IPR**)[2] estimation. In particular, we propose to use a simple IPR indicator based on expected *absolute prediction errors*. This is motivated by an observation that the performance of existing IPR estimators are usually evaluated by how well they are aligned with actual errors of outcome prediction models. As a result, an indicator based on the actual prediction errors can provide a more direct measurement of IPR, i.e., high estimated error indicates low IPR (see Section 3.3.1 for a detailed description). On the whole, such an indicator has the benefits of being intuitive and providing highly interpretable information to the decision makers as well as allowing for more precise evaluation of reliability estimation quality. Even more importantly, the proposed IPR indicator also allows us to reframe reliability estimation itself as a canonical numeric prediction problem (of the absolute prediction error). Specifically, estimating the proposed IPR indicator is equivalent to learning a numeric prediction model where targets are actual absolute prediction errors and then applying the model for error prediction given any new observations. The error prediction model aims at

[2]We use acronym IPR to refer to "individual prediction reliability" throughout the paper.

capturing the relationship between individual inputs and their prediction errors. The reframing of the problem makes the proposed approach general-purpose (i.e., can work in conjunction with any outcome prediction model), alleviates the need for any statistical/distributional assumptions, and enables the use of advanced, state-of-the-art machine learning techniques to learn IPR patterns directly from data. Advantages of the proposed approach are demonstrated using comprehensive computational experiments on seven real-world datasets and in comparison to multiple techniques from prior work. Error estimation performance measured by two different evaluation metrics show that the proposed machine-learning-based approaches, especially ensemble-based methods like XGBoost and Random Forest, provide significantly better reliability estimation over various baselines, especially in more complex application settings (e.g., datasets with more input features).

3.2 Related Work

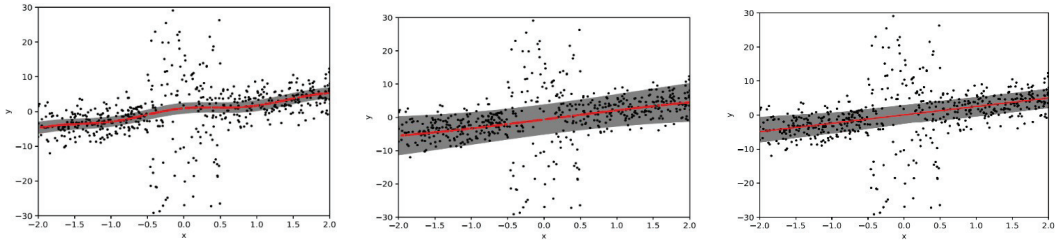
Given the popularity of (and reliance on) predictive modeling techniques in many aspects of everyday life, in general a more comprehensive and nuanced understanding of predictive model performance represents an increasingly important issue. Ability to provide IPR estimates is an important aspect for both application and interpretation of predictive models (Briesemeister et al. 2012, Bosnić and Kononenko 2008a, Shrestha and Solomatine 2006). In particular, for a given predictive model, IPR estimates would provide better understanding for which data points the model is expected to perform better vs. worse (i.e., have higher vs. lower reliability). This connects well to the topic of *error analysis*, which helps to find opportunities for substantial increase in predictive performance. For example, in biomedical informatics, the error models of individual cells can discern new subpopulations within complex mixtures of cells and derive more robust measures for cell classification (Kharchenko et al. 2014). In medical diagnosis,

analyzing inaccurate predictions are important to find out what cases can confuse machine learning models even when the overall predictive performance is impressive (Choi et al. 2017). In biological natural language processing (Hakala et al. 2013), analyzing inaccurate predictions helps diagnosing whether false predictions of the event type (e.g., gene expression, transcription, etc.) is due to missing or incorrectly constructed features. In speech recognition (Qian et al. 2018), error analysis is used to identify top types of errors (substitution, deletion, etc.) that the system makes under different noise contexts, which is valuable in informing prediction application as well as system adaptation. In online recommender systems, examining rating prediction errors at individual level can inform designing of meta-learning algorithms for different users or user groups (Collins et al. 2018). IPR estimates are also relevant to the important research topic of algorithmic bias (Datta et al. 2015, Simoiu et al. 2016, Hosanagar 2019, Johndrow et al. 2019), as they could provide early detection signals of potential systematic bias of predictive models. Finally, as mentioned earlier, IPR provides extra information, which is important for facilitating better decision making across a broad array of applications in chemical and pharmaceutical research (Briesemeister et al. 2012, Liu et al. 2018, Toplak et al. 2014, Cortés-Ciriano and Bender 2018), financial markets (Dash et al. 2015, Huang et al. 2018, Solares et al. 2019), medical diagnosis (Lebedev et al. 2014, Iorio et al. 2015, Tomassetti et al. 2016), and many others.

In terms of methodologies for the IPR representation and calculation, traditional approaches could be summarized into two broad categories: (i) *distribution-based*, i.e., estimating an entire distribution of the outcome variable predictions for any given input value x, which can then be provided to the decision makers directly or in some aggregate form (such as confidence interval) as information about prediction reliability or confidence, and (ii) *indicator-based*, i.e., providing a simple, single-numeric-value-based indicator of IPR for given x, often based on some heuristic.

Distributional, or confidence-interval-based (Wonnacott and Wonnacott 1990), approaches are rooted in statistical properties of prediction models, especially regression models, and represent an intuitive way to indicating IPR – predictions with wider confidence intervals (for a given confidence level) indicate higher model uncertainty. Distributional approaches also tend to be model-specific, i.e., designed specifically for a particular outcome prediction model, and rely on certain statistical assumptions. In both least-squares-based and likelihood-based learning of regression models, generation of confidence intervals or other confidence metrics is based on the assumption of independent and identical distribution of errors across the input space (i.e., homoscedasticity) (Halperin 1963, Knafl et al. 1985). However, this homoscedasticity assumption is usually violated in many real-world settings which is explicitly the focus of this study (reflecting situations similar to the one illustrated in Fig. 3.1, and thus, the derived confidence intervals would fail to reflect actual IPR. More sophisticated regression-based distributional approaches draw on the flexible Gaussian process (Rasmussen 2003), which allows to incorporate information on similarity between data points into the model building. Although the probabilistic Gaussian process regression model facilitates the derivation of predictive distribution for the regression outcome, the key characteristic of this modeling technique is that the variance of the distribution for new observation x (i.e., the indicator of its prediction reliability) *only depends on the input features of* x and, in particular, on the relative location (e.g., distance calculated using feature values) of x to other observations in the training data, and *not on the observed target (outcome) values* (Rasmussen 2003). Because of the latter fact, it is unlikely to capture the magnitude of error in the prediction that is due to variability in the outcomes, which makes it a less informative measurement of IPR. Fig. 3.2a emphasizes this by presenting the 95% prediction intervals of Gaussian process regression learned from the synthetic dataset used in Fig. 3.1 – the widths of individual prediction intervals are similar across the input

(x) space, not reflecting the actual variability in the outcomes. There have been other distributional approaches that extend certain specific learning techniques to make predictions together with corresponding probabilistic reliability estimates (Khosravi et al. 2010, Papadopoulos et al. 2001). For example, Hwang and Ding (1997) use an asymptotic approach to build confidence intervals for neural networks; however, similarly to what has been discussed above, due to traditional statistical assumptions on errors (e.g., homoscedasticity) and model parameters, the prediction intervals generated by this approach are not designed to reflect the variability in the actual outcomes (but rather the variability in model predictions), as illustrated in Figure 3.2b on the same stylized dataset.



(a) `Gaussian Process-based` (b) `Neural Networks-specific` (c) `Bootstrapping-based`

Figure 3.2: IPR Representation Based on 95% Confidence / Prediction Interval
(X/Y axis: input/outcome variable. Dots: data points. Solid line: model based on different techniques.)

Another standard approach to construct prediction distributions and corresponding prediction intervals is bootstrapping (Efron 1992). This approach has some clear benefits in that it can be used with all kinds of predictive models (i.e., it is not model-specific) and generate distributions without relying on statistical assumptions; however, heteroscedasticity still poses a significant challenge for bootstrapping-derived IPR representation in certain situations. To illustrate, in Fig. 3.2c we plot confidence intervals obtained from this approach on data presented in Fig. 3.1 [3]. Due to the similarity of

[3]Each bootstrap sample was generated from original training data by randomly sampling (with replacement) 500 data points, on which a linear regression model was learned and then applied to make predictions for test data. We repeat this process 100 times – resulting in 100 predictions for each data point, from which 95% prediction interval is empirically constructed.

data patterns in the bootstrap samples, the predictions of *all* linear models across the entire input space are similar. This means that, across the entire range of input values (x), the width of point-wise confidence intervals derived from the prediction variance would be similar too and not reflective of the actual underlying variability of data, as indicated in Fig. 2c.

To summarize, the existing distributional approaches (many of which are model-specific and rely on restrictive statistical assumptions) have been designed mainly to reflect the distribution of model predictions and, therefore, are less well-suited for capturing the actual underlying variability of ground truth data (and, hence, actual prediction errors), i.e., for capturing IPR in heteroscedastic environments. As an alternative, a number of prior studies addressed this issue by turning to simpler, yet more flexible, indicator-based approaches to IPR estimation, which we discuss next.

Indicator-based approaches typically represent general-purpose (i.e., applicable with any outcome prediction model, free from statistical assumptions) IPR estimators that provide a simple numeric value as an indicator of IPR for any individual input value. Among these approaches, early work focused on using nonparametric bootstrapping techniques (Carney et al. 1999, Heskes 1997) and summarizing the individual prediction variability across samples (e.g., by using confidence/prediction interval widths or prediction variance) as reliability indicators, or estimating errors based on the covariance among data points (Efron 2004). Several other methods are based on heuristics that try to exploit local information of individual data points in order to directly capture the actual variability of underlying data, e.g., using prediction errors (Briesemeister et al. 2012), prediction variance of the nearest neighbors of the focal data point (Clark 2009), or the density of the input space in close proximity to the focal data point (Bosnić and Kononenko 2008a), as surrogates of IPR. These approaches are based on intuition that the uncertainty of individual predictions should be higher around data points with high

prediction errors or high prediction variance, or for points around which there is not much training data available. One can see that some of these heuristics – in particular, density-based – would not be very useful in heteroscedastic settings (such as the one illustrated by Fig. 3.1). Somewhat similarly, Shrestha and Solomatine (2006) propose to partition the input space into different clusters and then construct prediction intervals based on the empirical distributions of the errors associated with instances in the same cluster. In terms of specific IPR indicators, (Briesemeister et al. 2012) designed two statistics based on the local properties of training data, while another related study (Bosnić and Kononenko 2008b) proposed several empirical measures based on sensitivity analysis.

For our computational experiments, we use nine commonly used indicator-based reliability approaches as baselines for comparison: VarBag (Breiman 1996), VarA, MSE (Briesemeister et al. 2012), VarP, AvgDiff (Bosnić and Kononenko 2008a), AvgDist (Sheridan et al. 2004), LCV (Demut 2010), SAV and SAB (Bosnić and Kononenko 2008b). The relevant notation and the formal definitions of these approaches are provided in Tables 3.1 and 3.2, respectively; note that all measures are calculated for a given individual data point (x, y), where x represents an input feature vector and y is an outcome (target) value. We narrowed down our choice to this particular set of approaches as most promising baseline candidates due to their potential flexibility for capturing IPR in heteroscedastic environments and for their advantageous performance reported in prior studies and observed in our pilot experiments.

Finally, evaluation is necessary to test and compare the effectiveness of different methods for IPR estimation. As observed in prior literature, for an IPR indicator to be meaningful and useful, the estimates that it produces should be "aligned" with actual individual prediction errors; i.e., predictions estimated to be more reliable should exhibit smaller errors (and vice versa). Based on this intuition, previous studies typically use

Table 3.1: Common Notations for Describing Reliability Estimation Methods

Symbol	Definition
x	input vector (of different features) of a given example
y	outcome value of a given example
x_i	input vector of ith nearest neighbor in heuristic-based methods
y_i	actual outcome of ith nearest neighbor in heuristic-based methods
$\hat{y}_i$	predicted outcome of ith nearest neighbor in heuristic-based methods
$\hat{\epsilon}_i$	$\epsilon_i = y_i - \hat{y}_i$,prediction error of ith nearest neighbor in heuristic-based methods
m	number of random samples in bootstrapping-based methods
M_j	prediction for x made by the model learned from the jth sample in bootstrapping-based methods
n	number of nearest neighbors selected in heuristic-based methods
$d(x_i, x)$	distance between the example x and its ith nearest neighbor in heuristic-based methods
$\hat{y}_{-i}$	leave-one-out prediction of ith nearest neighbor in heuristic-based methods
τ	sensitivity parameters ($\tau \in [0, 1]$)
S	set of sensitivity parameters. An example of $S = 0.01, 0.1, 0.5, 1$
t_{max}/t_{min}	Maximum/minimum value of outcome in the training data
$\hat{y}_\tau$	predicted outcome of x using models trained using training data (X, Y) plus augmented sample of $(x, y + \tau * (t_{max} - t_{min}))$ in sensitivity based methods
$\hat{y}_{-\tau}$	predicted outcome of x using models trained using training data (X, Y) plus augmented sample of $(x, y - \tau * (t_{max} - t_{min}))$ in sensitivity based methods

Table 3.2: Description of Baseline Reliability Estimation Methods

Baseline	Calculation and Description
VarBag	$\frac{1}{m}\sum_{j=1}^{m}(M_j - \hat{y})^2$, $\hat{y} = \frac{\sum_{j=1}^{m} M_j}{m}$.Variance of example x's predictions M_js made by models learned from different random samples.
VarA	$\frac{1}{n}\sum_{i=1}^{n}(\bar{y} - y_i)^2$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.Variance of example x's nearest neighbors' actual values (y_is).
VarP	$\frac{1}{n}\sum_{i=1}^{n}(\bar{\hat{y}} - \hat{y}_i)^2$, $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$.Variance of example x's nearest neighbors' predictions ($\hat{y}_i$s).
AvgDiff	$\lvert\frac{\sum_{i=1}^{n} y_i}{n} - \hat{y}\rvert$. Difference between the average of nearest neighbors' actual values (y_is) and the example x's prediction ()
MSE	$\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i)^2$,$\epsilon_i = y_i - \hat{y}_i$. Mean squared error of example x's nearest neighbors' predictions ($\hat{y}_i$s)
AvgDist	$\frac{1}{n}\sum_{i=1}^{n} d(x_i, x)$. Average distance between the example x and its nearest neighbors (x_is).
LCV	$\frac{\sum_{i=1}^{n} d(x_i,x) * E_i}{\sum_{i=1}^{n} d(x_i,x)}$, $E_i = \lvert y_i - \hat{y}_{-i}\rvert$. Weighted average errors of nearest neighbors' leave-one-out predictions ($\hat{y}_{-i}$).
SAV	$\frac{\sum_{\tau \in S}(\hat{y}_\tau - \hat{y}_{-\tau})}{\lvert S\rvert}$. Average difference between sensitivity predictions $\hat{y}_\tau$ and $\hat{y}_{-\tau}$ over different sensitivity parameters in set S.
SAB	$\frac{\sum_{\tau \in S}(\hat{y}_\tau - \hat{y}) + (\hat{y}_{-\tau} - \hat{y})}{2 * \lvert S\rvert}$. Average difference between sensitivity predictions ($\hat{y}_\tau$ or $\hat{y}_{-\tau}$) and original prediction ($\hat{y}$).

the correlation coefficient (between the reliability estimates and actual prediction errors) as the measure of "alignment" to evaluate the performance of proposed IPR indicators for numeric prediction models (Bosnić and Kononenko 2009, Briesemeister et al. 2012), where higher correlation indicates better IPR estimation performance.

In summary, the general structure and contribution of many existing individual prediction reliability estimation studies can be outlined as: (i) defining some reliability indicator; (ii) demonstrating how it can be computed/derived; and (iii) showing its quality by showing that its values are well "aligned" with the actual outcome prediction errors (using some "alignment" measure, typically correlation coefficient). Our study follows the same general structure to provide further improvements to the current state of the art in this area, as discussed in the next section.

3.3 Machine Learning Approach to Individual Prediction Reliability Estimation

3.3.1 Individual Prediction Reliability Indicator: General Overview

In this study, we propose a novel indicator-based approach to IPR representation and calculation. The key motivation for the proposed method was the observation that, while the existing IPR indicators have been defined in a variety of different ways (e.g., as variance or density of certain data, etc.), their performance is always judged by how well the IPR estimates are aligned with actual errors of the outcome prediction model. Therefore, we propose to use a simple and intuitive reliability indicator that is designed to be directly related to errors of the outcome prediction model, i.e., an indicator based on *expected absolute prediction error* for a given individual prediction.

More specifically, prediction uncertainty can come from different sources that are often hard to disentangle, e.g., random noise/variability, inappropriate model selection, or suboptimal model parameters. The actual prediction errors, i.e., the discrepancies between observed y and prediction $\hat{y}$ from some predictive model, provide the most reliable signals of the level of prediction uncertainty. Higher prediction error typically indicates lower IPR, and prediction errors could arguably be used in at least two distinct ways under different contexts.

In particular, one could use *absolute* prediction errors, i.e., $e = |\hat{y} - y|$, vs. *direct* prediction errors, i.e., $e = \hat{y} - y$; the latter would reflect not only the absolute magnitude of discrepancy, but also its *direction*, in other words, whether y is overestimated or underestimated by the outcome prediction model. In this study, we focus on the absolute-error-based IPR indicator for the following key reason. The situations where the model's prediction errors are highly imbalanced (model under-predicts and over-predicts in numerous portions of the data space) typically reflect the fact that the

outcome prediction model poorly represents the underlying generative process of the data (i.e., the model is biased, poorly fit), and the first goal typically is to improve the overall model fit. These situations often can be readily diagnosed with standard, traditional aggregate model evaluation metrics; of course, such situations could be remedied by looking at direct errors as well, and there are entire machine learning approaches dedicated to that[4]. However, once the outcome prediction model fit is improved using detected systematic direct errors (no significant over- or under-prediction), the remaining patterns of direct errors would be impossible to learn (essentially being random noise of different magnitudes), yet the key IPR problem as stated in the paper would still be highly relevant (as motivated by Fig. 3.1). And, insightful and actionable IPR information can still be mined from data.

Thus, abstracting away from the directionality of errors, we propose to view the reliability of a given individual prediction as the expected *absolute* prediction error. As mentioned earlier, this allows us to address the IPR estimation problem as a canonical, *meta-algorithmic* numeric prediction problem, i.e., it can to use any advanced machine learning technique for reliability estimation. More specifically, IPR represented by absolute prediction error, i.e., $e = |y - \hat{y}|$, could be directly modeled as a function of input variables x, i.e., as $e = F(x)$, to capture the structural relationships between the input space and the prediction reliability for any given outcome prediction model. Building machine learning model F (i.e., the reliability estimator) does require labeled training data $\{(x, e)\}$. It is important to point out that this data usually is readily available, because the outcome prediction models (i.e., models predicting) are typically evaluated

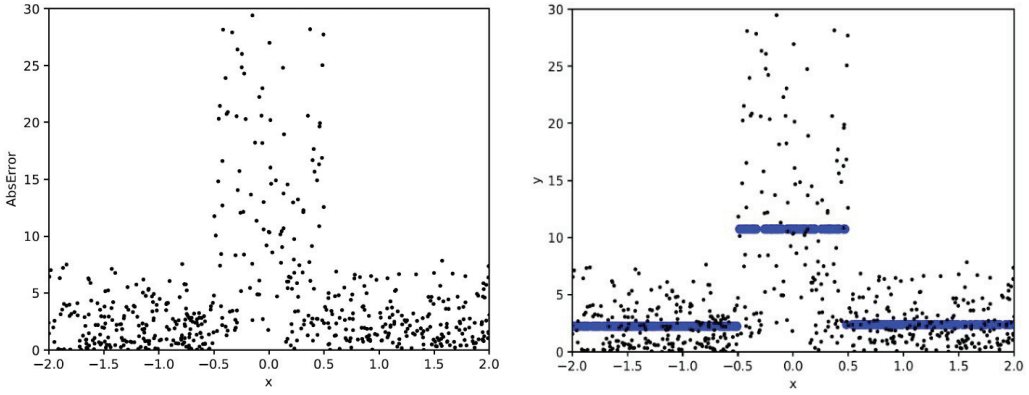
[4]In cases where prediction errors are not balanced (i.e., when the outcome prediction model is significantly biased), there are substantial opportunities to improve the outcome predictive models themselves first, before performing reliability estimation. In fact, some boosting-based machine learning techniques, e.g., XGBoost (Chen and Guestrin 2016), use this idea: they build an ensemble of models sequentially one-by-one and take advantage of the unbalanced direct prediction errors from models learned in earlier stages iteratively to improve the ultimate outcome prediction performance of the entire ensemble model.

on some *hold-out test data* $\{(x, y)\}$ which can then be straightforwardly reused to construct the ground truth for reliability estimation; i.e., every data point (x, y) together with corresponding outcome prediction $\hat{y}$ can be converted to (x, e), where $e = |\hat{y} - y|$.

Taking the data shown in Fig. 3.1 as an example, the absolute prediction error (and, hence, the IPR) of the best outcome prediction model is consistently higher in certain areas. This is illustrated in Fig. 3.3a, where x axis still represents the input features, while vertical axis now represents absolute error (i.e., e) of the outcome prediction model. As the figure shows, the absolute prediction error is much higher within interval $x \in [-0.5, 0.5]$ than elsewhere, which can be learned by machine learning techniques. For example, using the data plotted in Fig. 3.3a, a regression tree model can learn to predict e from x, and we show the pointwise prediction errors estimated from this regression tree in Fig. 3.3b. Each blue dot in Fig. 3.3b represents an estimated absolute prediction error for given x, which shows that the prediction of the errors, i.e., the IPR indicators $(\hat{e})$, are quite informative. As can be seen in Fig. 3.3b, estimated reliability is able to accurately differentiate the levels of outcome model's prediction uncertainty across different intervals, i.e., the uncertainty is lower for $x \in [-2.0, -0.5]$ and $x \in [0.5, 2.0]$ and higher for $x \in [-0.5, 0.5]$.

