Essays on Digital Transformation: Turning Data Into Assets

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sandeep Kumar Gangarapu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ravi Bapna, Edward McFowland III

Nov, 2022

Acknowledgements

I want to thank my advisor Dr. Ravi Bapna, who patiently guided me throughout the Ph.D. He was like a sherpa who helped me climb my own Mount Everest. He showed me the path whenever I was lost, encouraged me whenever I was tired, and stuck with me throughout the entire route, even when I made rookie mistakes. I counted my stars multiple times throughout the program for finding an advisor like him. I also want to thank Dr. Edward McFowland III for trusting me and taking me under his wing. I learned what it means to be a rigorous academic, being true to self and true to science from him. He was the perfect balance needed for mine and Ravi's academic personalities. Together, we formed a triangle where the sum of any two sides is greater than the third and the difference less than the third.

I also want to thank my committee members who provided feedback and help improve the thesis - Mochen Yang, Steven Wu, Colleen Flaherty Manchester, Jason Kerwin. I want to thank the incredible Ph.D. cohort of the Information and Decision Sciences department at Carlson School, specifically Scott Schanke, Yash Babar, and Jeff Clement. Scott and I felt comfort in each other's misery (sounds weird but helps) and celebrated the smallest of wins to the fullest extent. We promised each other in our second year that we would somehow see this through, and I am glad that day has come. Yash always spoke his mind and kept me close

to reality. Jeff passively encouraged me to work hard and produce quality research by just being Jeff.

I am also indebted to the entire faculty of Information and Decision Sciences department for accepting different flavors of research and always setting a high standard of research output and collaboration. They helped me understand the true meaning of scholarship, and I am proud to stand on their shoulders.

I want to thank the pillars like friends I leaned on for 5+ years - Sri Harsha, Anmol, Dheeraj, Aditi, Devleena, Kadidja, Ken, Celestine, Shashank, Rahul, Sandeep, Gautam, Sravan, Chandu, Shukla, Srikanth, Rashmitha, and Meghana.

Finally, I want to thank my mom for always pushing me to be better and being a constant source of inspiration, my dad who showered me with enough love even when I make it hard to do so, and my younger brother who I learn from every day and helps me hold the fort down.

See what I am up to these days on my website.

Dedication

To my mother, Padmavathi who showed that a strong woman can transform a generation and paved us a path.

Abstract

Digital transformation is defined as the use of digital technologies to transform every area of an existing business. A 2019 McKinsey report (Bughin et al. 2020) found that the top 10% of the digitized incumbents earned 80% of the revenue in their industries. One of the key strategic themes of digital transformation is to turn data the company collects into assets (Rogers 2016). In this thesis, we look at two different contexts, randomized control trials and referral marketing. In the first essay, we define a prescriptive analytics framework that addresses the needs of a budget constrained decision-maker facing, ex-ante, unknown costs and benefits of multiple policy levers to maximize overall utility from the randomized control trial data that the company already stores. we find a targeting strategy that produces an order of magnitude improvement in expected total utility compared to existing methods. In the second essay, we solve some of the challenges associated with referral marketing by turning the referral data that the company has into an asset. We investigate how referral targeting compares to algorithmic targeting in its effectiveness. We understand the mechanisms behind a referral and why they are valuable. We also unpack the effects of 'information' and 'influence' that play a role in the purchase decision of a referred customer.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Digital transformation is defined as the use of digital technologies to transform every area of an existing business. A 2019 McKinsey quarterly report (Bughin et al. 2020) found that the top 10% of the digitized incumbents earned as much as 80% of the revenue in their industries. One of the key strategic themes of digital transformation is to turn data the company collects into assets that help the company grow (Rogers 2016). For example, Netflix transformed from an online DVD rental company to the streaming giant it is today by collecting granular data on user watch behavior and preferences, then using that data to make decisions on what TV shows to produce and which celebrities to cast. In 2017, Netflix renewed 93% of original TV shows after the first season, compared to 35% for cable television. In another example, UPS started digitizing it deliveries by giving every driver a device to log deliveries and track GPS routes. Subsequently, UPS made route optimizations with all the GPS data, which alone saves them millions of dollars every year. An added benefit is that live GPS location information also provides real-time tracking for packages, which creates a better customer experience. This eventually led to UPS creating a new service—a "my

choice premium subscription"—that allows customers to change delivery dates, drop off packages at different locations, and even earns UPS additional revenue. This is a great example of how data can be turned into an asset. There is no question that companies can use raw data to optimize processes, drive innovations, and create new products; however, there is a dissonance between the pace at which the data are created and the pace of development of methods used to make sense of the data and turn them into actionable insights (Economist 2010). Historically, algorithmic development and managerial decision-making research have been disconnected. Addressing this gap and using algorithmic development to solve critical managerial challenges will help turn data into assets. In this thesis, I investigate this problem in the domains of randomized control trials (A/B testing) and social referral marketing.

Companies use RCTs to make data-driven decisions and continuously iterate to improve the product. There is clear evidence that start-ups that adopt A/B Testing to deploy features have increased performance on several critical dimensions, including page views and new product features (Koning et al. 2022). In chapter 2, we look at how existing RCT data can further be used in downstream processes and turned into assets. We define a prescriptive analytics framework that addresses the needs of a constrained decision-maker facing ex-ante, unknown costs and benefits of multiple policy levers. The framework is general and can be deployed in any utility-maximizing context. It relies on RCTs for causal inference, machine learning for estimating heterogeneous treatment effects, and the optimization of an integer linear program for converting predictions into decisions. The net result is the discovery of individual-level targeting of policy interventions to maximize overall utility under a budget constraint. This framework can be used by companies using existing RCT data and turned into an asset by finding a

targeting strategy that produces an order of magnitude improvement in expected total utility compared to existing targeting approaches such as average treatment effect and uplift modeling. We also show how policies like ours can help address the problem of the adoption of experimentation in companies. Though there is evidence suggesting the benefits of experimentation in a company's revenue, growth, and user retention, we do not see its widespread use in the industry. Industry experts attribute this problem to experimentation culture, infrastructure costs, and technical ability. This chapter provides methodological ways to increase utility from experimentation, thereby encouraging adoption.