It is important to reiterate that reframing IPR estimation as a data-driven numeric prediction problem makes the proposed approach *general-purpose* (i.e., reliability estimation can be done for any outcome prediction model) and alleviates the need for distributional modeling assumptions. An added benefit of the proposed IPR indicator is its clear *interpretability* to end-users and decision makers, which may not be the case with some existing approaches that require probabilistic assumptions (e.g., distribution-based approaches) and non-intuitive quantifications (e.g., density-based heuristic indicators). Specifically, the reliability score of a given prediction simply represents the

expected absolute error for this prediction, along the lines of "for given x, the outcome prediction model is expected to be off by this much, on average".



(a) `Actual absolute prediction error` (b) `Regression-tree-based error estimation`

Figure 3.3: Pointwise Prediction Error Estimation of Linear Regression Model from Fig. 3.1
(X axis: input variable. Y axis: absolute prediction error.
Black dots: actual abs. prediction error. Blue dots: estimated abs. prediction error.)

Finally, the proposed approach also allows for a more precise and informative evaluation. As mentioned earlier, a popular reliability evaluation metric has been the *correlation* between IPR values and actual prediction errors. Even though correlation coefficient is not a very precise measure in the sense that it captures only very high-level patterns (general trends), it has been widely used largely because the existing IPR indicators have been defined in highly differing ways (as variance, density, or average of certain data, etc.) – a more direct comparison to actual prediction errors was not feasible. In other words, even with high correlation, it is possible that the magnitude of IPR estimates might be significantly different than the one of the actual errors, thus, reducing diagnosticity (or interpretability) of IPR indicators. In contrast, the proposed IPR indicator (i.e., expected absolute prediction error) is, by design, "on the same scale" as the ground truth (i.e., actual absolute prediction error) against which the reliability is judged. This allows for an even more precise performance measurement (going well

beyond correlation coefficient), e.g., using canonical numeric accuracy measures such as *root mean squared error* (RMSE).

3.3.2 Estimating and Evaluating the Proposed Reliability Indicator: ML-Based Framework

In this subsection, we more formally describe the details of machine-learning-based framework, which can be used for estimating and evaluating the proposed absolute-prediction-error-based IPR indicator. We also use this framework for the computational experiments in our study.

As a quick summary, the proposed framework follows a two-stage process. Because IPR estimation is done for some given outcome prediction model, the overarching goal of Stage 1 is to use the outcome prediction model (build it first, if necessary) and produce the data about its errors, which is typically achieved by deploying the outcome prediction model on a representative, hold-out data sample (i.e., test data). This data then serves as the ground truth for Stage 2, where the actual IPR estimation is done – i.e., the actual absolute prediction errors from Stage 1 are used as outcome variables to build an error prediction model using best machine learning practices. The overview of the entire framework is depicted in Fig. 3.4.

In terms of data, as depicted in the first row of Fig. 3.4 and as is typical in predictive modeling, we assume the existence of two datasets (often based on a random split of an underlying database with known outcome values) $\{(x_o, y_o)\}$ and $\{(x_{test}, y_{test})\}$, where the former is used for outcome prediction model learning and the latter for outcome model evaluation. In both datasets, for each data point, x represents the input feature vector, and y represents the corresponding outcome variable. In Stage 1, given $\{(x_o, y_o)\}$, outcome prediction model f is built using any desired numeric prediction modeling technique, e.g., Neural Network, Regression Tree, Random Forest, etc., as is

shown in the second row of Fig. 3.4 [5]. This is a standard model learning process where the best machine-learning practices and procedures, such as cross-validation method (Kohavi et al. 1995, Picard and Cook 1984), can be used to properly build and fine-tune outcome prediction models. In our experiments (discussed later in the paper), we use multiple different machine learning techniques to explore the effectiveness of the proposed prediction reliability approach in conjunction with various outcome prediction models.

After model f is trained, naturally it can be deployed for outcome prediction purposes, i.e., to make outcome predictions for any input x as $f(x)$. Stage 1 concludes by deploying f on outcome evaluation data $\{(x_{test}, y_{test})\}$, i.e., prediction for each observation (x_{test}, y_{test}) is constructed as $\hat{y}_{test} = f(x_{test})$, and corresponding (absolute) prediction error is derived as $e = |\hat{y}_{test} - y_{test}|$. This newly generated data e – the set of actual prediction errors of f on the outcome evaluation dataset – has traditionally been used for the final, authoritative evaluation of model f performance on hold-out data. However, it also carries specific information about the performance on individual predictions (i.e., actual errors) by model f and, thus, we use this information in the form of labeled error learning dataset (x_{test}, e) in Stage 2 for building models for IPR estimation.

Stage 2 represents our proposed machine-learning approach to IPR estimation. As discussed earlier, we propose to use machine-learning techniques to estimate absolute prediction errors (representing IPR) of any given outcome prediction model directly as a function of input features x. The ground truth labels for this machine learning task are obtained from Stage 1, which results in the labeled error learning dataset

[5] In our study, during Stage 1, we use outcome prediction models built using *machine learning* techniques, as is done in many advanced real-world applications. However, *any* outcome-predicting model can be used in this framework, e.g., an already existing rule-based expert system or some black-box approach which may not require separate outcome learning data at this point.

$\{(x_{test}, e)\}$, as discussed above and shown in Fig. 3.4. Following standard machine-learning practices, dataset $\{(x_{test}, e)\}$ is randomly split into training dataset $\{(x_t, e_t)\}$ and validation dataset $\{(x_v, e_v)\}$. Based on training dataset $\{(x_t, e_t)\}$, to encapsulate the underlying relationships between input feature vector x_t and the absolute error e_t, error prediction model f_e is constructed as $\hat{e}_t = f_e(x_t)$, where $\hat{e}_t$ denotes the model prediction. Model f_e could be produced by any available machine learning technique using best model-building and fine-tuning practices (e.g., cross-validation), and the best choice of the technique ultimately will depend on the context of each specific prediction problem, e.g., complexity of underlying relationships in data, availability of the data, etc. In our experiments, we use numerous machine learning techniques to explore their performance under different contexts.

Once error prediction model f_e is trained, it can be deployed for reliability estimation purposes, and Stage 2 concludes by deploying f_e to error validation data $\{(x_v, e_v)\}$, as shown in Fig. 3.4. In particular, the actual absolute prediction error e_v of outcome prediction model f for any data point (x_v, y_v), i.e., $e_v = |\hat{y}_v - y_v|$, would be estimated by f_e as $\hat{e}_v = f_e(x_v)$. In other words, $f_e(x_v)$ represents the proposed IPR indicator for the corresponding individual outcome prediction $f(x_v)$, for any input x_v. In summary, based on the proposed approach and framework, *for any given input x, the two models (f and f_e) would be able to provide the essential prediction-related information: the outcome prediction as $f(x)$ and the estimated reliability of this prediction as $f_e(x)$.*

Finally, error validation dataset $\{(x_v, e_v)\}$ can also be used to properly evaluate the performance of error prediction model f_e, as this data has been used to build neither outcome prediction nor IPR estimation models. Thus, as shown in Fig. 3.4, the final evaluation of reliability estimation performance is done by comparing the obtained IPR estimates $\{\hat{e}_v\}$ with actual prediction errors $\{e_v\}$. As discussed earlier, for an IPR indicator to be meaningful, the IPR estimates ideally should be "aligned" with actual errors

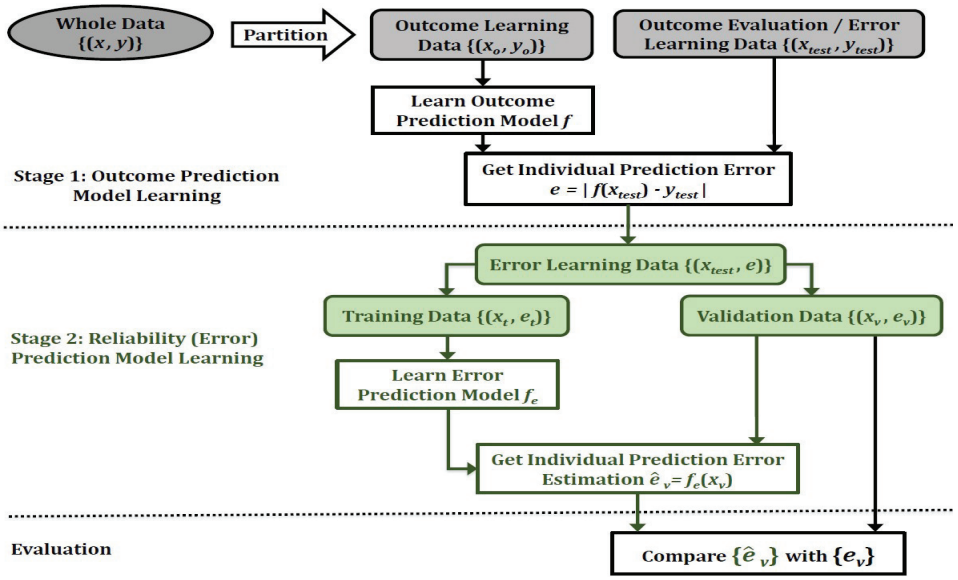


Figure 3.4: Two-Stage Machine-Learning-Based Framework for Reliability Estimation

of the outcome prediction model; thus, one relevant and widely used IPR estimation performance metric is correlation coefficient, i.e., corr($\{e_v\}$,$\{\hat{e}_v\}$); a higher correlation value indicates better performance. Importantly, as the proposed IPR indicator has been designed to be "on the same scale" as the ground truth, it allows to use more precise numeric prediction accuracy measures as well, such as root mean squared error, i.e., $RMSE(\{e_v\},\{\hat{e}_v\}) = \sqrt{\sum_v (e_v - \hat{e}_v)^2 / |\{e_v\}|}$, where a lower RMSE value indicates better performance.

3.4 Experiments

3.4.1 Experimental Setup

We demonstrate the effectiveness of our approach through comprehensive computational experiments following the general two-stage framework described in Section 3.3.2 and Fig. 3.4.

Seven public data sets from UCI Machine Learning Repository [6](as summarized in Table 3.3) are used to test the performance of the proposed approach. Selected data sets vary by application domain, size (number of records), and complexity (number of input attributes). For Stage 1, each data set is randomly split into two parts, i.e., outcome learning data and outcome evaluation data, with percentages of 40% and 60%, respectively. For Stage 2, the latter part is used as error learning data and is further randomly split into equal-sized (i.e., 50%-50%) error training and error validation datasets. Note that we do the performance evaluation 30 times for each dataset (by generating a different random split into outcome learning, error learning, and error validation datasets). All results are based on the average performance of the 30 runs, and all techniques (ML-based and baselines) were evaluated on the same evaluation data within each run.

For Stage 1, i.e., to build outcome prediction models, we chose seven machine learning techniques widely used for predicting numeric outcomes, i.e., KNN (k nearest neighbors), NN (neural network), LR (linear regression), RT (regression tree), RF (random forest), SVR (support vector regression), and XGB (extreme gradient boosting). The use of different predictive modeling techniques highlights the general-purpose applicability of the proposed reliability estimation approach for use in conjunction with a wide variety of outcome prediction models. For Stage 2, i.e., to build absolute error prediction models, we used the same set of machine learning techniques to explore whether some of them might be more advantageous for the reliability estimation task.

To benchmark the proposed approach, we use nine baseline algorithms for comparison (summarized in Table 3.2): one bootstrapping-based and eight heuristic-based reliability estimators. For parameter setting, specifically, in our experiments, we set $m = 20$ (number of random samples generated from original data in bootstrapping) and $n = 20$ (number of nearest neighbors chosen to calculate those heuristic-based baselines).

[6]https://archive.ics.uci.edu/ml/datasets.html

Table 3.3: Overview of Data Sets Used in Computational Experiments

Data Set	#Obs	#Attributes	Output description	Output range
Power Plant	9568	4	Hourly electrical energy output.	[420,495]
ISE	536	7	Istanbul Stock Exchange 100 index.	[-8.5,10]
Housing	506	13	Value of houses in $1000s.	[6,50]
Bike Rental	17389	16	Daily count of rental bikes.	[1,977]
Parkinsons	5875	26	Parkinson's disease symptom score.	[7,29]
Posts Comments	40949	54	Log of number of Facebook posts comments.	[0,8]
News Popularity	39797	58	Log of number of total shares of news.	[3,13]

The reliability estimation performance of different approaches is evaluated using both correlation-based and predictive-accuracy-based metrics, i.e., correlation coefficient and RMSE; as mentioned earlier, higher correlation and lower discrepancy between actual and estimated prediction errors indicate better reliability estimation.

For expositional completeness, we first present the predictive accuracy measured by RMSE of different outcome prediction models (i.e., models built in Stage 1), as shown in Table 3.4, where each row and column represents each outcome prediction model and dataset, respectively. Note that, for each technique, the results represent the best performance of each model achieved after optimizing model parameters, e.g., the number of nearest neighbors in KNN, depth of the tree in RT and RF, number of neurons and hidden layers in NN, number of estimators and size of subsample in XGB, length scale and gamma parameters of kernel functions in SVR, and many other parameters. Best performance, i.e., lowest RMSE, on each dataset is highlighted in bold, which shows that XGB generally tends to perform well on different data sets, followed by RF and NN. The one exception is the simple ISE dataset, where arguably the simplest model – linear regression – is sufficient to capture predictive relationships in the data.

The next two subsections discuss the results of IPR estimation (i.e., Stage 2) experiments.

Table 3.4: Predictive Accuracy (RMSE) of Different Outcome Prediction Models.
(Average performance based on 30 runs; best performance on each data set is shown in bold.)

Model	Power Plant	ISE	Housing	Bike Rental	Parkinsons	Posts Comments	News Popularity
KNN	3.985	1.548	6.040	0.896	0.329	0.701	0.882
LR	4.578	**1.404**	5.600	1.072	0.378	0.814	0.874
NN	4.285	1.431	4.914	0.431	0.364	0.647	0.869
RF	3.775	1.524	4.643	0.369	0.277	0.509	0.864
RT	4.415	1.688	5.696	0.501	0.382	0.555	0.888
SVR	4.535	1.526	7.095	0.906	0.400	0.634	0.889
XGB	**3.574**	1.514	**4.601**	**0.341**	**0.226**	**0.493**	**0.852**

Table 3.5: Comparison of Reliability Estimation Performance (Correlation Coefficient)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better)

Outcome Prediction Model	KNN		LR		NN		RF		RT		SVR		XGB	
Prediction Reliability Estimator	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML
Power Plant	0.23	**0.29*** **	0.23	**0.41*** **	0.18	**0.37*** **	0.19	**0.23*** **	0.16	**0.31*** **	0.23	**0.42*** **	0.18	**0.21***
ISE	**0.27*** **	0.12	**0.34*** **	0.16	**0.32*** **	0.13	**0.28*** **	0.17	**0.32*** **	0.23	**0.32*** **	0.18	**0.30*** **	0.16
Housing	0.40	**0.54*** **	0.36	**0.51*** **	0.42	**0.46*** **	0.42	**0.43**	0.39	**0.41**	0.45	**0.64*** **	0.44	**0.45**
Bike Rental	0.52	**0.75*** **	0.46	**0.87*** **	0.31	**0.50*** **	0.38	**0.51*** **	0.32	**0.48*** **	0.52	**0.84*** **	0.38	**0.50*** **
Parkinsons	0.56	**0.64*** **	0.55	**0.65*** **	0.45	**0.66*** **	0.53	**0.63***	0.62	**0.71*** **	0.57	**0.70*** **	0.45	**0.58*** **
Comments	0.47	**0.64*** **	0.50	**0.68*** **	0.40	**0.63*** **	0.43	**0.54*** **	0.38	**0.55*** **	0.41	**0.62*** **	0.32	**0.53*** **
News Pop	0.14	**0.23*** **	0.14	**0.24*** **	0.17	**0.25*** **	0.15	**0.26*** **	0.13	**0.24*** **	0.13	**0.23*** **	0.18	**0.26*** **

*significant at 5%;**significant at 1%;***significant at 0.1%

3.4.2 Performance Comparison Based on Correlation

We first focus on the performance comparisons in terms of correlation coefficient – the widely used metric for evaluating IPR indicators, as discussed earlier. We first show the detailed performance of each machine-learning-based method for our proposed absolute-error-based IPR indicator as well as each baseline method, and then compare the effectiveness of these two classes of methods using summarized results.

In particular, Table 3.6 compares the reliability estimation performance among seven machine learning techniques. The bold numbers represent best performance for a given outcome prediction model in terms correlation coefficient. A closer look at the results shows that the XGB approach produced the best (or near best) reliability performance among the ML-based approaches. Specifically, in the majority (33 out of 49) of settings that were explored XGB outperforms other techniques, followed closely by RF which performs the best in the rest (15 out of 49) of the settings. An interesting pattern observed from the results is that RF is better than or competitive with XGB only on

Table 3.6: Performance of Machine-Learning-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	**KNN**	**LR**	**NN**	**RF**	**RT**	**SVR**	**XGB**
Power Plant	KNN	0.253	0.336	0.304	0.215	0.240	0.337	0.202
	LR	0.071	0.089	0.104	0.074	0.077	0.121	0.062
	NN	0.094	0.194	0.156	0.099	0.105	0.201	0.094
	RF	**0.291**	**0.413**	**0.374**	0.219	**0.311**	**0.416**	0.191
	RT	0.059	0.182	0.137	0.068	0.097	0.175	0.046
	SVR	0.160	0.286	0.271	0.157	0.185	0.265	0.155
	XGB	0.278	0.407	0.373	**0.229**	0.301	0.415	**0.210**
ISE	KNN	0.122	0.099	0.089	0.135	0.182	0.150	0.123
	LR	0.015	0.026	0.020	0.040	0.000	0.000	0.015
	NN	0.043	0.052	0.047	0.107	0.115	0.080	0.046
	RF	**0.123**	**0.163**	**0.129**	**0.173**	**0.231**	**0.183**	0.160
	RT	0.045	0.075	0.112	0.062	0.157	0.097	**0.199**
	SVR	0.058	0.035	0.100	0.028	0.103	0.083	0.023
	XGB	0.110	0.100	0.066	0.133	0.199	0.141	0.111
Housing	KNN	0.419	0.384	0.333	0.327	0.339	0.454	0.306
	LR	0.381	0.326	0.340	0.308	0.293	0.442	0.326
	NN	0.490	0.484	0.357	0.310	0.321	0.635	0.341
	RF	**0.538**	0.500	0.457	**0.433**	**0.424**	**0.638**	0.440
	RT	0.398	0.409	0.360	0.320	0.298	0.518	0.350
	SVR	0.301	0.318	0.279	0.253	0.264	0.398	0.227
	XGB	0.522	**0.505**	**0.459**	0.431	0.417	0.636	**0.453**
Bike Rental	KNN	0.523	0.490	0.392	0.460	0.408	0.563	0.447
	LR	0.460	0.382	0.342	0.398	0.349	0.527	0.396
	NN	0.557	0.567	0.345	0.405	0.357	0.617	0.400
	RF	0.742	0.858	0.482	0.499	0.462	0.840	0.488
	RT	0.684	0.798	0.365	0.410	0.361	0.801	0.419
	SVR	0.479	0.441	0.370	0.443	0.383	0.541	0.434
	XGB	**0.748**	**0.865**	**0.494**	**0.505**	**0.480**	**0.844**	**0.498**
Parkinsons	KNN	0.508	0.478	0.476	0.484	0.538	0.560	0.419
	LR	0.267	0.205	0.235	0.257	0.279	0.285	0.237
	NN	0.285	0.232	0.255	0.277	0.307	0.345	0.226
	RF	0.496	0.377	0.430	0.498	0.456	0.404	0.401
	RT	0.240	0.209	0.239	0.283	0.254	0.248	0.242
	SVR	0.424	0.223	0.255	0.246	0.196	0.428	0.283
	XGB	**0.637**	**0.653**	**0.664**	**0.625**	**0.714**	**0.697**	**0.575**
Comments	KNN	0.528	0.519	0.478	0.428	0.438	0.495	0.420
	LR	0.489	0.556	0.449	0.374	0.395	0.456	0.368
	NN	0.539	0.599	0.451	0.460	0.459	0.506	0.458
	RF	0.634	0.672	0.622	0.532	0.541	0.611	0.525
	RT	0.574	0.589	0.566	0.496	0.512	0.551	0.491
	SVR	0.549	0.526	0.492	0.435	0.448	0.491	0.427
	XGB	**0.640**	**0.676**	**0.629**	**0.540**	**0.549**	**0.618**	**0.527**
News Pop	KNN	0.185	0.197	0.201	0.204	0.182	0.183	0.209
	LR	0.203	0.229	0.227	0.226	0.200	0.204	0.233
	NN	0.205	0.226	0.232	0.232	0.207	0.206	0.237
	RF	0.213	0.227	0.236	0.243	0.221	0.215	0.246
	RT	0.175	0.181	0.194	0.199	0.179	0.171	0.202
	SVR	0.052	0.175	0.198	0.179	0.179	0.181	0.181
	XGB	**0.229**	**0.242**	**0.251**	**0.258**	**0.236**	**0.233**	**0.262**

data with simpler structure (having fewer input features), e.g., Housing, ISE, and Power Plant; that is, XGB consistently has the edge over all approaches on more complex datasets.

Similarly, in Table 3.7 we show the comparison among nine baseline techniques. Although no one baseline predominantly outperforms others, MSE (heuristic-based) and VarBag (bootstrapping-based) tend to have higher correlation coefficient with actual errors in 20 and 12 (out of 49) settings, respectively.

Another observation is that VarBag is more competitive on data with less complex structure (having fewer input features), e.g., Power Plant, ISE, Housing, and Bike Rental, while on data sets like Parkinsons, Posts Comments, and News Popularity, heuristic-based estimators like MSE and VarA generally perform better.