We investigate another context of digital transformation: the area of social referral marketing, especially how companies can use existing data on social referral mechanisms as valuable assets. Referral-based programs that leverage existing customers' social networks to reach potential new customers are effective, as customers acquired through referrals are more valuable than other channels (Schmitt et al. 2011), although there are some problems with referral marketing. Referral campaigns soliciting information from existing customers place the burden on those customers to take action, which increases email overload and annoyance (Grevet et al. 2014) and leads to decreased loyalty (e.g., unsubscribes). The incentive mechanism for new customers can prove costly for companies if those customers will not return without the incentive. This is a problem, especially for those companies with low-profit margins and those that sell high-ticket items. The conversion rates in referral campaigns are usually low, making them cost-ineffective and non-scalable. In chapter 3, we investigate how referral targeting compares to algorithmic (machine-learning or matched) targeting in its effectiveness. We see if we can use existing machine-learning or econometric methods to turn data stored on previous referral marketing campaigns into assets we call

matched targeting to compare the assets' effectiveness to referral targeting. We understand the mechanisms behind a referral and why they are valuable. We also unpack the effects of 'information' and 'influence' that play a role in the purchase decision of a referred customer.

Randomized control trials are the gold standard for causal inference. Both essays use RCT's to understand the causal impact of a policy at hand on the outcome of interest. But, in both essays, we demonstrate the value of observational data that the company store increasing the utility of those experiments by feeding that data into machine learning algorithms that are used either in the downstream process (creating a heterogeneous treatment effect model) or in the upstream process (creating a machine learning model as a targeting policy).

In the following paragraphs, we give brief summaries of each chapter along with findings.

1.1 A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects

In this chapter, we define a prescriptive analytics framework that addresses the needs of a constrained decision-maker facing ex-ante, unknown costs and benefits of multiple policy levers. The framework is general in nature and can be deployed in any utility maximizing context, whether public or private. It relies on randomized field experiments for causal inference, machine learning for estimating heterogeneous treatment effects, and the optimization of an integer linear program for converting predictions into decisions. The net result is the discovery of

individual-level targeting of policy interventions to maximize overall utility under a budget constraint. The framework is set in the context of the four pillars of analytics and is especially valuable for companies that already have an existing practice of running A/B tests. The key contribution of this work is to develop and operationalize a framework to exploit both within and between-treatment arm heterogeneity in the utility response function in order to derive benefits from future (optimized) prescriptions. We demonstrate the value of this framework as compared to benchmark practices—i.e., the use of the average treatment effect as well as uplift modeling—in two different settings. Unlike these standard approaches, our framework can recognize, adapt to, and exploit the (potential) presence of different subpopulations that experience varying costs and benefits within a treatment arm while also exhibiting differential costs and benefits across treatment arms. As a result, we find a targeting strategy that produces an order of magnitude improvement in expected total utility, in cases where significant within- and between-treatment arm heterogeneity exists.

1.2 Turning referral data into assets: Evidence from an RCT on algorithmic vs. social targeting

In this chapter, we solve some of the challenges associated with referral marketing, such as email overload and annoyance, costly incentives, and low conversion rates by turning a company's referral data into assets. We investigate how referral targeting compares to algorithmic (machine-learning or matched) targeting in its

effectiveness. We want to see if we can use existing machine-learning or econometric methods to turn data stored on previous referral marketing campaigns into assets we call matched targeting, and then compare the assets' effectiveness to referral targeting. We want to understand the mechanisms behind a referral and why they are valuable. We also unpack the effects of 'information' and 'influence' that play a role in the purchase decision of a referred customer. We conduct two field experiments with the University of Minnesota Alumni Association by designing a referral campaign to ask its existing members to refer their friends to subscribe to the UMAA 's annual membership program. We find that referral targeting has a better uptake rate than machine-learning targeting for the same offer. We also designed a matched targeting method using coarsened exact matching (CEM) on existing referral data and targeted people with similar characteristics (information motivation) to that of the referred members. We find that both referral and machine-learning targeting have better uptake rates than matched targeting. This also shows that social influence plays a significant role in a referral mechanism compared to just straight information, though we do not claim any causality.

Chapter 2

A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects

2.1 Introduction

We address the general problem of a budget-constrained decision maker facing, ex-ante, unknown costs and benefits from multiple policy levers that she can potentially deploy to optimize an organizational goal. We define and deploy a prescriptive analytics framework as one that folds together the use of a) randomized field experiments for causal inference around the estimation of ex-ante unknown costs and benefits; b) machine learning to identify heterogeneity in treatment effects, and thereby, go beyond inference around average treatment effects; and c) constrained optimization to optimally decide which subpopulation of individuals

to treat with which policy levers, maximizing profit in the presence of organizational and individual level constraints. While each of the above three folds–causal inference, supervised machine learning and constrained optimization–are academic disciplines on their own, the distinct contribution of this chapter is to stitch specific aspects of them together into a decision-making framework. In doing so we advance the thinking around each of the individual disciplines as well. By using supervised machine learning on data generated by experiments, originally deployed to derive causal inference, we are able to discover patterns of within- and between-treatment arm heterogeneity in the individual and group response functions. This allows our work to go beyond using the average treatment effect as the decision lever from randomized experiments. In estimating unbiased individual level costs and benefits of multiple levers we parameterize the inputs, or 'the data', that feeds into the optimization problem. This is generally not the purview of the optimization literature, which assumes that the 'data' exists, and focuses on deriving, for example, efficient algorithms to solve computationally complex (often NP-hard) problems.