We summarize the comparison of correlation coefficient results between ML-based reliability estimators and baselines in Table 3.5.

Specifically, we compare the *best* baseline (BL) technique (chosen among the nine baseline techniques discussed earlier) and the *best* machine learning (ML) model (chosen from the seven ML techniques used earlier) in terms of correlation. The bold and red numbers represent significantly higher correlation coefficients from ML-based methods, and the bold and blue numbers represent better results from baselines. The results show that, in 42 out of 49 (85.7%) predictive task configurations in our experiments, the best ML-based estimator is a better IPR indicator (exhibiting higher correlation with actual prediction errors), and in 39 out of these 42 cases the advantage is statistically significant, emphasizing the advantages of using the proposed approach over baselines for IPR estimation.

Also note that ML-based IPR estimators were outperformed by baselines only on the ISE dataset, i.e., the simple dataset (536 observations and 7 input features) with less complex predictive relationships, where a simpler outcome prediction model like

linear regression was sufficient to guarantee high prediction accuracy, as mentioned earlier. In other words, the heuristic-based IPR estimators may be sufficient for simpler datasets; however, more sophisticated approaches are advantageous when more complex predictive settings must be considered.

3.4.3 Performance Comparison Based on RMSE

As discussed in Section 3.3.1, while correlation coefficient is able to capture general variability patterns, it is not designed to reflect the situations where the magnitude of IPR estimates might be significantly different than that of the actual errors, reducing our ability to make more precise judgements about IPR estimation performance. The proposed absolute-prediction-error-based IPR indicator provides for more precise and informative performance evaluation, due to being "on the same scale" as the ground truth, which allows us to bring in standard numeric prediction accuracy measures – specifically, *root mean squared error* (RMSE) – and provide a much clearer picture of true IPR estimation performance, as discussed below. Here we follow the same structure as in the previous subsection, where we first show the detailed performance of each machine-learning-based method for our proposed absolute-error-based reliability indicator, followed by detailed performance of each baseline method, and then compare the effectiveness of these two classes of methods using summarized results.

In Table 3.8, we compare the IPR estimation performance in terms of RMSE among machine learning methods, bold numbers representing best performance for each outcome prediction model. The results show similar patterns as in the correlation coefficient comparisons. In particular, XGB still performs the best, i.e., exhibits lower discrepancy with actual prediction errors, among all machine learning techniques in most cases. Specifically, in 29 out of 49 cases, XGB produces most accurate reliability estimation, followed by RF and KNN which perform better in the rest 14 and 6 cases, respectively.

Also, detailed results show that RF and KNN tend to outperform XGB on datasets with fewer input features, i.e., Power Plant, ISE, and Housing, while XGB is more advantageous on more complex datasets.

As mentioned earlier, RMSE should be calculated when IPR indicator and actual prediction error are on the same scale. However, this is not the case for heuristic-based IPR indicators, and computing RMSE based on raw values of heuristic-based indicators would put them at a significant disadvantage in terms of their performance comparison with the proposed approach.

Therefore, we take a broader view of the heuristic-based indicators by observing that some of them are calculated by aggregating (e.g., as variance) a certain set of discrepancies (errors), and we aggregated these discrepancies by averaging their absolute values to provide the best-effort estimation of an absolute prediction error. In particular, only five (i.e., VarBag, VarA, VarP, AvgDiff, MSE) out of nine baseline IPR indicators could be converted to estimates of absolute prediction errors (i.e., VarBag.AE, VarA.AE, VarP.AE, AvgDiff.AE, MSE.AE) and, thus, could be used for RMSE comparisons. The other baselines are heuristics that provide a numeric index indicating the degree of prediction reliability but have no direct connection to prediction errors.

As a result, in Table 3.10 we provide the comparison of reliability estimation performance in terms of RMSE among the four aforementioned baselines. As with performance comparisons based on correlation coefficient, no single heuristic-based indicator dominates all others, but MSE.AE and VarA.AE provide best performance in 24 and 21 (out of 49) settings, respectively.

Finally, we summarize the comparison of error estimation accuracy (RMSE) between ML-based estimators and baselines in Table 3.9. Similar to Table 3.5, we compare RMSE of the best baseline (BL) technique chosen among the four baselines discussed earlier and the best machine learning (ML) model of the seven machine learning techniques

used earlier. Significantly lower RMSEs from machine learning based methods are highlighted in bold and red, while significantly lower RMSEs from baselines are highlighted in bold and blue. The results show that, in 40 out of 49 (81.6%) predictive settings in our experiments, ML approaches constitute better IPR indicators, i.e., exhibit lower discrepancy with actual prediction errors as measured by RMSE. Furthermore, in 35 out of 49 cases, best ML-based IPR indicators provide statistically significantly better performance than the heuristic approaches. In contrast, only in 1 out of 49 settings, baselines were statistically significantly better than ML-based approaches. Even on the simpler ISE dataset (where heuristic-based approaches demonstrated better correlation performance), with a more precise performance evaluation using RMSE no statistically significant performance differences are observed between ML-based approaches and baselines. In aggregate, all the experimental results indicate substantial advantages of using machine learning techniques to estimate IPR.

3.5 Conclusions

Estimating individual prediction reliability (IPR) is important for both interpretation and application of predictive models. Going beyond global prediction performance, it gives a finer-grained evaluation for predictive models. In particular, by providing extra information on the potential error of individual predictions, it gives practitioners more confidence in making decisions, which can be helpful in a variety of application domains. For example, reliability of a certain disease prediction can help healthcare providers better evaluate a patient's health condition and make a customized treatment plan. Similarly, individual prediction reliability of a stock return can better inform an investor of the investment risk and, thus, facilitate rational financial decision making. Or, in recommendation applications, when deciding between several highly recommended items, the final selection can be informed by which of these recommendations is estimated

to be more reliable. Moreover, even when the outcome prediction model is relatively accurate in general, it may be important to know that, under some circumstances, some predictions objectively are expected to be less reliable than others. Such knowledge not only can provide signals of potential bias of the prediction models, but also point to how model prediction performance can be improved, e.g., by collecting more data related to less reliable predictions. More generally, IPR can be used as part of the criteria for identifying informative data points as candidates for subsequent actions, e.g., identifying most reliable predictions for decision making or least reliable predictions for error analysis, additional data collection, and model refinement.

While the awareness of how reliable the specific individual predictions are can be important in many complex real-world numeric predictive modeling applications, this issue has been under-explored in research literature. In this study, we propose to estimate IPR for any given numerical outcome prediction model by using machine learning techniques. Specifically, we reconceptualize the reliability estimation problem to a numeric prediction problem by proposing to use *absolute prediction error* as a simple IPR indicator due to its merits of higher interpretability and easy evaluation. The study also describes a general-purpose framework for implementing the proposed reliability estimation approach, which takes can take advantage of any state-of-the-art machine learning methods to directly learn the relationships between input features of a given data point and absolute prediction errors (i.e., reliability indicators) obtained from the outcome prediction model. In addition to providing an intuitive reliability indicator, the proposed machine-learning-based approach is *general-purpose* (i.e., reliability estimation can be done for *any* outcome prediction model), reduces the need for statistical modeling assumptions that some distributional approaches require, and allows for more precise and informative performance evaluation.

The general-purpose framework was also used in comprehensive computational experiments designed to test the proposed approach. Specifically, we observed that machine learning methods can significantly improve IPR estimation, especially in more complex settings, i.e., on datasets that are larger both in the number of examples and input features. We compared the proposed approach with numerous heuristic approaches used in prior work on seven different public datasets based on two different evaluation metrics. The performance advantages of the proposed machine-learning-based approach (over heuristic-based indicators) can be observed across different outcome prediction models, which further emphasizes the generality of the proposed approach.

In addition to introducing a machine-learning-based approach to estimating IPR and demonstrating its effectiveness, this study provides a number of directions for future research. One such direction would be to understand the impact of *dataset characteristics* on the performance of simpler (heuristic-based) vs. more complex (machine-learning-based) reliability estimators. Another direction would be to explore the impact of different *sources* of prediction uncertainty, e.g., whether low reliability of an individual prediction is due to noisy data, model misfit, etc. Revisiting the possibilities of designing additional, more sophisticated and accurate reliability indicators of different types (indicator-based vs. distribution-based) and levels of applicability (general-purpose vs. building specifically on the strengths of some specific outcome prediction model) also represent important direction for follow-up investigations. Advancing our understanding of these issues should not only make reliability estimation increasingly relevant and valuable in real-world predictive modeling applications, but should also lead to deeper, more significant developments of reliability estimation theory.

Table 3.7: Performance of Heuristic-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	KNN	LR	NN	RF	RT	SVR	XGB
Power Plant	VarBag	**0.232**	0.030	0.018	**0.187**	0.102	0.123	**0.184**
	VarA	0.173	0.104	0.139	0.133	0.136	0.150	0.117
	VarP	0.058	-0.001	0.002	0.096	0.070	0.025	0.091
	AvgDiff	0.025	0.023	0.015	0.014	0.044	0.009	0.018
	MSE	0.187	**0.231**	0.173	0.124	**0.160**	**0.225**	0.098
	AvgDist	0.028	0.002	-0.007	-0.010	0.018	0.045	-0.010
	LCV	-0.026	-0.045	0.141	-0.015	-0.027	-0.016	-0.009
	SAV	-0.003	0.012	0.132	-0.006	0.024	0.103	-0.078
	SAB	0.017	0.002	**0.180**	-0.001	0.012	0.001	0.111
ISE	VarBag	0.166	**0.342**	**0.321**	0.253	0.250	0.227	**0.296**
	VarA	0.107	0.093	0.084	0.112	0.152	0.100	0.122
	VarP	0.084	0.153	0.132	0.126	0.174	0.167	0.137
	AvgDiff	-0.024	-0.023	-0.001	-0.010	0.024	0.008	0.005
	MSE	0.100	0.075	-0.087	0.066	0.078	0.073	0.071
	AvgDist	**0.273**	0.252	0.261	**0.284**	**0.318**	**0.315**	0.287
	LCV	0.040	0.024	0.094	-0.016	-0.005	0.013	-0.009
	SAV	0.007	0.259	0.007	0.082	0.181	0.311	0.078
	SAB	0.020	-0.064	0.020	0.023	-0.003	0.004	-0.046
Housing	VarBag	**0.396**	0.342	0.404	**0.418**	**0.385**	0.266	**0.437**
	VarA	0.373	0.326	0.349	0.306	0.280	0.397	0.306
	VarP	0.272	0.130	0.329	0.294	0.256	0.108	0.305
	AvgDiff	-0.316	-0.323	**-0.417**	-0.360	-0.268	-0.318	-0.373
	MSE	0.368	**0.353**	0.245	0.303	0.232	**0.452**	0.271
	AvgDist	0.178	0.204	0.130	0.071	0.027	0.174	0.082
	LCV	-0.148	0.006	0.240	-0.061	-0.074	-0.358	-0.033
	SAV	0.007	0.143	0.106	0.012	0.162	0.255	0.122
	SAB	-0.123	-0.037	-0.002	-0.002	0.132	0.053	0.217
Bike Rental	VarBag	0.445	0.052	0.295	0.376	**0.323**	0.299	**0.376**
	VarA	**0.523**	0.461	0.242	0.241	0.180	0.505	0.261
	VarP	0.395	0.038	0.234	0.221	0.177	0.390	0.253
	AvgDiff	0.149	0.049	0.172	0.249	0.161	0.162	0.275
	MSE	0.479	**0.463**	**0.305**	**0.384**	0.265	**0.522**	0.272
	AvgDist	0.052	0.025	0.172	0.233	0.238	0.004	0.202
	LCV	-0.014	0.014	0.035	0.072	0.073	0.078	0.011
	SAV	-0.014	0.022	0.188	-0.002	0.233	0.169	0.225
	SAB	0.052	-0.012	-0.131	-0.021	-0.003	0.046	0.212
Parkinsons	VarBag	0.470	0.053	0.055	0.293	0.022	0.102	0.399
	VarA	0.531	0.413	0.446	0.315	0.423	0.434	0.361
	VarP	0.276	-0.012	0.134	0.149	0.116	0.124	0.263
	AvgDiff	0.120	0.013	-0.026	0.184	0.162	0.013	0.192
	MSE	**0.557**	**0.552**	**0.452**	**0.532**	**0.615**	**0.570**	**0.445**
	AvgDist	-0.085	-0.045	-0.080	-0.048	-0.010	-0.100	-0.069
	LCV	0.112	-0.007	-0.107	0.059	0.164	0.188	0.054
	SAV	-0.002	-0.053	-0.047	-0.157	-0.163	0.132	-0.026
	SAB	-0.082	0.001	-0.043	-0.048	0.014	0.064	0.034
Comments	VarBag	0.384	0.321	0.263	0.310	0.283	0.238	0.309
	VarA	0.470	0.379	**0.404**	0.330	0.353	**0.410**	**0.318**
	VarP	0.346	0.276	0.346	0.292	0.310	0.297	0.279
	AvgDiff	-0.213	-0.130	-0.137	-0.188	-0.201	-0.178	-0.178
	MSE	**0.473**	**0.493**	0.393	0.326	**0.383**	0.379	0.317
	AvgDist	0.230	0.361	0.214	0.159	0.171	0.257	0.149
	LCV	-0.083	0.006	0.043	-0.065	-0.051	-0.02	-0.075
	SAV	-0.045	0.365	0.320	**0.429**	0.251	0.312	0.279
	SAB	-0.095	-0.047	0.227	0.067	0.100	-0.062	0.237
News Pop	VarBag	0.135	0.063	0.050	0.107	0.047	0.077	0.130
	VarA	**0.137**	0.139	0.155	**0.146**	**0.130**	**0.134**	0.150
	VarP	0.125	0.132	0.129	0.111	0.068	0.119	0.125
	AvgDiff	-0.085	-0.116	-0.129	-0.125	-0.106	-0.087	-0.128
	MSE	0.135	**0.142**	0.142	0.128	0.130	0.115	0.103
	AvgDist	0.106	0.125	0.119	0.120	0.106	0.108	0.116
	LCV	-0.016	0.001	**0.172**	-0.007	-0.035	-0.042	0.148
	SAV	-0.005	0.083	0.080	0.143	0.026	0.05	**0.182**
	SAB	0.068	0.008	0.007	0.050	-0.011	0.046	0.095

Table 3.8: Reliability Estimation Performance of Machine-Learning-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	**KNN**	**LR**	**NN**	**RF**	**RT**	**SVR**	**XGB**
Power Plant	KNN	2.606	2.613	2.552	**2.520**	2.806	2.680	**2.407**
	LR	2.680	2.759	2.658	2.566	2.875	2.902	2.446
	NN	2.683	2.728	2.642	2.565	2.862	2.790	2.442
	RF	**2.590**	**2.535**	**2.498**	2.530	**2.756**	2.594	2.417
	RT	2.693	2.739	2.654	2.576	2.890	2.824	2.456
	SVR	2.684	2.672	2.599	2.566	2.858	2.839	2.450
	XGB	2.598	2.550	2.501	2.524	2.768	**2.595**	2.415
ISE	KNN	**1.014**	**0.909**	**0.954**	1.018	1.133	1.043	**1.019**
	LR	1.049	0.931	0.973	1.043	1.179	1.075	1.051
	NN	1.037	0.925	0.962	1.024	1.146	1.053	1.034
	RF	1.033	0.912	0.958	**1.014**	**1.115**	**1.031**	1.020
	RT	1.073	0.942	0.981	1.054	1.161	1.081	1.055
	SVR	1.024	0.918	0.961	1.026	1.146	1.039	1.025
	XGB	1.059	0.940	0.995	1.047	1.150	1.061	1.054
Housing	KNN	4.292	3.750	3.429	3.346	4.023	5.025	3.372
	LR	4.377	3.869	3.433	3.397	4.125	5.011	3.395
	NN	4.076	3.536	3.342	3.367	4.033	4.308	3.324
	RF	**3.924**	**3.477**	**3.206**	**3.200**	**3.906**	**4.179**	**3.164**
	RT	4.408	3.700	3.451	3.395	4.086	4.868	3.276
	SVR	4.503	3.919	3.522	3.425	4.171	5.169	3.527
	XGB	4.046	3.526	3.295	3.275	4.041	4.315	3.259
Bike Rental	KNN	0.532	0.580	0.269	0.247	0.327	0.527	0.228
	LR	0.553	0.615	0.275	0.255	0.335	0.541	0.234
	NN	0.518	0.550	0.275	0.255	0.334	0.501	0.234
	RF	0.417	0.342	0.256	0.240	0.317	0.339	0.222
	RT	0.454	0.402	0.272	0.254	0.334	0.376	0.231
	SVR	0.558	0.604	0.276	0.254	0.338	0.546	0.233
	XGB	**0.414**	**0.335**	**0.254**	**0.240**	**0.314**	**0.337**	**0.221**
Parkinsons	KNN	0.188	0.210	0.205	0.144	0.198	0.217	0.131
	LR	0.207	0.229	0.223	0.159	0.219	0.244	0.140
	NN	0.206	0.228	0.223	0.158	0.217	0.239	0.141
	RF	0.188	0.216	0.207	0.139	0.206	0.234	0.130
	RT	0.207	0.228	0.222	0.157	0.220	0.246	0.140
	SVR	0.201	0.233	0.225	0.160	0.227	0.238	0.139
	XGB	**0.169**	**0.180**	**0.176**	**0.124**	**0.173**	**0.188**	**0.115**
Comments	KNN	0.442	0.526	0.407	0.333	0.362	0.400	0.324
	LR	0.455	0.471	0.415	0.343	0.372	0.408	0.333
	NN	0.443	0.517	0.462	0.329	0.360	0.427	0.318
	RF	0.401	0.459	0.362	0.311	0.338	0.360	0.303
	RT	0.425	0.496	0.382	0.320	0.346	0.382	0.311
	SVR	0.440	0.495	0.412	0.337	0.364	0.393	0.327
	XGB	**0.399**	**0.462**	**0.360**	**0.310**	**0.336**	**0.358**	**0.301**
News Pop	KNN	0.593	0.577	0.574	0.569	0.581	0.614	0.566
	LR	0.591	0.573	0.570	0.566	0.579	0.611	0.563
	NN	0.591	0.574	0.570	0.565	0.578	0.613	0.562
	RF	0.589	0.573	0.569	0.563	0.576	0.609	0.561
	RT	0.594	0.579	0.574	0.569	0.581	0.615	0.566
	SVR	0.604	0.586	0.584	0.575	0.591	0.621	0.575
	XGB	**0.587**	**0.571**	**0.567**	**0.561**	**0.574**	**0.607**	**0.558**

Table 3.9: Comparison of Reliability Estimation Performance (RMSE)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better)

Outcome Prediction Model	KNN		LR		NN		RF		RT		SVR		XGB	
Prediction Reliability Estimator	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML	Best BL	Best ML
Power Plant	2.64	**2.59**	2.65	**2.54****	2.79	**2.50****	2.80	**2.52****	2.86	**2.76****	2.78	**2.60****	2.78	**2.41****
ISE	1.02	**1.01**	0.92	**0.91**	0.96	**0.95**	1.02	1.01	**1.09**	1.12	**1.02**	1.03	1.02	1.02
Housing	4.44	**3.92***	3.80	**3.48***	3.70	**3.21***	3.95	**3.20****	4.56	**3.91***	5.04	**4.18****	4.24	**3.16****
Bike Rental	0.54	**0.41****	0.58	**0.34****	**0.24****	0.25	0.28	**0.24****	0.35	**0.31****	0.54	**0.34****	0.27	**0.22****
Parkinsons	0.17	0.17	0.18	0.18	0.18	0.18	0.14	**0.12***	0.17	0.17	0.31	**0.19****	0.13	**0.12****
Comments	0.45	**0.40****	**0.45**	0.46	0.44	**0.36****	0.35	**0.31****	0.37	**0.34****	0.52	**0.36****	0.34	**0.30****
News Pop	0.60	**0.59****	0.58	**0.57****	0.58	**0.57****	0.57	**0.56****	0.59	**0.57****	0.84	**0.61****	0.57	**0.56****

*significant at 5%;**significant at 1%;***significant at 0.1%

Table 3.10: Reliability Estimation Performance of Heuristic-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	KNN	LR	NN	RF	RT	SVR	XGB
Power Plant	VarBag.AE	3.977	4.569	4.366	3.741	4.078	4.535	3.609
	VarA.AE	2.822	2.877	2.815	2.964	2.967	2.894	2.941
	VarP.AE	2.865	3.149	3.151	**2.798**	3.241	3.310	**2.783**
	AvgDiff.AE	3.301	3.345	3.446	3.072	3.358	3.391	3.260
	MSE.AE	**2.641**	**2.646**	**2.791**	3.038	**2.858**	**2.778**	2.993
ISE	VarBag.AE	2.090	2.087	2.087	2.088	2.089	1.534	2.088
	VarA.AE	**1.021**	0.936	0.961	**1.020**	**1.091**	**1.022**	**1.019**
	VarP.AE	1.185	1.003	1.033	1.099	1.297	1.191	1.081
	AvgDiff.AE	1.031	0.915	1.154	1.164	1.176	1.166	1.170
	MSE.AE	1.031	**0.915**	**0.961**	1.164	1.176	1.066	1.215
Housing	VarBag.AE	6.011	5.693	5.356	4.699	4.981	7.085	4.647
	VarA.AE	**4.436**	**3.804**	4.261	4.557	5.111	**5.036**	4.657
	VarP.AE	4.439	3.859	4.124	4.377	5.184	6.033	4.546
	AvgDiff.AE	4.858	4.026	4.095	4.408	5.263	5.773	4.274
	MSE.AE	4.536	3.833	**3.690**	**3.953**	**4.564**	5.239	**4.238**
Bike Rental	VarBag.AE	0.902	1.072	**0.235**	0.367	0.414	0.902	0.343
	VarA.AE	0.552	**0.578**	0.862	0.915	0.882	0.587	0.921
	VarP.AE	0.605	0.823	0.826	0.895	0.849	0.626	0.895
	AvgDiff.AE	0.758	0.809	0.695	0.739	0.738	0.716	0.734
	MSE.AE	**0.542**	0.584	0.282	**0.302**	**0.361**	**0.538**	**0.278**
Parkinsons	VarBag.AE	0.330	0.378	0.272	0.275	0.381	0.399	0.229
	VarA.AE	**0.170**	**0.182**	**0.181**	0.184	0.175	**0.305**	0.181
	VarP.AE	0.281	0.326	0.251	0.260	0.356	0.390	0.197
	AvgDiff.AE	0.718	0.702	0.702	0.913	0.718	0.692	0.184
	MSE.AE	0.173	0.187	0.210	**0.136**	**0.173**	0.398	**0.131**
Comments	VarBag.AE	0.701	0.863	0.536	0.504	0.525	0.608	0.493
	VarA.AE	**0.446**	**0.452**	0.451	0.491	0.496	0.631	0.495
	VarP.AE	0.521	0.598	0.449	0.459	0.473	**0.515**	**0.462**
	AvgDiff.AE	0.498	0.523	0.452	0.500	0.482	0.575	0.462
	MSE.AE	0.447	0.529	**0.435**	**0.349**	**0.365**	0.629	**0.339**
News Pop	VarBag.AE	0.883	0.872	0.903	0.859	0.878	0.882	0.849
	VarA.AE	**0.601**	**0.582**	**0.579**	**0.574**	**0.586**	0.841	**0.571**
	VarP.AE	0.808	0.744	0.665	0.687	0.751	0.840	0.658
	AvgDiff.AE	0.605	0.733	0.590	0.690	0.723	0.850	0.714
	MSE.AE	0.602	0.585	0.585	0.590	0.589	**0.839**	0.605

Chapter 4

Essay 3: The Role of Physical Stores in the Digital Age

4.1 Introduction

With the increasing growth of e-commerce, the Internet has dramatically changed the landscape of retail shopping. Online retailing is currently the fastest growing sector in the US and will continue to grow at a compound annual rate of 12 percent through 2020, surpassing $1 trillion mark by 2027(Bose 2017). As shopping behaviors shift away from the offline purchasing of goods in brick-and-mortar stores, pundits are engaged in the ongoing debate on the role of traditional physical retailers in the digital economy. Mirroring this debate, we see a divergence of strategies among retailers in practice. On the one hand, several major retailers in the United States are closing physical stores. For instance, the iconic retailer Macy's which used to be a mainstay in America's malls, has closed over 15% of its 650 stores in recent years (Egan 2016). Similarly, other major retailers including Walmart and JCPenny have also pulled the plugs of hundreds of their stores (Gustafson and Reagan 2016). On the other hand, retailers like Nike and

Nordstrom are opening new physical stores and revamping them as "concept stores" to help customers experience and discover their products/services, so as to boost sales. Interestingly, pure online retailers like Amazon have begun to invest in their offline presence in recent years, as seen in their acquisition and launch of physical grocery stores and book stores, etc. (Bensinger 2014, Addady 2016). Similarly, the Chinese e-commerce giant, Alibaba, has invested more than $9.3 billion in offline stores since 2015 (Cadell 2017). These contradictory trends bring up questions on the role of physical stores in the digital age, with emphasis on whether they are still of value to retailers in face of the impending unstoppable digitization of the industry.