Consider, as a motivating example, a marketing manager deciding between three different types of call-to-actions (CTA) to initiators of a referral campaign. Jung et al. (2019) experiments with an altruistic, equitable and egoistic framing of the call-to-action, to activate potential senders of a referral. The target population in this situation is the set of past customers, for whom firms typically collect a vast amount of demographic and behavioral data. Further, promotions such as referral programs are incentive laden word-of-mouth mechanisms, and like all promotions, have underlying costs and benefits associated with them. In the context investigated by Jung et al. (2019), each sender and recipient of a referral is eligible to receive a free product (valued at $25) with free shipping included. Thus

depending on how many referrals a person sends out, how many of these referrals lead to recipient purchases, and the varying dollar amounts of these purchases, the firm can realize many different values of cost and benefit. Furthermore, different value patterns can emerge from different subpopulations of individuals, as a consequence of different treatment allocations. Consider the following scenario:

1. User A (targeted using an equitable CTA) invites a friend who buys a $200 item and receives a $25 gift. User A also receives a $25 gift for initiating the referral. The resulting utility to the firm is $150

2. User B (targeted using a selfish CTA) invites a friend who buys a $20 item and receives a $25 gift. User B also buys a $20 item and receives a $25 gift. The resulting utility to the firm is -$10

3. User C (targeted using an altruistic CTA) invites 8 friends, each of who buy a $35 item and each receives a $25 gift. User C also buys a $35 item and receives a $25 gift. The resulting utility to the firm is $90

The challenge is that *ex-ante* the marketer does not know whether any of these call-to-actions have a causal impact on profitability, let alone know whether the impact of the three different arms may be heterogeneous, as a function of existing customer "data" – e.g., characteristics and behaviors. Further, given organizational constraints (e.g., budget) or those at the individual level, the problem of allocating individuals to treatments in order to maximize utility is an NP-hard problem (Martelo & Toth 1990). As a motivating example of the generality of our framework, consider an individual-level constraint that restricts the availability of certain treatments to certain individuals based on past interactions with the company. For instance, if there are existing users who have already been targeted

with a particular call-to-action multiple times, the company may not want to overwhelm them with the same call-to-action yet again.

We demonstrate that challenges of this type can be addressed using a prescriptive analytics approach built atop the key pillars of analytics (Figure 2.1): causal inference, machine learning and optimization, to enable a budget constrained decision maker to optimally target interventions with *ex-ante* unknown costs and benefits. A variety of organizations already use some variant of the causal inference paradigm (called 'A/B testing' in the industry) to determine which ads to show, what features to deploy, or what type of incentives to provide in order to motivate users to perform an action (Kohavi & Thomke 2017). In the first stage of our approach, the intent is to layer on to this existing practice an ability to use machine learning methods to make inference beyond the average treatment effects. More specifically, we showcase how to use the data from A/B testing to develop and validate a robust heterogeneous treatment effect (HTE) procedure that can provide estimates of cost and benefit for each individual, in each treatment condition. The next stage of our approach utilizes the 'randomly optimal' targeting that occurs by chance from random assignment, as a means for validating the existence of sufficiently exploitable between-condition heterogeneity. As a result of random assignment, our HTE step gives an unbiased (selection free) ordering of treatments, for each individual. Therefore, we can rely on random assignment to (by chance) produce a targeting of individuals at varying degrees of optimality–e.g., some individuals will experience their optimal assignment. This variation in treatment optimality allows us to examine whether subjects who were randomly placed in their (predicted) optimal condition exhibited higher utility than those that were not. We call this the observed utility rank condition (OUR), a metric that serves as a necessary condition for progressing to

the final (optimization) stage of our approach. In the final stage we exploit the within- and between-treatment heterogeneity in cost and benefit for the development of optimal prescriptions, i.e., targeting treatments to individuals, while respecting organizational constraints. While formal definitions are laid out in the sections that follow, in essence, our process adapts to the existence of heterogeneity across individuals (or subpopulations) for a given treatment, as well as requires the existence of heterogeneity between treatment conditions within individuals (or subpopulations). This makes it worthwhile for the optimization to exploit such heterogeneity and match treatments to the most suitable sets of individuals. If this heterogeneity between treatment conditions does not exist, there is no value in progressing to the prescription phase. When this heterogeneity does exist, we validate this final stage by comparing the utility generated by our prescriptions (allocations) against the current practices of assigning all individuals to the condition that has the highest average treatment effect, as well as uplift modeling (Rzepakowski & Jaroszewicz 2010).

We demonstrate the value of our proposed three-stage process using two real-world settings. Firstly, we consider a public policy setting used to motivate blood donations, and find progressive value in our three-stage framework. In particular, we find that 1) our HTE model has low out-of-sample mean absolute error, indicating that it is able to capture the data-generating process of the underlying treatment arms; 2) there exists clear observed ordering of the OUR metric, indicating enough heterogeneity across treatment conditions to be exploited; and 3) there exists significant gains in utility from the optimal allocation as compared to ATE and uplift modeling. In contrast, in the context of a call-to-action experiment for referral marketing, while the analysis passes stage 1 (low MAE), it fails

stage 2. This implies that even though the model captures the underlying data-generation process of each of the treatment arms, there is not sufficient evidence for the existence of heterogeneity between conditions. Furthermore, the absence of between-condition heterogeneity indicates a lack of value to be gained in stage 3. Thus, our practical contribution stems from providing an overall process that can be applied to any experimental setting of multiple treatment conditions, with ex-ante unknown costs and benefits. Step 2, in particular, provides a pragmatic tollgate, that prevents organizations from undertaking a targeting strategy that lacks sufficient evidence for generating increased utility. Such a strategy may lead to prescriptions that are laden with costs (e.g., the cost of chosen sub-optimal treatment, opportunity costs, etc.) without sufficiently counterbalancing benefit.

Overall, we contribute by developing and operationalizing a prescriptive analytics framework that combines randomized field experiments for causal inference; machine learning to exploit heterogeneity and advance this inference beyond the average treatment effects; and constrained optimization to optimally decide which subjects to treat with which policy levers, to maximize profit in the presence of organizational and individual level constraints. We believe that this prescriptive analytics framework encourages companies to take an integrated view of the four pillars of analytics (Figure 2.1). In particular, for sake of completeness, our framework also depends on using managerial judgment and exploratory analytics to generate treatment conditions, where there is the potential for an effect. Our study is also the first to highlight, and develop a metric (i.e., the OUR metric) around the importance of between-treatment arm heterogeneity of cost and benefit treatment effects.

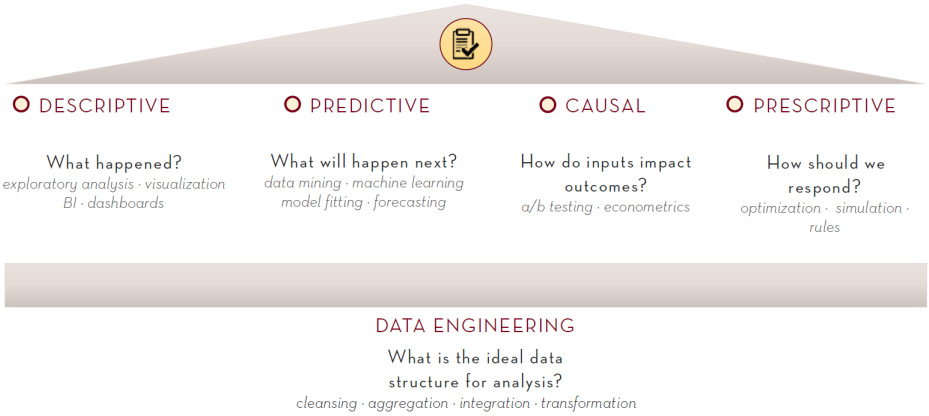

Figure 2.1: Four Pillars of Analytics

2.2 Background and related literature

Our work is adjacent to the emerging literature around the use of machine learning for heterogeneous treatment effects (Wager & Athey 2018, Athey & Imbens 2016, McFowland III et al. 2018). However, as in Imai & Strauss (2011), we focus on extending these ideas to the design of optimal policy decisions. Moreover, Imai & Strauss (2011) (which proposes a targeting approach in the context of "get-out-the-vote" campaigns) and our work argue that if policy makers solely rely on average treatment effects, they fail to exploit potentially valuable sources of treatment heterogeneity. However, while the overall objective in Imai & Strauss (2011) is similar to ours, we differ in the generality of the optimization approach to determine treatments assignment. In particular, Imai & Strauss (2011) adopts a Bayesian optimization approach, requiring knowledge of a suitable prior that is then subsequently revised with experimental data to determine the posterior distributions of the various treatments' effects on the probability of voting. Furthermore, the focus is maximizing the treatment effect on benefit (measured as a probability) subject to a cost constraint. It is not immediately obvious how to

optimize general utility (the difference in benefit and cost) where cost is not measured as probabilities (but instead as currency) and costs are unknown ex-ante; both of which occur in our setting.

Additionally our work relates to, but is significantly different from, the uplift modeling literature (Rzepakowski & Jaroszewicz 2010) and the contextual bandits literature (Auer 2002, Langford & Zhang 2008, Li et al. 2010) in machine learning and design of experiments. The main commonality lies in the end objective of personalizing decisions, such as marketing interventions, based on consumer characteristics and past behaviors. The key idea in uplift modeling is to go beyond building a model to simply predict who is likely to respond to an intervention, but to target those that will respond *because* they received the intervention. Intuitively, the goal is to develop a model whose estimation of treatment effects, and subsequent predictions of who should be targeted, will result in better responses relative to control group subjects that are randomly targeted. The practice of targeting customers who will respond even in the absence of the intervention, actually results in unnecessary costs. If we extend this setting to multiple treatments, with sequential allocation of subjects to treatments considering their demographic and behavioral data, we enter the setting of the contextual bandit problem. Therein, the key issue is dealing with the exploration-exploitation balance: trading off the focus on increasing learning with the focus on earning. Exploration is used to estimate individual rewards based on known contextual data, while exploitation simultaneously attempts to maximize cumulative returns in a sequential decision making framework. Multi-armed bandit approaches are gradually being adopted in more conventional management decision contexts, such as the pricing of multiple products (Misra et al. 2018) and the acquisition of customers using display advertising (Schwartz et al. 2017). Our framework borrows

ideas from both these streams of literature, and is consistent with what Bertsimas & Kallus (2014) describes as moving from predictions to decisions. A key aspect of this move is the incorporation of organizational and individual level constraints; for example, budgets and individual level ineligibilities, respectively. Further, to demonstrate the value of our proposed approach, we do not require the additional complexity of a sequential decision-making policy optimization formulation, that is characteristic of the known algorithms for the contextual bandit problem (Auer 2002, Li et al. 2010). In our work, we decouple the joint optimization of the exploration-exploitation by treating the learning phase–which result from utilizing heterogeneous treatment effect procedures to process and extract information from randomized experiments–as an input into the optimal targeting phase. From a practical point of view, our approach requires less supporting infrastructure and is much easier to deploy than bandit approaches, improving its accessibility to less technically sophisticated organizations. Moreover, our approach can provide immediate value for organizations that have already developed some causal analytics capabilities, and potentially have data from past experimentation contained in their administrative records.

It is important to note that the popularity of multi-armed bandit based approaches (Chu et al. 2011, Ding et al. 2013, Joulani et al. 2013) has risen with the popularity of A/B testing on digital platforms in what some call an online experimentation model (think content testing on a web page, or showing an ad on the Facebook wall, and observing subsequent engagement). We highlight this, because the use of multi-armed bandits is (essentially) predicated on the ability to update the policy of treatment delivery throughout the course of the experimentation process. While such conditions may be satisfied in certain contexts

where the customer may immediately respond to the treatment (e.g. click outcome in sponsored ads or website design application) or the response is observed with a delay (e.g. the decision to purchase a product might be observed hours or days after the assignment of treatment), there are many contexts when outcomes are not able to be observed instantaneously and the treatment effects cannot be measured prior to the treatment of subsequent subjects: often, outcomes for all subjects are observed long after the treatment. Take, for example, a blood donation experiment that serves as one of our case studies later in this chapter. Here, treatments can be given weeks (or even earlier) before outcomes are observed. Moreover, outcomes are not observed in sequences, rather they are collected at once on the day of the blood-donation. Therefore, all treatments are given before any outcomes can be observed, defeating the purpose and benefit of multi-armed bandit approaches. These are similar scenarios faced in a collage.com experiment (the second case study we discuss in detail later in this chapter) were purchases are made over a one-month period. Moreover, almost all interventions designed to move the long-term outcome (e.g. projected customer lifetime value) face certain constraints when using multi-armed bandit methods. While we understand that there are advances made in Multi-armed Bandit research to incorporate delayed feedback (Vernade et al. 2018), we believe that our approach is still useful for those applications and could complement the multi-armed bandit approach to improve firms' practice of randomization experiments. In general, we envision firms developing a cyclical organizational discipline of hypotheses generation using exploratory analytics, judgment, and intuition; the adoption of the scientific method, to validate and test these hypotheses using causal analytics; the use of predictive modeling, to go beyond ATE exploiting heterogeneity; and the use of constrained optimization, for eventual decision making.