Apart from inconsistent strategies from different retailers, past academic studies have yielded different insights on the effect of physical stores on online sales. Some studies have found the presence of substitutive effects wherein newly opened physical stores compete with online stores for sales (Brynjolfsson et al. 2009, Forman et al. 2009, Choi and Bell 2011), while others found a complementary effect among the two channels (Balasubramanian et al. 2005, Ansari et al. 2008). As e-commerce technologies and practices mature, it is imperative to reexamine the inter-relationships between the sales channels as the effects of physical stores might have evolved over time as consumer mindsets is likely to shift with increasing digitization. To that end, marketing scholars have recently attempted to investigate the cross-channel effects between online and offline stores to understand the mechanisms that underlie these relationships (Avery et al. 2012, Wang and Goldfarb 2017). Specifically, these studies proposed that stores fulfilled roles in the purchase funnel through two channel capabilities: 1) conspicuous capabilities (e.g., immediate satisfaction of on-site purchase, product information gathering through direct interaction with products and salespersons, no shipping cost, etc.) and experiential capabilities (e.g., generating brand awareness and association, increasing reach and frequency of brand message over time, etc.). In particular, both

studies found that physical stores mainly played the latter role of generating awareness of the existence of the retailer's brand and building its brand associations, i.e., a billboard effect. Further questions on the role of stores follows from these findings: are the conspicuous aspects of stores less important in the increasingly digitized world of retailing, if not, when would these capabilities matter? Recently, Bell et al. (2017) and Kumar et al. (2019) provide evidence showing that the conspicuous benefits of product sampling and in-store interactions can be helpful in spurring the purchase of certain products under the right conditions.

Despite these body of works, several gaps in the literature continue to exist. First, the studied impacts of physical stores till date were all based on observations of the US market. Consequently, little is known about the dynamics of the offline-online consumer behavior in retail markets beyond the western world. Of particular interest is the Chinese retail market which has hit US$5.8 trillion in 2018, making it on par with the size of the US market. This sizable market continues to grow at 10 percent annual rate and is projected to be one of the world's largest consumer market, as over 60% of world's middle-class population will reside in Asia in the next ten years. More importantly, consumers in this market are known to possess shopping habits distinct from western consumers (Ettenson and Wagner 1991, Wang and Lin 2009), making it hard to determine if past results would apply to the retail market in China. Second, with the exception of Bell et al. (2017), most existing work are faced with the challenge of not being able to directly observe if the exposure to products showcased in physical stores led to subsequent purchase behavior online. Understandably, this is an empirical limitation due to a lack of detailed data. Finally, extant findings on the multi-channel relationships are derived from on retailers that carry a specific type of product, e.g., eye glasses (Bell et al. 2017), and apparels and home products (Kumar et al. 2019). Thus, it is unclear if these findings may generalize to products beyond these studied goods.

Noting these gaps, we study the impact of newly open stores of a large Chinese retailer on the online sales of customers living near these stores. The retailer under study serves as a good institutional setting to investigate the multi-channel relationship for a few reasons. First, it's a traditional offline-first retailer which carries a wide variety of products in store. Findings from the analysis of this data provide the much-needed insights to large pool of offline-first retailers, across different products. Second, the retailer has been exploring multi-channel, offline-online strategy since 2011 to which its customers have become relatively familiar with fulfilling purchases on both offline and online channels. This mature ecommerce setting allows us to abstract away from learning effects and heterogeneity in terms of ecommerce adoption preferences. Third, we were able to make use of a quasi-experimental setup in our estimation, as the retailer expanded its pool of physical stores during the study period. No new stores were launched in the twelve months preceding this expansion plan. Fourth, the retailer provided data that tracks online purchases at the household-level, which is more finer-grained compared to that in most previous studies. More importantly, we were able to observe which products are showcased in each physical store, allowing us to establish a stronger link between online purchases of consumers and the newly opened store. These highly detailed information also facilitates the ability to test for the mechanisms underlying the link between physical stores and online purchase. Finally, the retailer provided a complete view of the transactions that occur during the study period, covering the entire range of merchandise carried. Not only does this allow us to derive estimates mirroring the actual magnitude of the impact of physical stores on online sales, we are also able to arrive at a more generalized view of the store effect across various product types.

Using difference-in-differences identification coupled with propensity score matching, we find that physical stores can complement online channel in terms of boosting sales in the Chinese market. Specifically, our analyses reveal that following the opening of

a new store, the amount of online purchases increases by 26%, which translates to an additional online sales worth $40 thousand per week (i.e., approximately $2 million per year). This increase in online sales occur in both tract-level and household-level analyses. A triple difference analysis at the product-level provides further evidence of the complementarity between physical stores and online purchases. Specifically, our analysis shows that the increase in online sales happens for showcased products and not for non-showcased products, indicating that the positive impact of stores materializes through the showcasing of products in physical stores. We further find that online purchases for products showcased in physical stores increase for both high and low involvement products, which we interpret as evidence for conspicuous and experiential capabilities of stores at work. When segmenting locations by virgin territories and tracts with existing stores, we find that the new stores in virgin territories exert a larger impact on online purchases, though the opening of new stores in both types of locations yield positive and significant spillovers on online sales. Analysis conducted on different groups of customers show that physical stores drive inactive customers to purchase significantly more low involvement products, while they encourage active customers to make more online purchases across all product types.

Our work contributes to the multi-channel marketing literature from several aspects. First, we provide stronger evidence of the complementarity of physical stores to online purchases. In particular, we are able to link the online purchases of certain products to the offline showcasing of these products, an evidence that has not been shown in previous studies. This results help to dispel the wide spread belief that physical stores are less valuable in the digital age, particularly in the Chinese retail market. Second, evidence of the conspicuous and experiential benefits of physical stores being simultaneously present in a single setting clarifies the role of the offline channel in the multichannel strategy. This finding provide insights on the mechanisms through which the value of physical

stores manifest in the digital age. Third, our results also reveal nuanced insights on how the impact of new stores on online sales varies across locations and customer types. Knowledge of these details provide guidance to academics and practitioners alike on how best to deploy physical stores to enhance online sales.

4.2 Related Work and Theoretical Background

4.2.1 Multi-channel Retailing

The rapid development of the Internet technologies and their applications to the retailing industry over the past decade has made the online channel an imperative consumer touchpoint. This new marketing touchpoint bears several advantages, by which the most crucial one lies in its ability of overcoming geographical constraints in their communications with potential customers. Retailers have experimented with multi-channel strategies by having online storefronts to supplement their existing pool of offline stores, to understand if this new digital channel would improve overall business performance (Verhoef et al. 2015, Narang and Shankar 2019). These are non-trivial decisions to make as adding new channels can affect existing channels since different channels operates quite differently in the customer conversion process. Brick-and-mortars work by allowing consumers to touch-and-feel merchandise, reducing their uncertainties about products, and providing instant gratification of on-site purchase. Internet channels, on the other hand, convey product information through product reviews and user ratings to support decision making, and enhance shopping satisfaction by providing a broad selection of products (Brynjolfsson et al. 2013).

Previous studies have investigated the contribution of different channels on firm performance. Earlier studies mainly focused on the impact of adding an online channel, and documented positive effects on sales volume (Ansari et al. 2008), financial performance

(Geyskens et al. 2002) and brand awareness (Halligan and Shah 2009). More recently, scholars started to study the impact of physical stores on online sales performance of traditional offline and online-first retailers. These studies yielded different insights. For instance, it was found that substitution occurs as newly opened physical stores leads to fewer purchases and activity online, indicating the presence of cannibalization (Brynjolfsson et al. 2009, Forman et al. 2009, Choi and Bell 2011). This is especially true for those customers who live close to store locations (Shriver and Bollinger 2015). Yet, another set of studies found the opposite to be true, that is, offline stores compliment online channels by driving up online sales (Balasubramanian et al. 2005, Ansari et al. 2008).

Motivated by those contradictory findings, further studies dug deeper into the question by investigating the underlying mechanisms of the multi-channel effect. In particular, Avery et al. (2012) posit that the nature of the economic relationship between offline and online stores (i.e., substitution or complementarity) is dependent on how closely the channel capabilities are related. If a new channel duplicates existing capabilities or offers superior capabilities, it would cannibalize sales from the existing channel. In contrast, additional demand in existing channels will be generated if the new channel provides capabilities complementary to existing channel. Avery et al. (2012) theorize that offline stores can have conspicuous or experiential capabilities, which manifest in short-term and long-term periods, respectively. They show that having a denser store population in an area can enhance branding effects leading to increased online purchases over time. Drawing on the framework of Avery et al. (2012), Wang and Goldfarb (2017) emphasize the role of physical stores in providing the information about a retailer's existence. Using the data from a US-based specialty retailer, they found that store openings tend to attract more first-time customers and that the increase in online sales

is mainly driven by these new customers, suggesting that the physical stores exert a billboard effect to draw in sales from new customers.

On the other hand, Bell et al. (2017) found that online sales of products is driven by the opening of offline showrooms reduce uncertainty in the purchase process by providing tactile product information to consumers. This result suggests that the conspicuous capabilities of physical stores play an important role in helping consumers converge in their purchase decisions. We note that the evidence presented in Bell et al. (2017) holds for an online-first retailer specializing in eye-wear. Pauwels and Neslin (2015) found physical stores play a conspicuous role of facilitating product returns and exchanges. Kim et al. (2019) showed that access to physical stores reduces uncertainty about product fit and quality, which manifests through impacts on product returns. Under the context of apparels and home-products (i.e., goods with non-digital attributes), Kumar et al. (2019) found that that physical stores increase online sales through a store engagement effect, which again demonstrates a conspicuous capability at work. In particular, the authors demonstrate the store engagement effect by showing that the proportion of consumers who made purchases in both online and offline stores increase at a greater rate than those who purchase online only, especially in locations where new stores are opened relative to locations where there are no new stores.

Thus far, we have seen one set of works providing evidence of physical stores mainly working through its experiential capability, while another set of works showing that brick-and-mortars increase online sales via its conspicuous capabilities. Taken jointly, these evidence coincidentally show the impact of offline stores on online purchase manifesting through one type of capability, but not simultaneously through both experiential and conspicuous capabilities. Thus, the question on whether physical stores can simultaneously work through both qualities remains as an open puzzle. Furthermore, With the exception of Bell et al. (2017), inferences are made without directly observing whether

physical stores influence subsequent online purchasing behaviors through the specific showcasing of certain products. Additionally, in the latter set of works, the conspicuous capabilities of physical store are demonstrated to hold largely on fit-and-feel products (i.e., eye-wear and apparel). Apart from this narrow set of products, we are unaware if retailers carrying other types of merchandise would stand to benefit from conspicuous aspects of opening physical stores.

4.2.2 Roles of Physical Stores in Chinese Markets

Apart of the gaps identified above, the multi-channel literature is silent on the the role of physical stores in markets beyond the western world. Yet, non-Western markets, especially that of the Chinese market, are rapidly growing in size to the point of overtaking the US in terms of sales volume. Cultural value systems are recognized as a powerful force in shaping consumers' motivations, lifestyles and product choices (Yau 1988, Tse et al. 1989, Wang et al. 2000). In particular, Chinese values have served as a clear and consistent system for generations, to which the learning and mastery of these value is a crucial prerequisite for achieving social status (Kindel 1985). Given the disparity between western and oriental cultural values, it is plausible that the causal mechanisms underlying the purchase behaviors of the US consumer may not generalize to that of the Chinese consumer (Kindel 1985, Yau 1988), which meant that past findings on the economic nature of physical stores may not materialize similarly in the Chinese retail market. In fact, industry experts have opined that physical shopping complexes in China have a good chance of surviving and thriving in the digital age, given that the lifestyle favored by Chinese consumers support shopping visits to physical stores (Shepard 2017). Among the numerous conceptualizations of the Chinese cultural values, the values of *thriftiness* and *relational orientation* are most pertinent to our discussion on the potential impacts of stores on online sales.

Synergies with Conspicuous Capability

First, the Chinese consumer is influenced by the oriental values which emphasize thrift and value consciousness (Wang and Lin 2009). Individuals living in collectivist cultures tend to adopt frugal mindsets as they value inter-generational over individual interests, and are likely to maintain a longer time horizon with respect to their individual consumption (West 1989). A direct consequence of thriftiness and value consciousness meant that Chinese consumers, relative to their American counterparts, are more willing to invest significant amount of efforts in obtaining product information before making a purchase (Ackerman and Tellis 2001, Pattaratanakun and Mak 2015), so as to reduce the chances of making a bad purchase decision. Such strategic consumers are willing to delay their purchases when they deemed product value to be uncertain (Swinney 2011), and would actively seek out product information from various sources to affirm their value. While the online channel offer consumers a wide variety of benefits, e.g., access to large product variety and enables side-by-side comparison of products (Smith and Brynjolfsson 2001, Brynjolfsson et al. 2003), it does not allow for the evaluation of non-digital attributes of the product. Offline stores fill this gap by moving undecided consumers toward purchase decisions by allowing a "full inspection" of the product through the touching and feeling of the merchandise (Avery et al. 2012, Bell et al. 2017, Ching and Ishihara 2012). Empirical evidence suggests that Chinese consumers actively seek out physical stores to shop at, especially at those that are conveniently located near their living vicinity (Ettenson and Wagner 1991). This suggests that Chinese customers are likely to place a strong emphasis on the physical examination of products to extract tactile information that are not readily available in online descriptions.

Second, the Chinese consumer is immersed in a culture that has a strong relational orientation, by which a distinctive set of marketing implications ensues (Wang and Lin 2009). Compared to the Western counterpart, oriental individuals are more

consensual-driven and tend to perceive themselves as more connected to others, such as the family and the community (Wang et al. 2000). This difference between Western and non-Western self-construal approximates the distinction in individualist and collectivist selves (Hsu and Marsella 1985, Markus and Oyserman 1989). Due to this distinction in self-construal, Chinese consumers react more favorably to marketing stimulus with connected appeals (i.e., relational ties with significant others and interdependence), while American participants favor marketing approaches with separated appeals (i.e., uniqueness, individuality, independence, autonomy) (Wang et al. 2000). A strong relational orientation also influences how Chinese consumers views the act of shopping. Because Chinese individuals value spending time with friends and family, shopping is deemed as an important social activity as it allows people to get together naturally (Wang and Lin 2009). Moreover, a relational orientation meant that peer pressure and opinion leaders often serve as key influential factors in purchase decisions. This social nature of shopping indicates that Chinese consumers prefer to develop personal relationships with salespersons, to which greater emphasis is placed on a salesperson's mannerism and less on the tangible aspects of shopping (e.g., store return policy) (Ettenson and Wagner 1991). This emphasis on interpersonal relationships is built on the principle of doing favors for others, which is considered a form of social investment by which reciprocity is expected (Yau 1988). As long as shopping is perceived to be a social occasion, physical stores and salesmanship will continue to play a significant role in the Chinese market, even with the increasing growth of digital commerce.

Thus far, we have seen how Chinese cultural values can influence purchasing process of consumers in China. The values of thriftiness and relational orientation meant that Chinese consumers are likely to value the conspicuous capabilities of physical stores i.e., being able to physically examine merchandise, shop with friends, establish relationships with salespersons in stores. The direct interactions with the products and salespersons

build up product interest, spur purchase intentions, and solidify purchase decisions. Having experienced a product in store in various dimensions, customers are likely to follow up with further product searches online (Verhoef et al. 2007), especially at the same retailer given that the prior relational links built up with their in-store salespersons. Moreover, with limitations on inventory size of physical stores, shopping trips can lead to online purchase at the same retailer for a particular product variant that is not available in store (certain models, sizes or colors are available online). These are practical instances of how physical stores complement and encourage online sales in the Chinese market.

Synergies with Experiential Capability

In addition to the conspicuous capabilities at work, physical stores can also play important experiential roles in generating awareness and shaping perceptions of brand images (Jacoby and Mazursky 1984, Keller 1993). Valuable brand associations built through the offline channel may be transferred to other channels. Wang and Goldfarb (2017) showed that it is possible that the brand salience conferred by physical stores can stimulate consumers' purchases from the same retailer from a different channel. On top of a "billboard effect", it is quite possible that the experiential aspect of physical stores may also influence brand perceptions at the product level. The awareness and familiarity of certain products is heightened when consumers are exposed to products showcased in physical stores. The increased familiarity leads consumers to pay more attention to these products when they are browsing through the online store. Specifically, the engagement with products on display in stores produces an exposure effect which is shown to enhance consumers' affective and evaluative responses to these brands (Obermiller 1985). Recent exposure to a brand or product increases its salience and recall rate, producing greater levels of top-of-mind awareness for showcased merchandise, which can in

turn inhibit the recall of competing products (Gruber 1969, Alba and Chattopadhyay 1986).

The experiential capability of physical stores in enhancing the brand image of products is particularly important in satisfying the Chinese consumers. The value conscious and consensual nature of these consumers meant that they are inclined to adopt a risk adverse approach in making purchase decisions, and would prefer to buy from brands that are established as the normative standard for their reference group (Wang and Lin 2009, Yau 1988). Here, having a wider exposure via in-store displays to gain a top-of-mind awareness increases the likelihood of purchase conversion, as the first brand in the Chinese consumer mind is her most probable choice (Xu 1992). At the same time, Chinese individuals tend to conform to what others are buying as a result of their social consensual tendency. Thus, a broader exposure through product showcasing in multiple physical stores would be helpful in spurring purchase intents, as product familiarity is increased across a larger consumer base. In sum, product placements in offline stores are likely to spur the purchase volume on the online channel, via the experiential capabilities of physical stores in the Chinese context.

Sales Impact on Different Products

Heightened sales due to the conspicuous and experiential capabilities of offline stores are likely to manifest in the purchase of different product types. Prior marketing work documents that consumer decision process differ significantly with the level of product involvement (Clarke and Belk 1979, Engel and Roger 1995). *Product involvement* refers to a consumer's perceived importance or complexity of a product (Traylor 1981, Richins and Bloch 1986). Consumers typically engage in an "extensive search for information or a comprehensive evaluation of the choice alternatives" (Zaichkowsky 1985)

to make the right decision for high-involvement products, while customers are less engaged in their search process when purchasing low-involvement products. Examples of high-involvement products include costlier, durable products such as electronics, large appliances, and furniture, to which consumers tend to expend effort in collecting product information through product inspection and/or communicating with in-store salespersons (Laurent and Kapferer 1985). Thus, the conspicuous aspect of physical stores facilitates the effortful information gathering process that consumers engage in the purchase of high involvement products. Should Chinese consumers value the the conspicuous aspects of physical stores, an increase in high involvement products should ensue upon the launch of new physical stores.