2.3 Prescriptive analytics framework

We begin by providing notation for establishing the setting of our prescriptive analytics framework. Let $\mathcal{N}$ be a sample of n independent and identically distributed units from the population of interest $\mathcal{P}$, such that sample units are indexed by $i \in \{1, \ldots, n\}$, and for each unit we observe characteristics $X_i \in \mathcal{X}$. Let the decision maker have a collection of J treatments $\mathcal{T} = \{T_0, \ldots, T_J\}$. Following the potential outcomes framework, we posit for each unit i the existence of potential *ex-ante* unknown outcomes $(B_i(j), C_i(j), U_i(j) \in \mathbb{R})$ which are the respective benefits, costs, and utility $(U_i(j) = B_i(j) - C_i(j))$ that would be realized by assigning unit i to treatment j. Let the decision maker's budget be M. Let $A_{ij} \in \{0, 1\}$ be the decision variable, such that $A_{ij} = 1$ assigns unit i to receive treatment T_j (note that A_{ij} is turned 'on' for only one j for each i). We elaborate on each of the key aspects of our framework in the following sections. We start with highlighting the growing value of causal analytics in organizations.

2.3.1 In-vivo experimentation

Two assumptions (preconditions) that are implicit to our framework are that firms can use exploratory analytics to discover new policy levers, and that firms can adopt the randomized control trial to causally estimate the expected effects of different policy levers on benefits and costs. The first assumption, by its exploratory nature, is difficult to explicate precisely or conduct mechanically. Furthermore, in this work, we utilize pre-existing large-scale randomized experiments for conducting the empirical analysis. Therefore, given the ex-post focus of our work, we do not cover how to develop an initial set of treatment conditions, conduct a randomized experiment with these conditions, explore the experimental results to

discover new conditions with potentially larger effects, or repeat this process as would an organization. However, we do highlight that recent advances, including subset scanning based approaches for subpopulation discovery (McFowland III et al. 2018, Somanchi et al. 2018), provide promising methods for organizations to engage in such exploratory hypothesis generation. Subset Scanning is a concept that originates in anomalous pattern detection (Neill 2012, McFowland III et al. 2013, Neill et al. 2013, Speakman et al. 2015), which has been adapted for heterogeneous treatment effects to identify the subpopulations with the most statistically significant treatment effects in randomized experiments (McFowland III et al. 2018) and field studies with multiple treatments (Somanchi et al. 2018). These method focus solely on identifying subpopulations with sufficient evidence of a treatment effect, which can be characterized as generating treatment hypothesis that are supported by the data. While, the idea of data-generating hypothesis is already prevalent in fields of (bio)informatics (Dopazo & Aloy 2006, Biesecker 2013); there has been a recent call for research in the IS community (Agarwal & Dhar 2014) to take advantage of the power of big data and machine learning in order to identify phenomena of interest, and use the results to investigate causal relationships by exploiting econometric techniques.

The second assumption is consistent with the emerging stream of IS literature that pinpoints the benefits of in-vivo large scale randomized field experiments, with respect to identifying nuanced mechanisms of interest in complex online systems, such as social networks and online dating markets (Aral & Walker 2011, Bapna & Umyarov 2015). It is also consistent with the growing recognition of 'A/B testing' as a valuable organizational capability to develop. Kohavi & Thomke (2017) describes not only how digital native organizations–e.g., Amazon,

Booking.com, Facebook, and Google–each conduct tens of thousands of experiments annually, but also how traditional companies–e.g., Walmart, Hertz, and Singapore Airlines–are beginning to rely on experimentation, albeit at a smaller scale. We treat this practice as a baseline capability for organizations and policy makers, and develop a framework for increasing the value obtained from such experimentation, by using machine learning and optimization to advance current practice.

2.3.2 Heterogeneous treatment effects

With the results of a randomized experiment as input, our framework begins by relying on machine learning to discern heterogeneous treatment effects (HTE). Most approaches for HTE in the literature are built atop the potential outcomes framework, with (as good as) random treatment assignment, enabling causal inference. More precisely, they begin with observing $\mathcal{N}_1$, a sample of n independent and identically distributed units from a population of interest $\mathcal{P}$. The units are indexed by $i \in \{1, \ldots, n\}$, and for each unit there is a binary assignment indicator $W_i \in \{0, 1\}$, where $W_i = 0$ indicates assignment to the control group, while $W_i = 1$ indicates assignment to the treatment group. Therefore, there exist two potential outcomes for each unit $(Y(0), Y(1) \in \mathbb{R})$, although in practice only one is realized. Additionally, each unit is described by $X_i \in \mathcal{X}$, a d-dimensional vector of covariates, whose support is the set $\mathcal{X}$.

In particular, there is interest in a causal population estimand τ that is a function of the potential outcome distributions and covariates, and measures the

treatment effect

$$\begin{aligned}\tau &= \tau(F_{Y(1)}, F_{Y(0)}, X) \\ &= Div(F_{Y(1)|X}, F_{Y(0)|X})\end{aligned}$$

where $Div\colon (F, F') \mapsto \mathbb{R}$ is a general measure of divergence between cumulative distribution functions (CDF) F and F'. The most common estimand of interest is the average treatment effect (ATE),

$$\begin{aligned}\tau_{\text{ATE}} &= \int y \; dF_{Y(1)}(y) - \int y \; dF_{Y(0)}(y) \\ &= \mathbb{E}[Y(1) - Y(0)],\end{aligned}$$

which computes the expected difference between the potential outcomes across the population. With the rising interest in HTE, the conditional average treatment effect (CATE)

$$\begin{aligned}\tau_{\text{CATE}}(x) &= \int y \; dF_{Y(1)|X}(y|x) - \int y \; dF_{Y(0)|X}(y|x) \\ &= \mathbb{E}[Y(1) - Y(0)|X = x],\end{aligned} \tag{2.1}$$

has also garnered much attention, as it considers how the potential outcome distributions vary for each covariate profile. However, the literature does offer alternative estimands (Grimmer et al. 2017), including a more general measures of divergence between distribution functions (McFowland III et al. 2018, Somanchi et al. 2018).

The CATE is the typical estimand of interest in the pursuit of HTE using

Machine Learning algorithms. More specifically, algorithms attempt to estimate

$$\tilde{\tau}_{\text{CATE}}(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]. \tag{2.2}$$

However, a challenge arises because $\tilde{\tau}_{\text{CATE}}$ is implicitly a function of both the observed and unobserved potential outcome values. At most one of the potential outcomes is observed for each unit in the sample:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1; \end{cases}$$

therefore machine learning algorithms cannot be trained directly to estimate $\tilde{\tau}_{\text{CATE}}$. Therefore, the literature commonly makes the additional assumption of unconfoundedness (Rosenbaum & Rubin 1983)

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \mid X_i \ \forall i, \tag{2.3}$$

because if we let $\gamma(x) = \mathbb{E}\left[W_i|X_i = x\right]$–the propensity of receiving treatment for a subject with covariate profile x–(2.3) implies

$$\mathbb{E}\left[Y_i^{\text{obs}}\left(\frac{W_i}{\gamma(x)} - \frac{1 - W_i}{1 - \gamma(x)}\right)|X_i = x\right] = \tilde{\tau}_{\text{CATE}}(x). \tag{2.4}$$

However, in our particular framework, we assume that the data is drawn from a randomized experiment, and therefore can conclude

$$Y_i(0), Y_i(1), X_i \perp\!\!\!\perp W_i \ \forall i. \tag{2.5}$$

It is clear that (2.5) is a stronger assumption which implies (2.3) and therefore

(2.4); more specifically, (2.4) becomes

$$\mathbb{E}\left[Y_i|X_i = x, W_i = 1\right] - \mathbb{E}\left[Y_i|X_i = x, W_i = 0\right] = \tilde{\tau}_{\text{CATE}}(x). \tag{2.6}$$

Now that $\tilde{\tau}_{\text{CATE}}$ can be expressed as a function of observed quantities, it can now be estimated empirically from data by

$$\hat{\tau}_{\text{CATE}}(x) = \frac{1}{|r(x,1)|} \sum_{i \in r(x,1)} Y_i - \frac{1}{|r(x,0)|} \sum_{i \in r(x,0)} Y_i \tag{2.7}$$

where $r(x, w) = \{i : X_i = x, W_i = w\}$. We know that theoretically $\hat{\tau}_{\text{CATE}}(x)$ possess good properties in terms of estimating $\tilde{\tau}_{\text{CATE}}(x)$ and subsequently $\tau_{\text{CATE}}(x)$, e.g., $\hat{\tau}_{\text{CATE}}(x)$ is an unbiased and consistent estimator. In practice, this empirical estimator is often substituted for an algorithm from the machine learning literature which has also been demonstrated to exhibit desirable properties theoretically (usually given regularity conditions) and empirically; for example, in this work we use the Random Forest estimation procedure. Finally, we note that although the HTE literature commonly considers one treatment group, and therefore defines the estimand and estimators of CATE as parameters by covariate profile x, this definition can be extended to account for multiple treatment groups. Essentially, in our previous definitions W_i is a binary indicator, which can be extended to $W_i \in \{0, 1, \ldots, J\}$ with the (2.1),(2.2),(2.7) being respectively redefined as

$$\begin{aligned}
\tau_{\text{CATE}}(x, j) &= \mathbb{E}[Y(j) - Y(0)|X = x], \\
\tilde{\tau}_{\text{CATE}}(x, j) &= \mathbb{E}[Y_i(j) - Y_i(0)|X_i = x], \\
\hat{\tau}_{\text{CATE}}(x, j) &= \frac{1}{|r(x,j)|} \sum_{i \in r(x,j)} Y_i - \frac{1}{|r(x,0)|} \sum_{i \in r(x,0)} Y_i.
\end{aligned} \tag{2.8}$$

We describe how organizations can implement and operationalize HTE in Section 2.4.

2.3.3 Optimal prescriptions with heterogeneous treatment effects

The final component of our framework utilizes the information gathered from estimating heterogeneous treatment effects in order to inform prescriptions for future subjects. As in the case of estimating HTE, we begin with observing $\mathcal{N}_2$, a sample of n independent and identically distributed units from the population of interest $\mathcal{P}$. Again, the sample units are indexed by $i \in \{1, \ldots, n\}$, and for each unit we observe $X_i \in \mathcal{X}$. Note that the sample observed here is different than the sample observed and utilized for estimating HTE ($\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$), although $\mathcal{P}$ and $\mathcal{X}$ must be the same. Furthermore, the decision maker has a collection of J treatments $\mathcal{T} = \{T_0, \ldots, T_J\}$, and must decide the value of each binary indicator $A_{ij} \in \{0, 1\}$, where $A_{ij} = 1$ assigns unit i to receive treatment T_j. Again, following the potential outcomes framework, we posit for each unit i the existence of potential outcomes $(B_i(j), C_i(j), U_i(j) \in \mathbb{R})$ which are the respective benefits, costs, and utility $(U_i(j) = B_i(j) - C_i(j))$ that would be realized by assigning unit i to treatment T_j. Therefore if the decision maker has a budget M, her objective

is to

$$\begin{aligned} \text{maximize:} \quad & \sum_{i=1}^{n}\sum_{j=0}^{J} U_i(j)A_{ij} \\ \text{subject to:} \quad & \sum_{i=1}^{n}\sum_{j=0}^{J} C_i(j)A_{ij} \leq M \\ & \sum_{j=1}^{J} A_{ij} \leq 1, \qquad \forall i \\ & A_{ij} \in \{0,1\}, \qquad \forall i,j. \end{aligned} \tag{2.9}$$