On the other hand, low-cost, non-durable products like groceries, household supplies, and sundry goods are instances of low-involvement products (Kannan et al. 2001, Gu et al. 2012) to which consumers are relatively less worried about making bad purchase decisions compared to high involvement products (Levy and Nebenzahl 2008, Mathwick and Rigdon 2004). Although consumers do not expend significant efforts in amassing information for the purchase of low involvement products, the experiential component of physical stores can play a role in increasing the purchase of these products by exposing consumers to these products and enhancing their top-of-mind awareness of these products. Hence, if the effects of experiential capabilities of stores were to exist in the Chinese retail market, physical stores would induce an increase in the purchase of low involvement products.

4.3 Empirical Context and Data Description

4.3.1 Empirical Context

We base our empirical analysis on one of the largest offline-first retailers in China. Similar to large US retailers (e.g., Walmart or Target), our focal retailer has thousands of physical stores across China.[1] On top of physical stores, the retailer has been investing in its online presence, by having a digital storefront since 2011. Products sold by our retailer include household appliances (e.g., TVs, laundry machines, refrigerators), electronic devices (e.g., laptops, cameras), general merchandise (e.g., Fast Moving Consumer Goods (FMCG), office supplies), and household commodities (e.g., furniture, kitchenware). In our data, we observe all online and offline purchases made by consumers between August 2014 and March 2016 in three major Chinese cities, i.e., Beijing, Shanghai and Nanjing. For each purchase, we have detailed information including the product, prices and quantity purchased. In addition, we have information on the consumer's address (i.e., latitude and longitude coordinates) for orders placed online. Based on the dataset provided by our retailer, we are also aware of the locations of all existing physical stores in the three cities. For each store, we are able to see what products were showcased during the observation period. During our study period, we observe 67 new stores launched in those three cities, which we also have information on their location coordinates.

4.3.2 Sales Tract Definition

The official geographical segmentation of the Chinese cities based administrative districts is overly broad for our purpose, as each district often contain several physical

[1] We are unable to disclose the identity of the retailer in writing due to a non-disclosure agreement.

stores making it hard to isolate the effect of each store on online sales.[2] To address this issue, we rely on a clustering-based approach to define sale tracts in our study context. To do so, we rely on a three step procedure. First, we amass mailing addresses (as represented by the latitude and longitude coordinates) of all customers who have placed online orders. Using these coordinates, we use K-means clustering (Forgy 1965) to group these locations into clusters. As a distance-based technique, K-means clustering works by assigning geographical locations into the same cluster if they are close to one another (relative to other coordinates). Based on the typical population size in an average neighborhood of our study cities (TMO Group 2018), we set the clustering parameter (i.e., total number of clusters) such that the end result of our clustering exercise produces neighborhood clusters containing 100 to 300 customers on average. Based on this criterion, our clustering algorithm generated 3000 different clusters. Fig. 4.1 illustrates the clusters identified in Nanjing: the original locations of households are represented by grey dots and the centroid of each cluster is represented by a red cross. Fig. 4.2 shows that the distribution of customers across different clusters is well-dispersed, validating that our choice of the clustering parameter is appropriate and does not produce skewed outcomes.

Having established the neighborhoods for our households, we next associate neighborhoods to the physical stores. This step allows us to understand which store is serving each customer from a distance perspective, to create sales tracts for our study. A 5 kilometers distance is set as an initial threshold in which customers are willing to travel to reach a store.[3] The distance is measured by the Euclidean distance between the location of a store and the center of that tract. In our final step, we cluster the leftover neighbourhoods that are not served by any close-by stores together so that they can

[2]In Beijing, there are 16 administrative districts containing over 150 physical stores. This meant that each area is served by multiple stores. The situation is similar in Shanghai and Nanjing.

[3]In our robustness checks, we also attempted alternative cut-offs of 2 and 10 kilometers in our definition of the sales tracts.

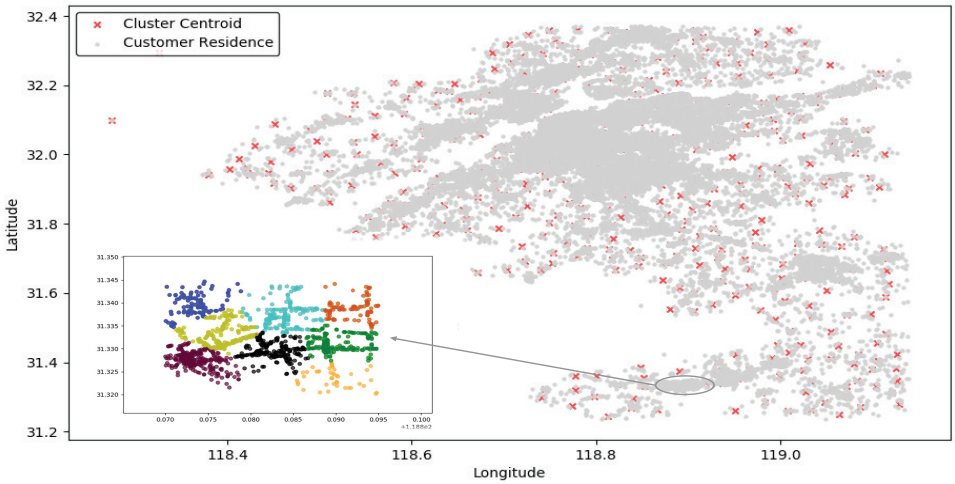


Figure 4.1: Clusters Discovered in Nanjing.

form sales tracts of their own. Through this three-step procedure, 585 sale tracts across three cities are formed. With this sale tracts define, we go on to examine the impact of a new store on the purchase behaviors of consumers living within the same sales tract.

4.3.3 Descriptive Statistics

Using defined sale tracts, we generate a panel of weekly online sales volume, matched with the indicator of whether a physical store is present in the tract in a given week. The weekly sales volume of a tract is the aggregate quantity of products ordered online by consumers living within it. During our study period, we observe 67 instances of new store openings and no instance of store closing. In total, our panel have 29,040 observations spanned across 86 weeks.

We include several control variables in our dataset. First, we control for the purchasing power of consumers in each sale tract by measuring its average monthly volume of online orders for the four largest product categories (i.e., household appliances, electronic devices, kitchen appliances, fast moving consumer goods) in a six-month window

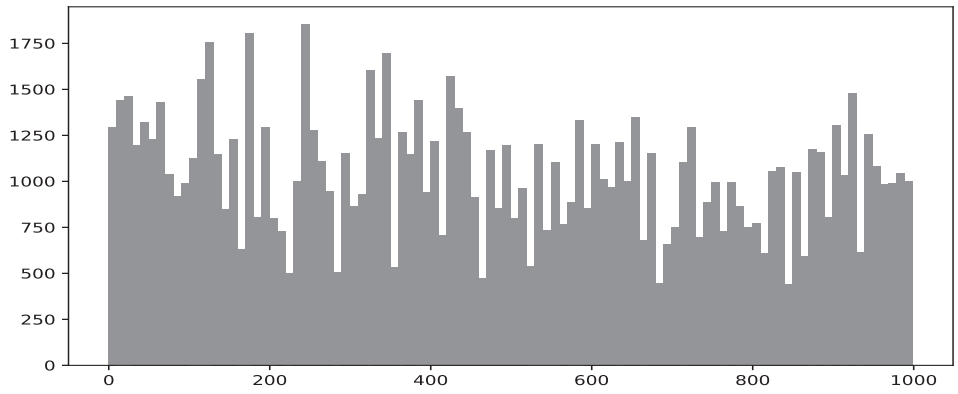


Figure 4.2: Distribution of Consumer Base of Different Clusters in Nanjing.

before our actual study period (i.e., February 2014 to July 2014). Second, we use the average housing price of neighborhoods in each tract as an another indicator of the purchasing ability of the customer base in each tract. That is, a higher housing price implies greater spending capabilities, and vice versa. Third, we control for the size of the online customer base for each tract by capturing the number of customers who have ever placed online orders in pre-observation period and the size of the tract (proxied using radius of tract in kilometers). Finally, we also include the distance to the city center as a control variable, given that locations closer to the city center are likely to experience greater flow of customer traffic in stores leading to greater purchase incidence. Descriptive statistics of all independent and dependent variables (excluding pre-observation period) are presented in Table 4.1.

4.3.4 Natural Experiment Involving Offline Store Expansion

In September 2015, our focal retailer collaborated with a property developer, to expand their offline operations. This collaboration entails the opening of new stores in about a hundred locations. Prior to this expansion effort, the retailer did not engage in any store openings in the last twelve months. As seen in Fig. 4.3, we can clearly see that several

Table 4.1: Descriptive Statistics

	Observations	Mean	Std.Dev.	Min	Max
Dependent Variables					
OnlineSales ($1K)	29040	80.02	885.49	0.00	40273.11
TotalSales ($1K)	29040	176.66	1087.82	0.00	42245.3
OnlineOrders (1K)	29040	1.20	13.12	0.00	1385.26
TotalOrders (1K)	29040	1.82	14.88	1.00	1386.26
Independent Variables					
StoreOpening	29040	0.03	0.17	0.00	1.00
Appliances	29040	16.88	59.50	0.00	839.67
Electronics	29040	121.34	702.48	0.00	11513.25
Houseware	29040	28.03	134.49	0.00	2453.00
FMCG	29040	295.46	588.10	0.17	5555.50
HousingPrice ($)	29040	6165.09	3131.90	919.54	16038.42
Userbase	29040	1088.25	1628.80	4.00	15735.00
TractRadius (km)	29040	3.04	1.71	0.34	9.78
DistanceToCityCenter (km)	29040	32.63	22.23	0.98	109.10

new brick-and-mortar stores were opened after the launch of the collaboration beginning in September 2015. The expansion plan represents a natural shock in our study period, to which the sales tracts across the three cities experience a greater incidence of store openings. Through this expansion effort, new stores are opened in both tracts that do not have stores ("virgin terrorities") and tracts that already have existing stores (from the same retailer). By exploiting the variation in store opening across sale tracts, we can quantify the impact of physical stores on online purchases, by contrasting changes in sale volume before and after store openings via a difference-in-difference framework.

The major challenge to empirically identifying the impact of store opening on online purchase is the endogeneity concerns of unobserved biases that are inherent in the store launch decisions. It is plausible that the retailer has strategically chosen to open stores in certain tracts due to considerations of their market potential which will have profitability impacts. As a result, the mere comparison of sales growth between those "treated"

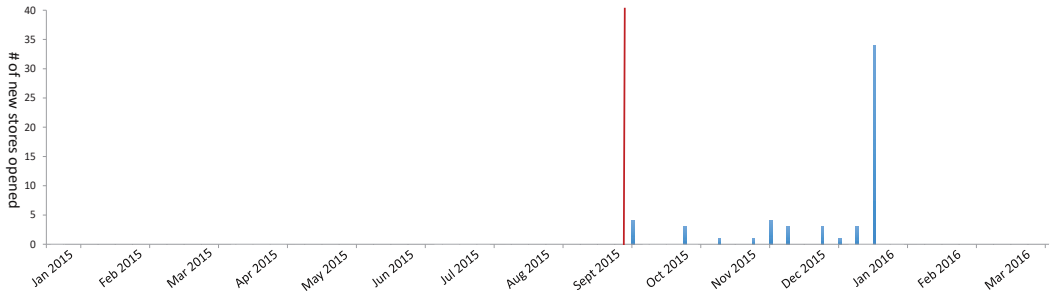


Figure 4.3: New Store Opening Over Time.

tracts and "control" tracts without accounting for such endogeneity can lead to biased estimation of the store opening effect. We employ several econometric techniques to account for this issue, which we described more detail next.

4.4 Econometric Model and Identification Strategies

4.4.1 Difference-in-Differences

We apply a difference-in-differences (henceforth referred to as DID) framework coupled with propensity score matching to measure the effect of offline stores on online purchases. The DID model is widely used for measuring treatment effect in a given period, by contrasting the outcomes of a treated group that received a certain intervention with the same outcomes of a control group that was not exposed to such intervention in the same period (Meyer 1995, Angrist and Pischke 2008). In our context, the "treated" units are tracts with a new store opened during our observation period, and the "control" group includes tracts without new stores. The DID framework has been generally accepted as a reliable model to estimate the impact of physical stores in past works (Bell et al. 2014, Wang and Goldfarb 2017), as it accounts for heterogeneous differences across locations and time. In essence, our DID estimates capture the average treatment effect of a new physical store on the treated tract after its opening on purchase behaviors of customers

living in the tract. Specifically, our main estimation equation for tract i in week t is:

$$\begin{aligned} Sales_{it} = \beta_0 + \beta_1 * (TreatedTract_i * AfterTreatment_t) + \\ \alpha * TreatedTract_i + \gamma * AfterTreatment_t + \beta_2 * X_{it} + \mu_t + \epsilon_{it}. \end{aligned} \tag{4.1}$$

Our main dependent variable is the logged number of products ordered online by customers living in tract i at time t, i.e., $Log(1 + OnlineOrders_{it})$. $TreatedTract_i$ is a indicator on whether a tract experiences a new store launch during the study period. It is assigned "1" if it is a treated tract i and "0" otherwise. This binary term controls for the time-invariant geographical differences that may exist between treated and untreated sale tracts. $AfterTreatment_t$ is a binary variable that indicates the post expansion period for all tracts, and is denoted as "1" after week 60, i.e., period when offline expansion starts, and "0" before that. This variable accounts for potential temporal-related factors that may simultaneously affect the timing of store openings and purchase behaviors of customers across tracts (e.g., marketing campaigns and advertising efforts to promote awareness of new store openings). The coefficient β_1 for the interaction term, i.e.,$TreatedTracts_i * Aftertreatment_t$, represents the DID estimator which captures the average incremental effect of store opening on online sales. More specifically, it quantifies how online purchase behaviors of customers in treated tracts change after the launch of offline stores, relative to the same difference of control tracts during the same period. Standard errors are clustered by tracts to allow for serial correlation over time (Bertrand et al. 2004).

In our model, we also control for observed covariates, X_{it}, which include each sale tract's average housing price, distance to central area of the city, size of customer base (via number of customers and tract radius), customers' past purchase behaviors, and week-fixed effects, u_t. These covariates and time fixed-effects control for the observed differences across tracts and systematic changes over time.

We also attempt an alternative DID specification by which we utilize an indicator to denote the launch of each store (i.e., concurrent variation across tract and time), so that the individual tract-level fixed effects can be included in the specification to account for unobserverable factors related to the location. Arguably, tract-fixed effects might be able to do a better job than the use of covariates, given that the use of covariates leaves questions on unobservables of tract differences not being accounted for. The alternative estimation equation for tract i in week t is as follows:

$$Outcome_{it} = \beta_0 + \beta_1 * Treated_{it} + \alpha_i + \mu_t + \epsilon_{it}. \tag{4.2}$$

where $Treated_{it}$ is denoted as "1" after a new store is opened in tract i after week t, and "0" otherwise. Similar to the previous model, the coefficient β_1 captures the marginal change in online purchases in sale tracts served by new stores relative sale tracts that do not have stores in the same time period. α_i controls for the unobserved characteristics of each tract that might have an impact on its online sales.

4.4.2 Propensity Score Matching

In addition to the use of the DID framework and covariates, we rely on propensity score matching as another defence against endogeneity concerns. The matching process helps to derive a set of control tracts that share similar characteristics with that of the treated tracts, such that each treated sale tract is paired up with a control tract that has a similar propensity of being treated (more details on matching are provided in Section 4.4.2). Assuming that matching is effective in accounting for effects from the unobservables, the regressions based on a matched set of tracts would abstract away from extraneous factors that are simultaneously correlated with the treatment status

and the outcome, allowing for an unbiased estimate to be derived (Rosenbaum and Rubin 1983).

Given that a major source of endogeneity lies in the strategic decisions of launching stores in certain tracts, we match tracts based on their profitability potential using the set of covariates listed earlier, i.e,. sales volume of four largest product categories, average housing prices, user base size, tract radius and distance to city center. At the same time, these factors are likely to affect customers' purchase behaviors and should be accounted for, so that tracts under consideration would be more similar. In the first three columns of Table 4.2, we present summary statistics for those variables before and after matching. In the pre-matched variables, we observe significant differences between control and treated tracts. For instance, the monthly average amount of online purchases of different types of products of customers who live in treated tracts tend to be higher than control tracts; the treated tracts tend to have a larger user base size than the control tracts; and treated tracts are generally closer to city centers relative control tracts. These differences are aligned with our intuition that retailers do not open stores in random locations, but are guided by a combination of market potential and growth opportunities of locations.

Our baseline matching uses one-to-one matching with replacement to derive the control tracts that are most similar to the treated tracts under a caliper size set to be the standard deviation of the propensity scores. In addition, we also tried three nearest neighbors matching to assess whether inclusion of more control tracts would affect the results. We find that under the baseline matching process, appropriate matches are derived. Columns (5) to (7) in Table 4.2 show that the differences between the treated and control means are greatly reduced and are no longer significantly different (i.e., based on t-test on mean difference) after matching. Following Haviland et al. (2007), we conduct a balance check by comparing the standardized bias before and after matching.

Table 4.2: Illustrative Summary Statistics-Mean(SD)

	BeforeMatch				AfterMatch				Std.bias
	(1) Treated	(2) Control	(3) Mean Diff	(4) Std.bias	(5) Treated	(6) Control	(7) Mean Diff	(8) Std.bias	(9) Improvement (%)
Appliances	2.56	1.74	0.82***	0.38	2.56	2.62	−0.06	0.02	93.8
	(1.16)	(1.25)	(73.83)		(1.16)	(1.30)	(3.37)		
Electronics	3.37	2.19	1.18***	0.42	3.37	3.31	0.06	0.02	95.7
	(2.06)	(1.82)	(59.76)		(2.06)	(1.88)	(1.89)		
Houseware	2.73	1.95	0.78***	0.33	2.73	2.80	−0.07	0.03	92.3
	(1.22)	(1.46)	(68.27)		(1.22)	(1.32)	(3.40)		
FMCG	5.34	4.17	1.17***	0.24	5.34	5.40	−0.06	0.01	95.4
	(1.45)	(1.85)	(86.06)		(1.45)	(1.41)	(3.92)		
HousingPrice ($)	5780.44	6217.82	−437.38**	0.07	5780.44	5904.31	−123.87	0.02	70.9
	(2997.26)	(3146.31)	(11.56)		(2997.26)	(3186.38)	(2.40)		
Userbase	1839.30	985.29	854.01***	0.58	1839.30	1982.77	−143.47	0.08	87.0
	(1663.19)	(1596.76)	(52.92)		(1663.19)	(2650.17)	(7.42)		
TractRadius (km)	3.87	2.39	1.48***	0.46	3.87	3.76	0.11	0.03	93.7
	(1.60)	(1.64)	(95.62)		(1.60)	(1.67)	(0.36)		
DistanceToCityCenter (km)	26.76	33.43	−6.67***	0.22	26.76	26.58	0.18	0.01	96.9
	(17.06)	(22.73)	(44.02)		(17.06)	(18.83)	(1.26)		

As shown in Columns (4) and (8), the biases between the two groups are greatly reduced after matching, with a 70 to 90 percent reduction across different covariates as shown in Column (9). These results affirm that those two groups are statistically similar after matching.

4.4.3 Individual and Product level Analysis

The aforementioned analysis aims at identifying aggregate online sales increase in a tract resulting from the new store openings. While the use of covariates and matching are helpful in accounting for extraneous factors at the tract level, there might still exist unobserved geographical effects that simultaneously affect the launch of new store in the tract and the average online purchase levels of the location. To further alleviate such endogeneity concerns, we perform a robustness check at a finer unit of analysis that examines that change in online purchase levels at the individual household level by contrasting "treated" households that have been experience a new store launch near them with "control" households that did not. In this analysis, we use stratified sampling

to randomly pick 20,000 households from different sale tracts. We utilize a population-weighted sampling strategy, so that more households are sampled from tracts with higher population density. We then use the weekly household online purchase volume as an dependent variable. The model specification used here is similar to Equations (1) and (2). Should the store entry have a positive impact on households' online purchase, the DID estimate would produce a significant coefficient.

In addition to performing a finer level analysis at the individual level, we also attempt to conduct our analysis at the product level. Given that the exposure to products showcased in physical store is one of the main mechanisms by which physical stores induce and stimulate online purchases, it would be reasonable to assess if a statistical relationship exists between showcased products and online purchase in each tract. A unique advantage we have through our unusually rich dataset is the ability to see the products that are showcased in each store at each period. By exploiting the variation in products showcased across stores and time, we apply a third layer of differencing to our DID framework to arrive at a triple-difference estimation approach (henceforth referred to as DDD) model (Matsa and Miller 2013). As an extension of DID, an additional layer of differences can make parallel trends assumptions across different groups more plausible. In our setting, the DDD estimator compares the changes in the outcome variables, e.g., amount of online purchases of each product before and after its showcasing in a tract with the corresponding difference in other tracts where those products never get showcased in the same time period. The DDD specification is as follows:

$$Outcome_{ijt} = \beta_0 + \beta_1^{DDD} * Showcased_{ijt} + \alpha_i + \mu_t + \epsilon_{ijt}. \tag{4.3}$$

where $Showcased_{ijt}$ is denoted as 1 after a product is showcased in tract i after week t and 0 otherwise. Under this setting, β_1^{DDD} measures the change in the difference in product sales before and after product showcasing in certain tracts, compared to sales difference of the same set of products over the same period in other tracts where those products are not on display. On top of the week fixed effects, we include a product level fixed effects in our specification to alleviate potential confounding trend arising from product-specific characteristics. This fixed effect controls for the selection of certain products that are more likely to be chosen to be showcased in stores. [4] The tract level fixed effects further helps to account for the fact that some products are better received in certain areas than others, with or without showcasing. For robustness, we include tract-product fixed effects to control for the possibility that some products might be more popular in certain locations. Similarly, we also utilize a product-time fixed effects to see if the results might be affected by the general popularity of certain products at certain time periods. In our DDD analysis, we utilize the universal product code (UPCs) to identify individual products based on their model and variant. However, an analysis at the UPC level is met with operational concerns given that the global set of products sold by our retailer is enormous. To resolve this operational issue, we randomly sample ten distinct UPC codes from each of the fourteen product categories[5] that the retailer carry, resulting in the use of 140 unique products in our DDD regressions. To check if the results are robust, we randomly sampled products thrice and repeated the respective regressions to arrive at three sets of DDD estimates.

While a DDD estimation based on product level showcasing is insightful, we are only able to cover a select set of products in this analysis. To further ascertain that the showcasing of a particular product in the physical store has an influence on online

[4] The products used in this analysis were showcased in some tracts and not in other tracts.

[5] 1. imported snacks and healthcare 2. sundry goods 3. food and drinks 4. skin care 5. household supplies 6. maternity 7. electronic devices 8. houseware 9. laundry machine and refrigerator 10. A/C 11. PC/laptop 12. kitchenware 13. TV 14. communication

sales of that product, we perform an additional check where we consider all products by splitting them into two groups based on whether they have ever been showcased in a tract before. Specifically, we tabulate the online sales volume of products that have been showcased in a tract and products that have not been showcased before in a tract, and use that as dependent variables in Equations (1) and (2). Should the showcasing of product play a role in inducing greater online purchase, we should see the DID estimate having a significant coefficient on the online sales of showcased products. An absence of a significant DID estimate for the online sales of non-showcased product will affirm that the increase in online sales comes mainly from exposing consumers to products showcased in stores.

4.5 Empirical Results

4.5.1 Impact of Stores on Online Sales

We first present our main results based on the DID setup. Table 4.3 shows the regression results from different models (i.e., OLS with log-transformed online purchase volume: Equation 4.1, Negative Binomial, and specification with tract fixed effects: Equation 4.2). We run each model using both unmatched and matched samples to see how the results might differ. From the results, we could see that the DID estimates from different models are positive and significant across various models, which indicate that online sales volume of a tract increases after the launch of a new store in the same tract. Results on matched samples also indicated that the launch of physical stores has a significant positive impact on online sales, albeit with a smaller magnitude. Based on the coefficient estimate of the main model using matched samples (i.e., Column 2 of Table 4.3), the online sales of tracts with new physical stores increases by 25.8% on average. Using average monetary sales values, this additional online sale volume

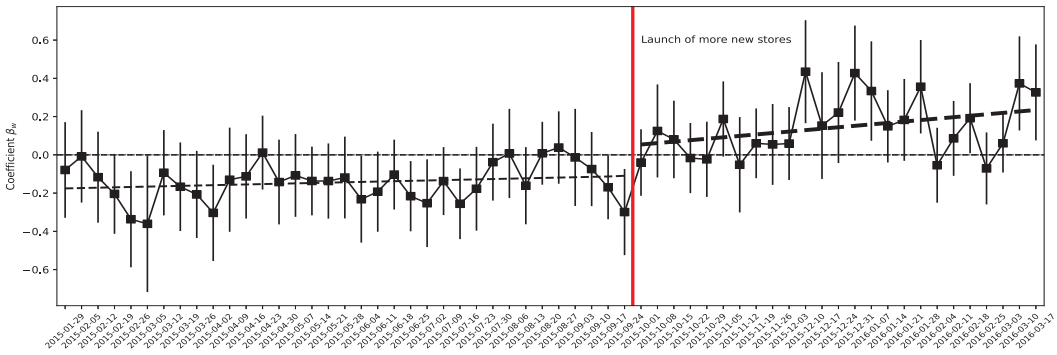
Table 4.3: Impact of Store Opening on Total Sales (Volume)

	OLS (logged)		NB		FixedEffect	
	(1) Unmatched	(2) Matched	(3) Unmatched	(4) Matched	(5) Unmatched	(6) Matched
$TreatedTracts_i$*	0.43***	0.41***	1.05***	0.88***		
$Aftertreatment_t$	(0.09)	(0.10)	(0.32)	(0.19)		
$TreatedTracts_i$	0.14	0.04	−0.01	0.15		
	(0.14)	(0.09)	(0.17)	(0.10)		
$Aftertreatment_t$	−0.23***	−0.25***	−0.04	−0.33*		
	(0.04)	(0.09)	(0.12)	(0.19)		
	(0.03)	(0.04)	(0.05)	(0.04)		
Treated					0.42***	0.35***
DistanceToCityCenter (km)	−0.02***	0.01*	−0.01***	0.00		
	(0.00)	(0.00)	(0.00)	(0.00)		
Userbase	0.00	0.00	−0.00**	0.00		
	(0.00)	(0.00)	(0.00)	(0.00)		
HousingPrice ($)	0.00***	0.00	0.00*	0.00		
	(0.00)	(0.00)	(0.00)	(0.00)		
Appliances	1.04***	0.10	1.29***	−0.01		
	(0.10)	(0.08)	(0.12)	(0.08)		
Electronics	−0.00	−0.08	−0.00**	−0.06		
	(0.00)	(0.05)	(0.00)	(0.07)		
Houseware	0.00	0.55***	−0.00	0.60***		
	(0.00)	(0.16)	(0.00)	(0.18)		
FMCG	0.00*	0.61***	0.00	0.66***		
	(0.00)	(0.10)	(0.00)	(0.10)		
Observations	29040	6395	29040	6395	29040	6395
City fixed effect	Yes	Yes	Yes	Yes		
Week fixed effect	Yes	Yes	Yes	Yes		
Tract fixed effect					Yes	Yes
R^2	0.77	0.83			0.43	0.50
Pseudo R^2			0.09	0.10		

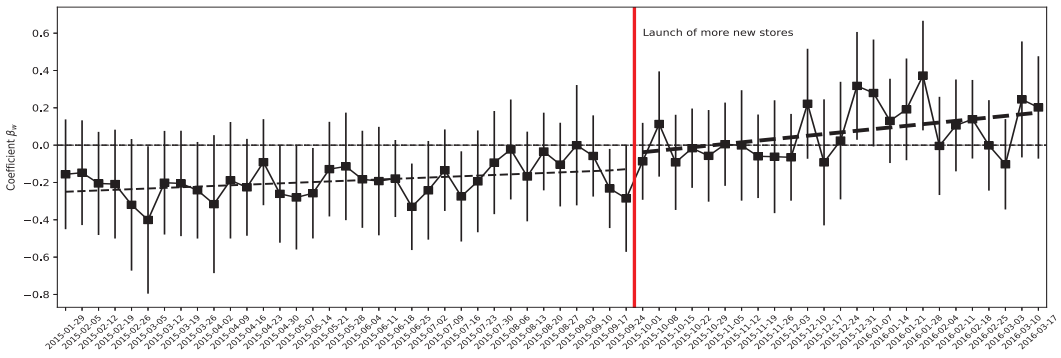
*significant at 10%;**significant at 5%;***significant at 1%

induced by new stores translates into approximately US$40 thousand per week, which equates to an average annual increase of US$2 million. We also note that the coefficient magnitudes derived under Equation 4.1 and Equation 4.2 are highly comparable, suggesting that the chosen set of covariates is as proficient as the tract fixed effects in capturing extraneous effects. In all, these results suggest evidence of a complementarity effect between offline and online channels.

In addition to online purchases, we also run a regression analysis for total purchase volume in each tract (sum of online and offline sales volumes). The results show that following the opening of a new store in a tract, total sales from that tract rises significantly ($\beta_1 = 0.41$, $p < .01$ based on the main model). Specifically, the total sales volume increases by about 50% on average, which is worth US$198.8 thousand on a weekly basis. Upon checking with the store managers from our retailers, their internal information show that the additional sales is able cover the cost of operating the new stores, i.e., positive profits are made.



(a) `Unmatched Sample`



(b) `Matched Sample`

Figure 4.4: Coefficients of the Weekly Interactions Before and After Offline Expansion.

Next, we assess the validity of the parallel trends assumption underlying the DID approach. To do so, we include the interactions of week dummies and the treatment indicator to our main regression specification (i.e., adding $\sum_{w=26,\ldots,85} \beta_w(TreatedTract_i * \mu_w)$ to Equation (1)) to capture the effect of store opening over time. Fig. 4.4 visualizes the estimated coefficients β_w before and after the launch of each store. From the figures of both the unmatched (Fig. 4.4 (a)) and matched sample (Fig. 4.4 (b)), the coefficients are mostly non-significant, indicating an absence of a pre-treatment trend. For coefficients that are significant, we note that they are negative in nature, which goes against the concern that treated tracts bear certain characteristics which allowed them to gain a greater sales volume even without the treatment. Trend lines plotted to approximate the sales volume before and after the store launches across tracts (i.e., dashed lines) showed that sales growth before retail expansion is rather flat, suggesting the chosen set of covariates is sufficiently effective in accounting for profitability potential of the tract locations, reduces the concern of unobserved tract effects at play in affecting the purchase patterns of consumers. We notice that some of the coefficient estimates become positive and significant in the period after stores are launched in treated tracts. More coefficients are significant after the stores are opened for over a month, which goes to show that the positive impact of physical stores do not surface immediately. This makes intuitive sense, given that it takes time for customers to gain awareness of new stores and to take time to make store visits. The fact that significant and positive impacts of physical stores are strictly restricted to the post-treatment period further assures us that a pre-treatment trend is absent and that the parallel trend assumption is satisfied. We also note that the positive impact on sales in the post treatment extends beyond six months, which we interpret as a sign that the complementary impact of physical stores is not simply a novelty effect that diminishes over time.

4.5.2 Robustness Checks

We conduct a series of robustness checks to validate the consistency of the main effect under different assumptions and conditions. In particular, we examine whether the results are robust with respect to a finer unit of analysis, alternative definitions of sale tracts and outcome variables, and accounting for additional external factors that might have driven online sales.

First, we conduct our analysis at the household level, wherein we examine how the volume of online purchases of each household change after a new store is opened nearby. In our analysis, we also consider households that are located in those matched tracts to address potential endogeneity concerns. Results of various models are displayed in panel A of Table 4.4. All of these results show a significant increase in online purchases at the household level, after offline stores are launched in the neighborhood. Based on these results, it is likely that our results are robust towards endogeneity concerns related to unobserved household level factors which we explicitly account for using a finer level of analysis.

For the second set of robustness check, we assess whether the main results are sensitive to the size of sale tract defined. Instead of assuming each store's reach to be effective in serving a distance of five kilometers, we attempted narrower and broader cutoff thresholds. In particular, we set the thresholds of 2 and 10 km to arrive at alternative sales tracts.[6] The results from these two extra thresholds are shown in panel B and C of Table 4.4. We can see that the effect of physical stores on online purchases remains positive and significant across different specifications and samples.

In our third robustness check, we adopt (Wang and Goldfarb 2017)'s definition of sale tracts by which we treat each existing store as the center of a tract, and then

[6] A 2 km scope can be seen as a store that is of walking distance, and a 10 km scope is able to serve consumers who are within driving distance.

defining its service area by including households that are within a 5 km radius of the store. For cases where two stores are less than 5km to each other, we would combine these stores as a single tract when they are less than 2km apart. If they are more than 2 km apart, we divide the area between them evenly to form two tracts.[7] Results based on this alternative tract definition are shown in panel D of Table 4.4. DID estimates are consistent with our main results in that they remain positive and significant.

In our fourth robustness check, we test the sensitivity of the main results to an alternative measurement of online sales. While the earlier results inform us that online sale volume has increase as a result of store openings, it does not inform us on whether the dollar value of these sales have also increased. It is plausible that the increased amount of online purchase may be driven by the greater purchase level of less expensive products. To assess for this possibility, we use the dollar value of purchases as an alternative outcome variable. Panel E of Table 4.4 shows that the DID estimate remains positive and significant when dollar amount is used as the dependent variable. In particular, we find that the launch of a new store leads to an average increase of 16% in the monetary amount transacted in the online channel of the retailer. Thus, the qualitative conclusion from the results remains similar to the main results under an alternative sales measure.

Finally, we consider whether the increase in sales might be an artifact of promotional events of the retailer. It is plausible that the launch of the new stores were intentionally made to coincide with festive occasions and holidays, where Chinese consumers have habits to purchase more. Based on the annual online and offline promotion plans provided by the retailer (i.e., 2015 Chinese New Year, the retailer's anniversary, Singles' Day:November 11), we remove observations from the weeks when major online promotions were held. Regression results of this check is documented in panel F of Table

[7]Only a few cases fall into this category in our dataset. Most stores are geographically well dispersed with little overlaps.

Table 4.4: Robustness Checks.

	OLS (logged)		FixedEffect	
	(1) Unmatched	(2) Matched	(3) Unmatched	(4) Matched
(A) Household Purchases	0.05***	0.03*		
No.obs = 110451/45876	(0.02)	(0.02)		
unmatched/matched			0.05***	0.03*
			(0.02)	(0.02)
(B) Service Radius=2km	0.26***	0.22***		
No.obs = 36384/4787	(0.07)	(0.07)		
unmatched/matched			0.24***	0.23***
			(0.07)	(0.08)
(C) Service Radius=10km	0.22***	0.22***		
No.obs = 21308/8974	(0.06)	(0.07)		
unmatched/matched			0.20***	0.18***
			(0.06)	(0.07)
(D) Alternative Tract Definition	0.32***	0.24***		
No.obs = 12918/5165	(0.06)	(0.07)		
unmatched/matched			0.30***	0.19***
			(0.06)	(0.07)
(E) Sales in Dollar Values	0.22***	0.15***		
No.obs = 29040/6395	(0.05)	(0.06)		
unmatched/matched			0.23***	0.15***
			(0.04)	(0.05)
(F) Removing Promotion Effect	0.24***	0.22***		
No.obs = 26626/5875	(0.05)	(0.07)		
unmatched/matched			0.25***	0.21***
			(0.06)	(0.07)

*significant at 10%;**significant at 5%;***significant at 1%

4.4, The coefficients from different models are positive and significant, after accounting the promotional effect. This test rules out the possibility that the observed increased in sales is driven by promotional efforts of the retailer, but is largely due to a true complementary impact from the physical store.

4.5.3 Product level Analysis

The exposure of showcased products in physical stores is a plausible mechanism of inducing greater level of online purchases. If evidence of this mechanism can be empirically verified, the causal link between offline stores and online purchases would be strengthened. As described earlier, we rely on a DDD framework that exploit the variation in product showcasing across tract and time to identify this effect. Our analysis results are reported in Table 4.5. We find that online purchases of those showcased products significantly increase after they are being showcased in offline stores. The findings are consistent across different model specifications, and with the inclusion of product-tract and product-time fixed effects.

As an additional check, we consider all products carried by the retailer to assess if the launch of a store affects the online sales of showcased and non-showcased products in Table 4.6. Results show that a newly opened store significantly affects the amount of online purchase of showcased products (Model 1), but not so for products that are not showcased (Model 2). This is another evidence showing that when products are displayed in stores, customers living in the vicinity of these stores are more likely to purchase those products online. The absence of a sales increase of products that are not showcased goes to show that the main mechanism through which the positive impact of physical stores work is through the exposure of products to consumers in stores.

4.6 Mechanisms and Heterogeneous Effects

4.6.1 Conspicuous and Experiential Roles of Stores

Thus far, we have seen evidence of a positive, complementary effect of physical stores on online sales, particularly through the exposure of products to consumers in stores.

Table 4.5: Effect of Product Showcase on Online Sales (Volume)

	Sample1			Sample2			Sample3		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Showcase	0.03***	0.03***	0.03***	0.05***	0.03***	0.04***	0.01**	0.01**	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Tract fixed effect	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Week fixed effect	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No
Product fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tract-Product fixed effect	No	Yes	No	No	Yes	No	No	Yes	No
Product-Week fixed effect	No	No	Yes	No	No	Yes	No	No	Yes
Observations	246992	246992	246992	265912	265912	265912	235812	235812	235812
R^2	0.04	0.11	0.18	0.06	0.20	0.18	0.03	0.10	0.12

*significant at 10%;**significant at 5%;***significant at 1%

Table 4.6: Effect of Offline Showcase on Online Sales (Volume)

	OLS(logged)		FixedEffect	
	(1)	(2)	(3)	(4)
	Showcased	NotShowcased	Showcased	NotShowcased
$TreatedTracts_i*$	0.22***	0.02		
$Aftertreatment_t$	(0.06)	(0.06)		
$TreatedTracts_i$	1.42***	−1.53***		
	(0.38)	(0.44)		
$Aftertreatment_t$	−0.41***	−0.20**		
	(0.06)	(0.09)		
Treated			0.18***	0.04
			(0.06)	(0.06)
Observations	6395	6395	6395	6395
R^2	0.57	0.30	0.44	0.39

*significant at 10%;**significant at 5%;***significant at 1%

Despite these findings, the finer mechanisms that drive this positive relationship remains unclear. As discussed in our literature review, there are two main ways in which physical store can facilitate more online purchases, namely through the conspicuous and experiential characteristics of stores. The purchase of high involvement products tend to require a more comprehensive evaluation as they are costlier and have a larger

number of attributes to consider (Hansen 1985). Here, the conspicuous aspect of stores would be helpful in the conversion process as it allows consumers to physically experience products and getting advice from salespersons about the products. On the other hand, the purchase of low involvement products involves lower risks, and tend to be largely driven by top-of-mind awareness of products that can be built up by product exposure facilitated through the experiential aspect of stores. Thus, by examining the types of products that experienced an increase in online sales in tracts that experience the launch of new stores, we are able to identifying which roles are at work in the purchase conversion process.

To do so, we first partition all products into high and low involvement types. We rely on two different methods to segment the products so that various definitions of product involvement are considered. We begin by using a simplistic approach of categorizing products based on their prices. Two thresholds were utilized: the the 10th percentile (representing products that cost $250 or less) and 25th percentile (representing products that cost $650 or less). Products that cost more than these threshold values are deemed as high involvement products, while those that are below the thresholds are classified as low involvement products. We also adopt a clustering approach to segment the products. Specifically, we tabulated product features including price and purchase volume. Based on these features, a K-means clustering method is used to group all products into two clusters[8]. A visualization of the clustering result in a two dimensional space involving *purchase volume* and *price* is shown in Fig. 4.5. From this figure we could see that, in general, products with higher prices and lower purchase volume (i.e., orange stars) form one cluster which is indicative of high involvement products, while those with relatively lower prices and higher purchase volume (i.e., blue dots) form the other cluster consisting

[8]We tried different values for the number of clusters, and plotted the line graph of Sum of Squared Errors of each value. The checks reveal that optimal number of clusters is 2 (i.e., at the elbow point of the graph), which is in line with our apriori expectation that the price and purchase volume would produce a natural split between high and low involvement products.

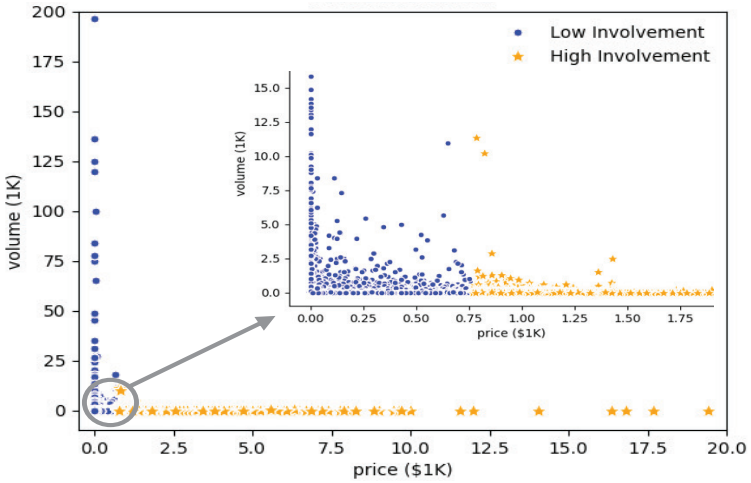


Figure 4.5: Discovered Product Clusters.

of low involvement products. After these product types are identified, we run separate regressions on each of these product types.

The results of our analysis are shown in Table 4.7. Under a price-based classification (i.e., first two rows), we see that DID estimator for all products and showcased products are positive and significant, for both the high and low involvement products. The results are qualitatively similar under a more elaborate classification scheme that considers multiple product attributes (i.e., third row). These results provide initial evidence that both the conspicuous and experiential aspects of physical stores are at work in driving online purchases.

The classification the products into two broad high-low categories can be rigid given that there might nuances underlying each product that can influence consumer's consideration in their purchase process. To gain deeper insights on how physical stores may influence online purchases of different products, we adopt a more fine-grained categorization by splitting products by store categories provided by the retailer. Through this categorization, there are eight major product categories, including laundry machines,

Table 4.7: Impact of Stores on Online Sales (Split by Product Involvement)

	High Involvement		Low Involvement	
	(1) All	(2) Showcased	(3) All	(4) Showcased
Store Opening (10th percentile price-based classification)	0.10* (0.05)	0.14** (0.05)	0.22*** (0.07)	0.21*** (0.06)
Store Opening (25th percentile price-based classification)	0.15** (0.06)	0.20*** (0.07)	0.21*** (0.07)	0.20*** (0.06)
Store Opening (Clustering-based classification)	0.12** (0.05)	0.20*** (0.06)	0.22*** (0.07)	0.20*** (0.06)
Observations	6395	6395	6395	6395

*significant at 10%;**significant at 5%;***significant at 1%

TVs, kitchen-based electronics, cell phones, kitchenware, household supplies, skin care products, and food and drinks. We conduct the same analysis on the sales volume of these product categories.