We label this optimal model in (2.9) as HTE-OPT, note that its solution is the result of a integer linear program (ILP), and recognize that it is purely theoretical as it is based on unknown (potential outcome) functions. Furthermore, these functions are not directly estimable from data as they rely on all of the potential outcomes, of which at most one can be observed. However, recall from §2.3.2 that we can use machine learning methods to obtain unbiased estimates of potential outcomes if we have data that follows a randomized experiment, as this implies (2.5). Therefore, we propose substituting the unknown potential outcomes in (2.9) for the following estimators

$$\begin{aligned} \hat{B}_i(j) &= \hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i \\ \hat{C}_i(j) &= \hat{\tau}^C_{\text{CATE}}(X_i, j) + \hat{C}_i(0)|X_i, \end{aligned} \tag{2.10}$$

which can be computed as in (2.8). We propose computing the estimators by explicitly modeling the heterogeneity, as the treatment effect is the facet that creates the variability of most interest. Moreover, it has been demonstrated that

the variables responsible for the treatment effect heterogeneity are not only qualitatively different from those that capture the response surface for the baseline condition, but they often have relatively weak predictive power (Imai & Ratkovic 2013, Imai & Strauss 2011). Therefore, attempting to estimate the entire function jointly will likely obscure components of the heterogeneity, especially when the true τ_{CATE} diverges from the assumptions of the chosen machine learning approach. Even beyond the particular learning approach, there are various forms of modeling for the $\hat{\tau}_{\text{CATE}}$ estimators (e.g., the meta-learners in Künzel et al. (2019)) which exhibit different finite sample properties depending on the complexity of the underlying treatment effect, further demonstrating the importance of properly capturing the uniqueness of this heterogeneity (variation). Therefore, our final proposed ILP optimization, which serves as an estimate of HTE-OPT, is labeled HTE-EST and is defined as follows:

$$\begin{aligned}
\text{maximize:} \quad & \sum_{i=1}^{n}\sum_{j=0}^{J} \hat{U}_i(j) A_{ij} \\
\text{subject to:} \quad & \sum_{i=1}^{n}\sum_{j=0}^{J} \hat{C}_i(j) A_{ij} \leq M \\
& \sum_{j=1}^{J} A_{ij} \leq 1, \qquad \forall i \\
& A_{ij} \in \{0,1\}, \qquad \forall i, j.
\end{aligned} \tag{2.11}$$

In order to solve this ILP, we use IBM CPLEX optimizer[1], an industrial level high-performance mathematical programming solver with state-of-the-art runtimes[2].

[1] `https://www.ibm.com/analytics/cplex-optimizer`

[2] Empirically, we find that this solver can optimize for over a million users in under five minutes and for ten million users in eight hours.

Having provided the theoretical and mathematical foundations of our framework, we now provide a practical operationalization for the determination of the heterogeneous treatment effects. This is an important aspect of our contribution where we leverage and develop key metrics (serving as tollgates) that determine whether the data generating process within a treatment arm can be properly modeled, and if there exists sufficiently exploitable heterogeneity across treatments. We need both conditions to be satisfied to provide value beyond average treatment effects or uplift modeling.

2.4 Operationalization of heterogeneous treatment effects

Above, we outlined how the four pillars of analytics can be unified to produce a general three-stage prescriptive analytics framework. In this section, we dive deeper into the facet of the framework that is built atop heterogeneous treatment effects. Specifically we outline how one may exploit HTE given a randomized control trial with a control group and multiple treatment arms.

The building and evaluation of the HTE comes from the ability to split the data into training and test datasets. The split percentage can be set by the decision maker. The training set is used to demonstrate the process an organization would take in their attempt to learn the costs and benefits of each intervention. The test set is used to demonstrate how the organization would implement our validation measures–i.e., computing out-of-sample model accuracy metrics and measuring the existence of exploitable heterogeneity across treatment conditions. Furthermore, the test set can also be used later to evaluate the expected utility the organization could obtain on future subjects (from the same population)

by following prescriptions derived from the learning. We describe our key steps formally using pseudocode where necessary.

2.4.1 Model fitting

The first objective of our process for operationalizing HTE is to fit a statistical model to the training data set, allowing us to develop an understanding of the treatment effects for each intervention. For each treatment intervention we compare the subjects who experience this treatment to those who experience control, in order to formulate an estimate of the treatment effect. For the ATE approach this reduces to simply computing the average difference in the outcomes of interest (e.g., firm benefit and firm cost) between the treatment and control groups. However, the prescriptive component of our framework, as well as uplift modeling, is built atop individual level treatment effects. Therefore, to support its prescriptions, we must build models that will provide individual-level treatment effect estimates. For each outcome of interest, using 5-fold cross validation, we build a model to capture its relationship with provided covariates, under each condition. The collection of these models form an estimate $\hat{m}(x, j, o)$ of the function which maps from covariate profile $x \in \mathcal{X}$, treatment condition $T_j \in \mathcal{T}$, and outcome of interest $o \in \mathcal{O} = \{\text{cost, benefit}\}$ to the expected realization of the outcome. We note that any class of statistical learning algorithm can be used derive $\hat{m}$; we select Random Forest for demonstrative purposes. Algorithm 1 describes the process of building machine learning models that represent the cost and benefit outcomes of the control group and the treatments groups.