The results are shown in Table 4.8. The results in this analysis generally agree with our earlier intuition. Products that require greater efforts in information collection via the physical examination and communications with salesperson (i.e., laundry machines, TVs, kitchen electronics and cell phones) are enjoying greater online purchase volumes in tracts that experience the launch of new stores. The more expensive items, such as laundry machines and TVs, experience an increase of about 15-17%, and the relatively less costly items such as kitchen electronics and cellphones see an increase of about 18-24%. This is reasonable given that the purchase decision process of the former two categories of products are likely to involve greater deliberation and consideration compared to the latter two, given that they are "riskier" purchases with greater costs involved. The remaining product categories, i.e., kitchenware, household supplies, skin care products, and food and drinks, are largely considered low involvement products

Table 4.8: Effect of Stores on Online Sales (Fine-grained Categorization)

	LaundaryMachine		TV		KitchenElectronics		Cellphone	
	(1) All	(2) Showcased	(3) All	(4) Showcased	(5) All	(6) Showcased	(7) All	(8) Showcased
Store Opening	0.12***	0.14***	0.12*	0.16***	0.03	0.17***	0.17***	0.24***
	(0.03)	(0.03)	(0.06)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
	Kitchenware		HouseholdSupplies		SkinCare		FoodDrinks	
	(9) All	(10) Showcased	(11) All	(12) Showcased	(13) All	(14) Showcased	(15) All	(16) Showcased
Store Opening	0.08*	0.08**	0.16**	0.17***	0.05	0.07	0.24***	0.25***
	(0.04)	(0.04)	(0.07)	(0.05)	(0.06)	(0.05)	(0.09)	(0.07)
Observations	6395	6395	6395	6395	6395	6395	6395	6395

*significant at 10%;**significant at 5%;***significant at 1%

that are bought with greater frequency and are generally of lower cost compared to the previous set of products. We find the online purchase of these items are heightened in treated tracts that experience the launch of physical stores, with the exception of skin care products. The lack of an effect for skin care products seems reasonable when we consider that the retailer is not known for carrying skin care products and they indeed dedicate a smaller proportion of store space for skin care products (less than 3% are for skin care products) relative to other products.

4.6.2 Impact in Different Territories

It is possible that the complementary effect of offline stores on online sales may be different across locations. In particular, we are interested in seeing if the effect of stores are constrained to virgin tracts that are not already served by an existing store from the retailer. To investigate this, we split the tracts into different two main categories, namely virgin territories and territories with existing stores. Results are displayed in Table 4.9. We first report that estimates for all locations in columns (1) and (2) for comparison. Columns (3) and (4) show the impact of new stores that are opened in

virgin locations, and Columns (5) and (6) show that for new stores opened in locations without existing stores. Across both types of territories, new stores generate positive and significant impacts on online sales. While we have contrasted the impact of new stores opening up in non-virgin tracts with tracts that do not have the retailer's presence in Columns (5) and (6), it is unclear if the effect of an additional store would result in an additional improvement in online sales when compared to tracts that are served by one operating physical store. We repeat the same analysis by utilizing non-virgin locations that did not experience the launch of new stores as the control tracts to see if there are incremental effects. The results in Columns (6) and (7) show a positive and significant impact on online sales, albeit smaller in magnitude compared to the previous estimates.

It is interesting to note that the positive effect of physical stores are not limited to the virgin territories but also in tracts that already have existing retailer presence. Here, our results are slightly different from that of past work by Wang and Goldfarb (2017), which did not detect an effect on online sales when stores are opened in non-Virgin locations. We interpret this set of results as supporting our earlier finding that stores in the Chinese retail context serve a conspicuous role, on top of an experiential role, in the conversion of online purchases. Given the relative larger population density in Chinese cities, additional stores within a sales tract can help to serve a larger base of customers at the same time. This aspect is likely to be important in the Chinese market, as consumers have a greater desire to seek out product information in stores before making purchase.

4.6.3 Impact on Customer Types

Next, we explore if the new stores have varying impacts on different customer types. The focal retailer is one of the largest retailers in China and is well known by the

Table 4.9: Effect of Stores Opened in Different Types of Locations

	All Locations		Virgin Locations		With Existing Stores		With Existing Stores (subset)	
	(1) Unmatched	(2) Matched	(3) Unmatched	(4) Matched	(5) Unmatched	(6) Matched	(7) Unmatched	(8) Matched
$TreatedTracts_i$*	0.25***	0.23***	0.29***	0.30**	0.15***	0.13*	0.22***	0.14*
$Aftertreatment_t$	(0.06)	(0.07)	(0.08)	(0.12)	(0.06)	(0.07)	(0.06)	(0.07)
$TreatedTracts_i$	−0.07	−0.04	−0.16*	−0.04	−0.07	−0.01	−0.11**	−0.05
	(0.05)	(0.06)	(0.08)	(0.11)	(0.05)	(0.05)	(0.05)	(0.05)
$Aftertreatment_t$	−0.32***	−0.41***	−0.24***	−0.37**	−0.31***	−0.40***	−0.45***	−0.45***
	(0.04)	(0.09)	(0.06)	(0.16)	(0.05)	(0.09)	(0.06)	(0.08)
Dist2cen (km)	0.05	0.07	−0.04	0.05	0.04	0.00	0.06	−0.01
	(0.05)	(0.08)	(0.07)	(0.20)	(0.05)	(0.08)	(0.06)	(0.07)
Userbase	0.29***	0.14	0.36***	0.17	0.29***	0.10	0.24***	0.12
	(0.05)	(0.09)	(0.06)	(0.13)	(0.05)	(0.11)	(0.06)	(0.11)
HousingPrice ($)	0.07	0.06	0.08	0.11	0.06	−0.01	−0.14*	−0.08
	(0.05)	(0.10)	(0.06)	(0.15)	(0.05)	(0.12)	(0.07)	(0.09)
Appliances	0.12***	0.11***	0.15**	0.11	0.11***	0.07*	0.13***	0.03
	(0.04)	(0.04)	(0.06)	(0.08)	(0.04)	(0.04)	(0.05)	(0.02)
Electronics	−0.01	−0.02	−0.05	−0.07	0.02	0.02	0.07***	0.06*
	(0.03)	(0.03)	(0.06)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
Houseware	0.19***	0.45***	0.29**	0.54***	0.16***	0.31**	0.17**	0.23
	(0.06)	(0.12)	(0.12)	(0.12)	(0.06)	(0.13)	(0.08)	(0.14)
FMCG	0.53***	0.41***	0.50***	0.39***	0.55***	0.54***	0.45***	0.58***
	(0.04)	(0.06)	(0.05)	(0.08)	(0.04)	(0.07)	(0.06)	(0.08)
Observations	29040	6395	18084	3192	27078	4433	10956	3203
R^2	0.87	0.87	0.78	0.79	0.87	0.90	0.78	0.89

*significant at 10%;**significant at 5%;***significant at 1%

Chinese consumers, which makes it less relevant to consider a "new vs old" dichotomy of customers. Indeed, an examination of the customer IDs in our dataset reveals that there are very few instances of customers who were new customers in our study period. Extant literature has been shown that the hiatus heuristic, based on whether customers have made a purchase within a set time period, works well in distinguishing customer groups in practice (Mumford 1995, Wübben and Wangenheim 2008). Guided by this categorization scheme, we distinguish customers based on their activity level, which we define as whether they are made a purchase from the retailer in the last six months (Gigerenzer and Gaissmaier 2011). Specifically, *active customers* purchase from our focal retailer in the last six months, while *inactive customers* did not. For these two groups of customers, we look at the impact of store openings on different outcomes

Table 4.10: Impact of Store Opening on Online Sales (Split by Customer Group)

	Inactive Customers (last purchase before 6 months)				Active Customers (last purchase within 6 months)			
	(1) Number	(2) Order_All	(3) Order_High	(4) Order_Low	(5) Number	(6) Order_All	(7) Order_High	(8) Order_Low
Store Opening	0.04*	0.17**	0.03	0.17**	0.10**	0.22***	0.11**	0.21***
	(0.02)	(0.08)	(0.03)	(0.08)	(0.04)	(0.07)	(0.05)	(0.07)
Observations	5753	5692	6323	5719	6395	6395	6395	6395

*significant at 10%;**significant at 5%;***significant at 1%

including the number of customers who made purchases, online sales volume, and online sales of high- and low-involvement products.

The results are reported in Table 4.10. From the results, we see that inactive customers are making online purchases after a new store is opened in their associated sales tract (Column (1)). Specifically, approximately 4% inactive customers are becoming active shoppers via online purchases. We also find that these customers are increasing their online sale volume (Column (2)) by 18.5%. Interestingly, this increase is largely attributed to the purchase of low involvement products (Columns (3) and (4)). We take this as a sign of inactive customers re-bonding with the retailer by making purchases involving less complex decisions. For active customers, our results show that following the addition of a new store, the number of online shoppers grows significantly, i.e., approximately 10% more active customers (Column (5)) are making online purchases, which accounts for roughly 25% more online sales (Column (6)). In contrast to inactive customers, active customers are purchasing both high and low involvement products (Columns (7) and (8)). Given that active customers are more loyal to begin with, the presence of a nearby store can help to further supports their purchase process, thereby generating a greater amount of online sales conversion for both types of products.

Finally, we conduct an analysis that considers both location and customer types on the online purchase of different product types in response to physical store openings.

The results are shown in Table 4.11. We see that the results for virgin locations are quite similar to the results in Table 4.10, in that inactive customers purchase significantly more low involvement products online (column (3)), and active customers are buying more of both types of products online, after a new store is opened in their tract (Columns (1) to (6)). In the case of store openings in tracts with existing stores, inactive customers remain inactive, while active customers are purchasing more high involvement products. Given that existing stores are not already inducing online purchases for inactive customers, it is reasonable to see that having an additional store would not be helpful in moving these consumers towards more online purchase. New stores in non-virgin territories work in a more specialized fashion of providing more opportunities for serving the active customers living in the vicinity. These additional physical stores serve as additional avenues to which customers are able to solicit product information through physical examination and question-asking. This set of results is consistent with our previous conjecture that additional stores is helpful in serving the densely populated customer base in Chinese retail market.

4.7 Implications and Future Research

The advent of digital commerce has spurred dramatic changes in consumers' shopping behaviors and the landscape of retailing industry in the past two decades. One of the most significant impacts is the growing prominence of online retailing shifting the value placed on physical stores. To adapt to the change in consumer purchase behaviors, many traditional retailers have adopted multi-channel strategies which involve adding an online channel to their existing set of offline touchpoints, and managing customers across channels. As the retailing industry progresses further in the digital age, the question of whether physical stores continue to matter and how they generate value becomes important to retailers who are competing for market share in this industry. In

Table 4.11: Customer Maintenance in Different Types of Locations

	Virgin Locations					
	Inactive Customers (last purchase before 6 months)			Active Customers (last purchase within 6 months)		
	(1) Order_All	(2) Order_High	(3) Order_Low	(4) Order_All	(5) Order_High	(6) Order_Low
Store Opening	0.23**	0.01	0.22*	0.29**	0.15***	0.28**
	(0.11)	(0.01)	(0.11)	(0.12)	(0.05)	(0.12)
Observations	3192	3192	3192	3192	3192	3192
	Locations with Existing Stores					
	Inactive Customers (last purchase before 6 months)			Active Customers (last purchase within 6 months)		
	(7) Order_All	(8) Order_High	(9) Order_Low	(10) Order_All	(11) Order_High	(12) Order_Low
Store Opening	0.22	0.07	0.21	0.12*	0.17**	0.09
	(0.14)	(0.07)	(0.14)	(0.07)	(0.08)	(0.07)
Observations	4433	4433	4433	4433	4433	4433

*significant at 10%;**significant at 5%;***significant at 1%

this work, we provide an updated view of the impact of offline store on retailers' online business in the under-studied Chinese retail market. Guided by the framework of Avery et al. (2012), we investigate the effects of physical stores on consumers' online purchases, under a quasi-experimental framework that involves variation in the launch of new stores across tracts and time. We augment this analysis framework by exploiting an additional layer of variation in product showcasing across stores and weeks in an effort to establish stronger evidence of the link between physical stores and online purchase.

Our results reveal that the launch of a new offline store enhances online weekly purchase by 26% per tract, on average. This positive impact appears to be long standing, which does not diminish over time. The increased amount of online purchases is economically substantial, representing approximately an increase of US$2 million in annual sales. Product level analysis reveals that the online purchase of the same product significantly increases in tracts where they are showcased, but not in tracts where they

are not showcased. We further find that the complementary impact of physical stores increases both high and low involvement products, which we interpret as evidence supporting the conspicuous and experiential role of stores at work in the conversion process. Our analysis indicate that the positive impact of stores on online sales manifest in both locations with or without existing stores, though a larger effect is observed in virgin territories. Finally, the impact of offline stores exert heterogeneous effects on different customers, inducing inactive customers to purchase more low involvement products and active customers to purchase more of high and low involvement products. Overall, our empirical results provide evidence that offline channels can have complementary impacts on online channels in the Chinese context, and marketing efforts should be coordinated to greater cross-channel synergies.

Our main results depicting a complementary offline-to-online effect is generally aligned with past works, including Avery et al. (2012), Wang and Goldfarb (2017), Bell et al. (2017) and Kumar et al. (2019). The distinct addition of our work to this literature is two-folds. Past works have mainly found evidence of the experiential influence of traditional retail stores at work in the Western market (e.g., Avery et al. (2012), Wang and Goldfarb (2017), but less comprehensive evidence of the effects of the conspicuous role of stores. Our work speaks to this topic by showing the simultaneous presence of both conspicuous and experiential roles of stores at work when a broader assortment of products are considered. At the same time, we proposed theoretical linkages of the Oriental culture that might influence the conspicuous and experiential roles of offline stores on the purchase behavior of Chinese consumers. We find that some of these theoretical propositions are helpful in explaining why the conspicuous aspect of stores is also found to be present in our study context, on top of the effects emanating from the experiential role of stores. The second contribution we provide pertains to providing finer empirical evidence that the positive impact of stores on online purchase. We do so by drawing

a connection between the products that were exposed to consumers in close-by stores, with the same products that the same set of consumers purchase online. This evidence circumvents criticisms of DID evidence based on geographical and temporal variation, thereby validating the conclusions of a positive impact of stores found in past works.

Our study provides several managerial insights for practitioners. First, our findings show that offline stores continue to play an important role in the digital age, and should still be given serious consideration by retailers to complement the online channel. This is especially in the Chinese retail scene. The confidence in Chinese market is also reflected in Lego's recent decision to double the number of shops in China in 2019 (Gronholt-Pedersen 2019). Second, our results indicate that offline stores enhances both the online sales of high and low involvement products, by which in-store product placement decisions should take into consideration. While high involvement products generate higher per-unit profit margin, these items constitute a lower purchase frequency compared to the low involvement counterparts. In certain cases, large purchase volume of items with small profit margins may sometimes be more helpful in generating profits compared to the infrequent sales of high-margin items. Retailers would need to balance these considerations when deciding how much store space to allocate to high and low involvement products that they carry. Finally, our study results find that offline stores are an important vehicle for maintaining customer relationship, in that it spurs inactive customers into making purchases with the retailer. Marketing strategies that allow for customer retention is of crucial value to marketers in the increasingly competitive retail landscape of retail, as consumers can easily contrast product offerings across providers via a quick online search.

There are a few limitations of this research. First, our result is based on one retailer in China by which caution is warranted when drawing conclusions to broader

contexts. Given that the company considered in our study is a well-known brick-and-mortar retailer, results that we get may not apply to other retailers that do not enjoy this brand recognition. Furthermore, due to lack of data on retailer's competitors, we are unable to draw any inferences related to the competitive dynamics between retailers and whether the focal retailer experiences the negative aspects of "showrooming" (i.e., shoppers browsing products in one retailer's physical store and making their purchase online from another retailer). Second, while we have significantly improved the empirical identification of the estimation of the effect of physical stores by conducting a store-product analysis, we still do not directly observe if a certain individual saw an interacted with a product in store before making a purchase of the same product online. We leave this as a research topic for future work, which will involve the use of video analytics and facial recognition techniques to capture and identify customers in stores and their interest in products during their shopping trips. This information when matched to online identities and online purchase behaviors will be instrumental in generating an even tighter identification of the effects of physical stores. Finally, our work is unable to directly assess and contrast the difference in effects across physical stores in different countries. Ideally, we would want to observe the same retailer operating in Western and Oriental countries and see how the launch of new stores might shift online purchase behaviors across cultures, so that we can pinpoint finer mechanisms that underlie behaviors in these geographies. Notwithstanding these limitations, our study documents a substantial economic impact of the offline stores on online purchase, using fine-tune data from a major Chinese retailer. We hope that our work would spur more research in this domain.

Chapter 5

Concluding Remarks

My thesis examined data-driven decision making problems encountered in contexts where multiple perspectives are required to inform intelligent data-driven decisions. While various data-driven analytics methods have been applied in numerous application domains to achieve efficient and effective decision making, adapting existing techniques to deal with the multifaceted nature of practical problems remains a significant challenge. This thesis takes a step further in addressing this issue by proposing a multi-perspective view for data-driven model adaptation and application. Specifically, in three essays, I respectively investigated: (1) the solutions to balance aspects of *accuracy* and *long tail recommendation* in real-world recommender systems; (2) the provision of complementary information, i.e., individual prediction reliability, for outcome prediction models for more nuanced application and better decision support; (3) the complementary role of offline channels for product information collection and consumer purchase decision support in e-commerce platform.

My thesis research contributes new methodologies and empirical understanding to decision support literature in the field of Information Systems. In the first essay, we

discussed the necessity of evaluating recommendation techniques beyond accuracy to address the *popularity bias* in traditional recommender systems. To improve the quality of recommendations on multiple aspects, we proposed a new method that can capture the key relationship among niche items for relevant long tail recommendations. Comprehensive experimental results compare the proposed approach to a wide variety of classical, widely-used recommendation algorithms and demonstrate its practical benefits in accuracy, flexibility, and scalability, in addition to the superior long-tail recommendation.

While the first essay adapts existing methods directly to balance multiple perspectives of predictive model quality for better decision support, the second essay proposed to provide an additional aspect of information, i.e., individual prediction reliability (or confidence, uncertainty), to complement aggregate predictive accuracy for the purpose of more nuanced decision support. A general machine learning-based framework is also designed to provide intuitive and highly interpretable prediction reliability estimation. Extensive experimental results on multiple real-world datasets show that the proposed machine-learning-based approach can significantly improve individual prediction reliability estimation as compared to a number of baselines from prior work, especially in more complex predictive scenarios.

Finally, the third essay investigated the role of retailers' offline channel (e.g., physical stores) in consumers' online purchase decision making using a quasi-experiment, taking place through a nationwide retailer that expanded its physical presence during the study period. Through a "triple-differences" framework, we provide a more direct data-driven evidence on the effect of the physical channel on consumers' online purchases. We also find that the *product showcase* process can significantly increase online transactions, suggesting the imperative function of offline channel in providing complementary information for consumer purchase decision support.

Findings from my thesis provide several managerial implications, both for various information systems that adopt data-driven methodologies to provide decision support and any businesses that use online platforms to build digital presence and facilitate online transactions. For example, it is imperative to be cognizant of the multifaceted nature of decision making problems in real-world applications. Decision support technologies failing to take into account multiple perspectives can lead to unintended consequences. For example, *popularity bias* occurs when recommendation techniques are accuracy-oriented, leaving out other aspects of recommendation evaluation. Incorporating multiple perspectives is key to the design of intelligent decision support and, thus, the viability and sustainability of the information system. In addition, platform or business designers can consider several ways to account for multiple perspectives for decision support, such as by directly adapting existing methods to balance various aspects or by integrating complementary sources of information. Further exploration of these insights and of new ways to facilitate multi-perspective decision making represent a problem-rich set of directions for future research.

References

Abdollahpouri H, Mansoury M, Burke R, Mobasher B (2019) The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* .

Ackerman D, Tellis G (2001) Can culture affect prices? a cross-cultural study of shopping and retail prices. *Journal of Retailing* 77(1):57–82.

Adamopoulos P, Tuzhilin A (2015) On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology* 5(4):54:1–54:32.

Addady M (2016) Amazon to open 2,000 grocery stores across the U.S. *Fortune, October* .

Adomavicius G, Kwon Y (2012) Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24(5):896–911.

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17(6):734–749.

Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.*, 207–216.

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Management of Data, VLDB*, 487–499.

Alba JW, Chattopadhyay A (1986) Salience effects in brand recall. *Journal of Marketing Research* 363–369.

Alshammari G, Jorro-Aragoneses JL, Kapetanakis S, Petridis M, Recio-García JA, Díaz-Agudo B (2017) A hybrid cbr approach for the long tail problem in recommender systems. *Proceedings of International Conference on Case-Based Reasoning*, 35–45 (Springer).

Anderson C (2006) *The long tail: Why the future of business is selling less of more* (Hachette Books).

Angrist JD, Pischke JS (2008) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. *Journal of Marketing Research* 45(1):60–76.

Avery J, Steenburgh TJ, Deighton J, Caravella M (2012) Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. *Journal of Marketing* 76(3):96–111.

Balasubramanian S, Raghunathan R, Mahajan V (2005) Consumers in a multichannel environment: Product utility, process utility, and channel choice. *Journal of Interactive Marketing* 19(2):12–30.

Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif. L. Rev.* 104:671.

Baumol WJ, Ide EA (1956) Variety in retailing. *Management Science* 3(1):93–101.

Bell DR, Gallino S, Moreno A (2014) The store is dead long live the store. *Sloan Management Review* .

Bell DR, Gallino S, Moreno A (2017) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science* 64(4):1629–1651.

Bensinger G (2014) Amazon to open first brick-and-mortar site. *Wall Street Journal* .

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119(1):249–275.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Bose N (2017) U.S. online retail sales likely to surpass $1 trillion by 2027: FTI. *Reuters* .

Bosnić Z, Kononenko I (2008a) Comparison of approaches for estimating reliability of individual regression predictions. *Data and Knowledge Engineering* 67(3):504–516.

Bosnić Z, Kononenko I (2008b) Estimation of individual prediction reliability using the local sensitivity analysis. *Applied intelligence* 29(3):187–203.

Bosnić Z, Kononenko I (2009) An overview of advances in reliability estimation of individual predictions in machine learning. *Intelligent Data Analysis* 13(2):385–401.

Breiman L (1996) Bagging predictors. *Machine learning* 24(2):123–140.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly weather review* 78(1):1–3.

Briesemeister S, Rahnenführer J, Kohlbacher O (2012) No longer confidential: estimating the confidence of individual regression predictions. *PloS one* 7(11):e48723.

Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: Generalizing association rules to correlations. *Proceedings ACM SIGMOD International Conference on Management of Data*, 265–276.

Brynjolfsson E, Hitt LM, Kim HH (2011a) Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486* .

Brynjolfsson E, Hu Y, Rahman MS (2009) Battle of the retail channels: How product selection and geography drive cross-channel competition. *Management Science* 55(11):1755–1765.

Brynjolfsson E, Hu Y, Simester D (2011b) Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* 57(8):1373–1386.

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49(11):1580–1596.

Brynjolfsson E, Hu YJ, Rahman MS (2013) *Competing in the age of omnichannel retailing* (MIT).

Brynjolfsson E, Hu YJ, Smith MD (2006) From niches to riches: Anatomy of the long tail. *Sloan Management Review* 47(4):67–71.

Buskirk EV (2016) The most streamed music from spotify discover weekly. URL `https://insights.spotify.com/no/2016/07/07/top-music-discover-weekly/`.

Cadell C (2017) Amazon's grocery push playing catch up with chinese e-commerce giants. *Reuters* .

Carney JG, Cunningham P, Bhagwan U (1999) Confidence and prediction intervals for neural network ensembles. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 2, 1215–1218 (IEEE).

Castells P, Hurley NJ, Vargas S (2015) Novelty and diversity in recommender systems. *Recommender Systems Handbook*, 881–918 (Springer).

Ceglar A, Roddick JF (2006) Association mining. *ACM Computing Surveys* 38(2):5.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Ching AT, Ishihara M (2012) Measuring the informative and persuasive roles of detailing on prescribing decisions. *Management Science* 58(7):1374–1387.

Choi E, Schuetz A, Stewart WF, Sun J (2017) Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* 24(2):361–370.

Choi J, Bell DR (2011) Preference minorities and the internet. *Journal of Marketing Research* 48(4):670–682.

Clark RD (2009) Dpress: localizing estimates of predictive uncertainty. *Journal of Cheminformatics* 1(1):11.

Clarke K, Belk RW (1979) The effects of product involvement and task definition on anticipated consumer effort. *ACR North American Advances* .

Collins A, Tkaczyk D, Beel J (2018) One-at-a-time: a meta-learning recommender-system for recommendation-algorithm selection on micro level. *arXiv preprint arXiv:1805.12118* .

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) *Introduction to Algorithms* (MIT press).

Cortés-Ciriano I, Bender A (2018) Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *Journal of chemical information and modeling* 59(3):1269–1281.

Craw S, Horsburgh B, Massie S (2015) Music recommendation: audio neighbourhoods to discover music in the long tail. *Proceedings of International Conference on Case-Based Reasoning*, 73–87 (Springer).

Dash R, Dash PK, Bisoi R (2015) A differential harmony search based hybrid interval type2 fuzzy egarch model for stock market volatility prediction. *International Journal of Approximate Reasoning* 59:81–104.

Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015(1):92–112.

Davidson J, Liebald B, Liu J, Nandy P, Vleet TV, Gargi U, Gupta S, He Y, Lambert M, Livingston B, Sampath D (2010) The youtube video recommendation system. *Proceedings of the 2010 ACM Conference on Recommender Systems*, 293–296.

Demut IR (2010) Reliability of predictions in regression models. *Doktorandske dny'10* .

Domingos P (2000) A unified bias-variance decomposition. *Proceedings of 17th International Conference on Machine Learning*, 231–238.

Efron B (1992) Bootstrap methods: another look at the jackknife. *Breakthroughs in statistics*, 569–593 (Springer).

Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467):619–632.

Egan M (2016) Macy's is closing another 100 stores. *CNN Money, August* 11.

Engel JF, Roger D (1995) Consumer behavior. *New York: Holt, Renehard, and Winston* .

Ettenson R, Wagner J (1991) Chinese (vs. US) consumer behavior: A cross-cultural comparison of the evaluation of retail stores. *Journal of International Consumer Marketing* 3(3):55–71.

Farquhar PH, Rao VR (1976) A balance model for evaluating subsets of multiattributed items. *Management Science* 22(5):528–539.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science* 55(5):697–712.

Fleder DM, Hosanagar K (2007) Recommender systems and their impact on sales diversity. *Proceedings of the 8th ACM Conference on Electronic Commerce*, 192–199 (ACM).

Forgy EW (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21:768–769.

Forman C, Ghose A, Goldfarb A (2009) Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management science* 55(1):47–57.

Funk S (2006) Netflix update: Try this at home. URL `http://sifter.org/~simon/journal/20061211.html`.

Gama GMC, Meira W, Carvalho ML, Guedes DO, Almeida VA (2001) Resource placement in distributed e-commerce servers. *Proceedings of IEEE Global Telecommunications Conference*, volume 3, 1677–1682.

Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. *Neural computation* 4(1):1–58.

Geyskens I, Gielens K, Dekimpe MG (2002) The market valuation of internet channel additions. *Journal of marketing* 66(2):102–119.

Ghoshal A, Sarkar S (2014) Association rules for recommendations with multiple items. *INFORMS Journal on Computing* 26(3):433–448.

Gigerenzer G, Gaissmaier W (2011) Heuristic decision making. *Annual review of psychology* 62:451–482.

Goldstein DG, Goldstein DC (2006) Profiting from the long tail. *Harvard Business Review* 84(6):24–28.

Graham RL (1969) Bounds on multiprocessing timing anomalies. *SIAM Journal on Applied Mathematics* 17(2):416–429.

Gronholt-Pedersen J (2019) Toymaker lego to open 80 new shops in china this year. *Reuters* .

Gruber A (1969) Top-of-mind awareness and share of families: An observation. *Journal of Marketing Research* 6(2):227–231.

Gu B, Park J, Konana P (2012) Research note—the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research* 23(1):182–196.

Gustafson K, Reagan C (2016) Walmart to close 269 stores as it retools fleet. *CNBC* .

Hakala K, Van Landeghem S, Salakoski T, Van de Peer Y, Ginter F (2013) Evex in st'13: Application of a large-scale text mining resource to event extraction and network construction. *Proceedings of the BioNLP Shared Task 2013 Workshop*, 26–34.

Halligan B, Shah D (2009) *Inbound Marketing.: Get Found Using Google, Social Media, and Blogs* (John Wiley & Sons).

Halperin M (1963) Confidence interval estimation in non-linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)* 25(2):330–333.

Hamedani EM, Kaedi M (2019) Recommending the long tail items through personalized diversification. *Knowledge-Based Systems* 164:348–357.

Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 1–12.

Hand DJ, Yu K (2001) Idiot's bayes—not so stupid after all? *International statistical review* 69(3):385–398.

Hansen F (1985) Involvement of interest or what? *Advances in Consumer Research* 12:257–260.

Hart MA (2007) The long tail: Why the future of business is selling less of more by chris anderson. *Journal of Product Innovation Management* 24(3):274–276.

Haviland A, Nagin DS, Rosenbaum PR (2007) Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological methods* 12(3):247.

Heskes T (1997) Practical confidence and prediction intervals. *Advances in neural information processing systems*, 176–182.

Ho SS, Wechsler H (2003) Transductive confidence machine for active learning. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 2, 1435–1440.

Hosanagar K (2019) *A Human's Guide to Machine Intelligence: How Algorithms are Shaping Our Lives and how We Can Stay in Control* (Viking).

Hosanagar K, Fleder D, Lee D, Buja A (2013) Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Science* 60(4):805–823.

Hsu FL, Marsella AJ (1985) *Culture and self: Asian and Western perspectives* (Tavistock Publications).

Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. *Proceedings of IEEE International Conference on Data Mining*, 263–272.

Huang J, Zhu L, Fan B, Chen Y, Jiang W, Li S (2018) Large-scale price interval prediction at ota sites. *IEEE Access* 6:69807–69817.

Hurley NJ (2013) Personalised ranking with diversity. *Proceedings of the 7th ACM Conference on Recommender Systems*, 379–382.

Hwang JG, Ding AA (1997) Prediction intervals for artificial neural networks. *Journal of the American Statistical Association* 92(438):748–757.

Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, McGinn T, Hayden J, Williams K, Shea B, et al. (2015) Use of grade for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *bmj* 350:h870.

Jacoby J, Mazursky D (1984) Linking brand and retailer images: Do the potential risks outweigh the potential benefits? *Journal of Retailing* 60(2):105–122.

Jannach D, Lerche L, Gedikli F, Bonnin G (2013) What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. *Proceedings of International Conference on User Modeling, Adaptation, and Personalization*, 25–37 (Springer).

Johndrow JE, Lum K, et al. (2019) An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13(1):189–220.

Kannan P, Chang AM, Whinston AB (2001) Wireless commerce: marketing issues and possibilities. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, 6–12.

Kazienko P (2009) Mining indirect association rules for web recommendation. *International Journal of Applied Mathematics and Computer Science* 19(1):165–186.

Kekre S, Srinivasan K (1990) Broader product line: a necessity to achieve success? *Management Science* 36(10):1216–1232.

Keller KL (1993) Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing* 57(1):1–22.

Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nature methods* 11(7):740–742.

Khosravi A, Nahavandi S, Creighton D (2010) Construction of optimal prediction intervals for load forecasting problems. *IEEE Transactions on Power Systems* 25(3):1496–1503.

Kim SY, Lewis VM, Wang Y (2019) The effects of a physical store opening on purchase and return behaviors: A quasi-experimental approach using the causal forest method. *Working Paper* .

Kindel TI (1985) Chinese consumer behavior: historical perspective plus an update on communication hypotheses. *Association for Consumer Research Special Volumes* 186–190.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *The quarterly journal of economics* 133(1):237–293.

Knafl G, Sacks J, Ylvisaker D (1985) Confidence bands for regression functions. *Journal of the American Statistical Association* 80(391):683–691.

Kohavi R, et al. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, volume 14, 1137–1145 (Montreal, Canada).

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.

Kresh D (2007) *The whole digital library handbook* (American Library Association).

Kukar M, Kononenko I (2002) Reliable classifications with machine learning. *European Conference on Machine Learning*, 219–231 (Springer).

Kumar A, Mehra A, Kumar S (2019) Why do stores drive online sales? Evidence of underlying mechanisms from a multichannel retailer. *Information Systems Research* 30(1):319–338.

Laurent G, Kapferer JN (1985) Measuring consumer involvement profiles. *Journal of Marketing Research* 22(1):41–53.

Lebedev A, Westman E, Van Westen G, Kramberger M, Lundervold A, Aarsland D, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, et al. (2014) Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical* 6:115–125.

Lee D, Hosanagar K (2019) How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research* 30(1):239–259.

Levy M, Bosteels K (2010) Music recommendation and the long tail. *Proceedings of the Workshop on Music Recommendation and Discovery*, 55–58.

Levy S, Nebenzahl ID (2008) The influence of product involvement on consumers' interactive processes in interactive television. *Marketing Letters* 19(1):65–77.

Lin W, Alvarez SA, Ruiz C (2002) Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery* 6(1):83–105.

Liu R, Glover KP, Feasel MG, Wallqvist A (2018) General approach to estimate error bars for quantitative structure–activity relationship predictions of molecular activity. *Journal of Chemical Information and Modeling* 58(8):1561–1575.

Markus H, Oyserman D (1989) Gender and thought: The role of the self-concept. *Gender and thought: Psychological perspectives*, 100–127 (Springer).

Mathwick C, Rigdon E (2004) Play, flow, and the online search experience. *Journal of Consumer Research* 31(2):324–332.

Matsa DA, Miller AR (2013) A female style in corporate leadership? Evidence from quotas. *American Economic Journal: Applied Economics* 5(3):136–69.

Melluish T, Saunders C, Nouretdinov I, Vovk V (2001) Comparing the bayes and typicalness frameworks. *European Conference on Machine Learning*, 360–371 (Springer).

Meyer BD (1995) Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 13(2):151–161.

Mobasher B, Dai H, Luo T, Nakagawa M (2001) Effective personalization based on association rule discovery from web usage data. *Proceedings of the 3rd International Workshop on Web Information and Data Management*, 9–15 (ACM).

Mumford A (1995) Intuition: The new frontiers of management. *Industrial and Commercial Training* 27(11).

Murakami T, Mori K, Orihara R (2007) Metrics for evaluating the serendipity of recommendation lists. *Proceedings of Annual Conference of the Japanese Society for Artificial Intelligence*, 40–46 (Springer).

Nakagawa M, Mobasher B (2003) A hybrid web personalization model based on site connectivity. *Proceedings of WebKDD*, 59–70.

Narang U, Shankar V (2019) Mobile app introduction and online and offline purchases and product returns. *Marketing Science* .

Niemann K, Wolpers M (2013) A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 955–963.

Nouretdinov I, Melluish T, Vovk V (2001) Ridge regression confidence machine. *ICML*, 385–392.

Obermiller C (1985) Varieties of mere exposure: The effects of processing style and repetition on affective response. *Journal of Consumer Research* 17–30.

Papadopoulos G, Edwards PJ, Murray AF (2001) Confidence estimation methods for neural networks: A practical comparison. *IEEE transactions on neural networks* 12(6):1278–1287.

Paraschakis D, Nilsson BJ, Holländer J (2015) Comparative evaluation of top-n recommenders in e-commerce: An industrial perspective. *2015 IEEE 14th International Conference on Machine Learning and Applications*, 1024–1031.

Park YJ (2013) The adaptive clustering method for the long tail problem of recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 25(8):1904–1915.

Park YJ, Tuzhilin A (2008) The long tail of recommender systems and how to leverage it. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 11–18 (ACM).

Pathak B, Garfinkel R, Gopal RD, Venkatesan R, Yin F (2010) Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems* 27(2):159–188.

Pattaratanakun JA, Mak V (2015) Culture moderates biases in search decisions. *Psychological Science* 26(8):1229–1240.

Pauwels K, Neslin S (2015) Building with bricks and mortar: The revenue impact of opening physical stores in a multichannel environment. *Journal of Retailing* 91(2):182—-197.

Pessemier EA (1978) Stochastic properties of changing preferences. *The American Economic Review* 68(2):380–385.

Picard RR, Cook RD (1984) Cross-validation of regression models. *Journal of the American Statistical Association* 79(387):575–583.

Proedrou K, Nouretdinov I, Vovk V, Gammerman A (2002) Transductive confidence machines for pattern recognition. *European Conference on Machine Learning*, 381–390 (Springer).

Qian Y, Tan T, Hu H, Liu Q (2018) Noise robust speech recognition on aurora4 by humans and machines. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5604–5608 (IEEE).

Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. *Health information science and systems* 2(1):3.

Rasmussen CE (2003) Gaussian processes in machine learning. *Summer School on Machine Learning*, 63–71 (Springer).

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: Bayesian personalized ranking from implicit feedback. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 452–461 (AUAI Press).

Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175–186.

Ribeiro MT, Ziviani N, Moura ESD, Hata I, Lacerda A, Veloso A (2015) Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology* 5(4):53:1–53:20.

Ricci F, Rokach L, Shapira B, Kantor PB (2015) *Recommender Systems Handbook* (Springer).

Richins ML, Bloch PH (1986) After the new wears off: The temporal context of product involvement. *Journal of Consumer Research* 13(2):280–285.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285–295.

Sarwar B, Karypis G, Konstan J, Riedl J, et al. (2000) Analysis of recommendation algorithms for e-commerce. *EC*, 158–167.

Saunders C, Gammerman A, Vovk V (1999) Transduction with confidence and credibility .

Shao J (1996) Bootstrap model selection. *Journal of the American Statistical Association* 91(434):655–665.

Shepard W (2017) How china's shopping malls survive and thrive in the e-commerce age. *Forbes, October* .

Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in qsar. *Journal of chemical information and computer sciences* 44(6):1912–1928.

Shi L (2013) Trading-off among accuracy, similarity, diversity, and long-tail: a graph-based recommendation approach. *Proceedings of the 7th ACM Conference on Recommender Systems*, 57–64 (ACM).

Shrestha DL, Solomatine DP (2006) Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks* 19(2):225–235.

Shriver S, Bollinger B (2015) A structural model of channel choice with implications for retail entry. *Shriver: Columbia Business School, Colubia University* .

Simoiu C, Corbett-Davies S, Goel S (2016) Testing for racial discrimination in police searches of motor vehicles. *SSRN Journal* .

Smith MD, Brynjolfsson E (2001) Consumer decision-making at an internet shopbot: Brand still matters. *The Journal of Industrial Economics* 49(4):541–558.

Solares E, Coello CAC, Fernandez E, Navarro J (2019) Handling uncertainty through confidence intervals in portfolio optimization. *Swarm and evolutionary computation* 44:774–787.

Su R, Yin L, Chen K, Yu Y (2013) Set-oriented personalized ranking for diversified top-n recommendation. *Proceedings of the 7th ACM Conference on Recommender Systems*, 415–418 (ACM).

Swinney R (2011) Selling to strategic consumers when product value is uncertain: The value of matching supply and demand. *Management Science* 57(10):1737–1751.

Tan PN, Kumar V, Srivastava J (2002) Selecting the right interestingness measure for association patterns. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32–41 (ACM).

Tan PN, Steinbach M, Kumar V (2006) *Introduction to Data Mining*.

Tan TF, Netessine S, Hitt L (2017) Is tom cruise threatened? an empirical study of the impact of product variety on demand concentration. *Information Systems Research* 28(3):643–660.

Taramigkou M, Bothos E, Christidis K, Apostolou D, Mentzas G (2013) Escape the bubble: Guided exploration of music preferences for serendipity and novelty. *Proceedings of the 7th ACM Conference on Recommender Systems*, 335–338 (ACM).

TMO Group TG (2018) 2018 china ecommerce insights. *TMO Group, March* .

Tomassetti S, Wells AU, Costabel U, Cavazza A, Colby TV, Rossi G, Sverzellati N, Carloni A, Carretta E, Buccioli M, et al. (2016) Bronchoscopic lung cryobiopsy increases diagnostic confidence in the multidisciplinary diagnosis of idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 193(7):745–752.

Toplak M, Mocnik R, Polajnar M, Bosnic Z, Carlsson L, Hasselgren C, Demsar J, Boyer S, Zupan B, Stalring J (2014) Assessment of machine learning reliability methods for quantifying the applicability domain of qsar regression models. *Journal of chemical information and modeling* 54(2):431–441.

Traylor MB (1981) Product involvement and brand commitment. *Journal of Advertising Research* .

Tsanas A, Little MA, McSharry PE, Ramig LO (2009) Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering* 57(4):884–893.

Tse DK, Belk RW, Zhou N (1989) Becoming a consumer society: A longitudinal and cross-cultural content analysis of print ads from hong kong, the people's republic of china, and taiwan. *Journal of consumer research* 15(4):457–472.

Tzikas D, Kukar M, Likas A (2007) Transductive reliability estimation for kernel based classifiers. *International Symposium on Intelligent Data Analysis*, 37–47 (Springer).

Valladares EC (2017) Zara intensifies expansion with new flagship stores. *Fashion Network* .

Verhoef PC, Kannan PK, Inman JJ (2015) From multi-channel retailing to omni-channel retailing: introduction to the special issue on multi-channel retailing. *Journal of Retailing* 91(2):174–181.

Verhoef PC, Neslin SA, Vroomen B (2007) Multichannel customer management: Understanding the research-shopper phenomenon. *International Journal of Research in Marketing* 24(2):129–148.

Walker SH, Duncan DB (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54(1-2):167–179.

Wang CL, Bristol T, Mowen JC, Chakraborty G (2000) Alternative modes of self-construal: Dimensions of connectedness–separateness and advertising appeals to the cultural and gender-specific self. *Journal of Consumer Psychology* 9(2):107–115.

Wang CL, Lin X (2009) Migration of chinese consumption values: traditions, modernization, and cultural renaissance. *Journal of business ethics* 88(3):399–409.

Wang K, Goldfarb A (2017) Can offline stores drive online sales? *Journal of Marketing Research* 54(5):706–719.

West P (1989) Cross-cultural literacy and the pacific rim. *Business Horizons* 32(2):3–14.

Wickramaratna K, Kubat M, Premaratne K (2009) Predicting missing items in shopping carts. *IEEE Transactions on Knowledge and Data Engineering* 21(7):985–998.

Wonnacott T (1987) Confidence intervals or hypothesis tests? *Journal of Applied Statistics* 14(3):195–201.

Wonnacott TH, Wonnacott RJ (1990) *Introductory statistics*, volume 5 (Wiley New York).

Wu J, Zhu S, Liu H, Xia G (2012) Cosine interesting pattern discovery. *Information Sciences* 184(1):176–195.

Wu Z, Cao J, Wu J, Wang Y, Liu C (2014) Detecting genuine communities from large-scale social networks: a pattern-based method. *The Computer Journal* 57(9):1343–1357.

Wübben M, Wangenheim Fv (2008) Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing* 72(3):82–93.

Xiong H, Tan PN, Kumar V (2006) Hyperclique pattern discovery. *Data Mining and Knowledge Discovery* 13(2):219–242.

Xu B (1992) Reaching the chinese consume. *China Business Review, November-December* .

Yau OH (1988) Chinese cultural values: Their dimensions and marketing implications. *European Journal of marketing* 22(5):44–57.

Yin H, Cui B, Li J, Yao J, Chen C (2012) Challenging the long tail recommendation. *Proceedings of the VLDB Endowment* 5(9):896–907.

Zaíane OR (2002) Building a recommender agent for e-learning systems. *Proceedings of International Conference on Computers in Education*, 55–59.

Zaichkowsky JL (1985) Measuring the involvement construct. *Journal of Consumer Research* 12(3):341–352.

Zhang M (2009) Enhancing diversity in top-n recommendation. *Proceedings of the 3rd ACM Conference on Recommender Systems*, 397–400.

Zhang M, Hurley N (2008) Avoiding monotony: improving the diversity of recommendation lists. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 123–130.

Zhang M, Hurley N (2009) Novel item recommendation by user profile partitioning. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 508–515.

Zhang YC, Séaghdha DÓ, Quercia D, Jambor T (2012) Auralist: introducing serendipity into music recommendation. *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 13–22.

Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, Zhang YC (2010) Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107(10):4511–4515.

Ziegler CN, McNee SM, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification. *Proceedings of the 14th International Conference on World Wide Web*, 22–32.