2.4.2 Causal inference and treatment effects

There are two ways to estimate outcomes in each treatment condition. The first approach is to model the baseline outcomes and treatment effect jointly–i.e. $\hat{B}_i(j) = \hat{m}(x_i, j, B)$–where as the second approach, which we select in this work, explicitly separates (from the baseline outcome) and estimates the treatment effect heterogeneity–$\hat{B}_i(j) = \hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$–as shown in (2.10). In Section 2.3.3, we provide commentary from the literature why the latter approach can be preferable for accurately modeling heterogeneity, irrespective of the chosen estimation algorithm (Imai & Ratkovic 2013, Imai & Strauss 2011). Moreover, it has been argued from a practical and empirical perspective that utilizing single tree methods–e.g., Causal Tree or Transformed Outcome Tree (Athey & Imbens 2016)–to capture treatment effects can lead to estimations that are partly fit on the error terms, which eventually find treatment effects even when they are not present (Berry et al. 2016). Therefore, Berry et al. (2016) proposes a two-stage process (consistent with the $\hat{\tau}^B_{\text{CATE}}(X_i, j)+\hat{B}_i(0)|X_i$ approach to estimation) which is demonstrated to have better performance than Causal Tree (Athey & Imbens 2016).

For our empirical analysis we follow the suggestions from the prior literature and therefore, select to model treatment effect heterogeneity separately from the baseline outcome ($\hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$). We utilize an approach similar to that described in Berry et al. (2016) to model the individual level treatment effects.[3] Assuming that the model accuracy metric computed on the test data indicates that the models ($\hat{m}$) are well estimated, we therefore can trust their predictions. Moreover, armed with $\hat{m}$, we can (in a sense) overcome the "the

[3]For the sake of completeness, in Section 2.6.2 we also conducted our analysis using Causal Forest (Wager & Athey 2018), which is an ensemble with a Causal Trees base learner. Therefore, we use it to represent models that estimate the baseline outcomes and treatment effect jointly.

fundamental problem of causal inference"–i.e., a subject is observed in at most one condition–because $\hat{m}$ provides estimates for each subject's outcomes across all conditions. Furthermore, because of randomization these estimates are unbiased. Therefore, for each subject in the training data, we predict their outcome under each condition and obtain subsequent estimates of individual level treatment effects: $\hat{\tau}_i(j,o) = \hat{m}(x_i,j,o) - \hat{m}(x_i,0,o) \;\; \forall i,j,o$. Note that $\hat{\tau}_i(j,o)$ is an unbiased and consistent estimate of $\tau_i(j,o)$, the true, unobservable, individual-level treatment effect for subject i, given treatment condition T_j and outcome o. Finally, using 5-fold cross validation, from the $\hat{\tau}_i(j,o)$ we learn a model to capture the heterogeneity in individual-level treatment effects, for each outcome of interest and treatment condition. We note that any class of statistical learning algorithm can be used in this context as well; we select decision trees for demonstrative purposes. The collection of these models form (2.8): an estimate $\hat{\tau}_{\text{CATE}}(x,j,o)$ of the function which maps from covariate profile x and treatment condition T_j to the expected treatment effect for outcome o. This process is in the same category of well documented approaches in machine learning, where a strong–but potentially difficult to interpret–learner (e.g., Random Forest) is utilized to best capture the underlying complex function, and then a more interpretable–but potentially weaker–learner (e.g., decision tree) is utilized to summarize the complex learner's decision function (Domingos 1997). Evaluation of the predictive modeling exercise described between Section 2.4.1 and 2.4.2 is outlined in Algorithm 1. The ultimate goal is to ensure that our models are rich enough to capture the (within-heterogeneity of the) data generation process for all treatment conditions.

Algorithm 1: HTE Algorithm

input : training data ($\mathcal{N}_1$), treatment condition set ($\mathcal{T}$)

/* $\mathcal{N}_1$ is composed by benefit outcomes ($\vec{B}$), cost outcome ($\vec{C}$), covariates ($\boldsymbol{X}$), treatment condition ($\vec{W}$), and set of Outcomes $o \in \mathcal{O} = \{$cost, benefit$\}$ */

for T_j *in* $\mathcal{T}$ **do**

  $D_j \leftarrow \{(B_i, C_i, X_i) | W_i = j\}$; // separate training data by treatment condition

  for o *in* $\mathcal{O}$ **do**

    Use D_j to learn $\hat{m}(x, j, o)$–via cross validation, optimizing MAE, for hyper parameter tuning–which estimates $m : (x, j, o) \mapsto \mathbb{R}$ the true potential outcomes function;

  end

end

for T_j *in* $\mathcal{T} \setminus \{T_0\}$ **do**

  for o *in* $\mathcal{O}$ **do**

    for i *in* $1 \dots |\mathcal{N}_1|$ **do**

      Compute $\hat{\tau}_i(j, o) = \hat{m}(x_i, j, o) - \hat{m}(x_i, 0, o)$, an estimate for the individual-level treatment effect in the outcome of interest;

    end

    Use $\{\hat{\tau}_i(j, o)\}_{i=1 \dots |\mathcal{N}_1|}$ to learn $\hat{\tau}_{\text{CATE}}(x, j, o)$–via cross validation, optimizing MAE, for hyper parameter tuning–which estimates $\tau_{\text{CATE}}(x, j, o)$;

  end

end

output: the functions $\hat{m}(x, j, o)$ and $\hat{\tau}_{\text{CATE}}(x, j, o)$

2.4.3 Model accuracy

In order to evaluate the accuracy of our chosen estimated functions (i.e., $\hat{m}$ and $\hat{\tau}_{\text{CATE}}$) we turn to the test data subjects ($\mathcal{N}_2$). More specifically, we measure **mean absolute error** (MAE) which is a standard metric in model fitting to measure the similarity between the estimated and observed quantities of interests, e.g., cost, benefits, and therefore utilities. This is computed across all subjects, comparing the observed quantity of interest to the estimate of this quantity, given the observed treatment condition for the subject. In our context, values of mean absolute error (on the test data) demonstrate to what degree the model is able to capture the true data-generating process of the underlying treatment arms. Furthermore, a small value of MAE is desirable, as it implies that the models are able to capture what true utility a subject would experience under any condition. It is still an open question what is a sufficiently small value of MAE on a given dataset to signify a sufficiently high quality of fit. We treat this threshold value as an external parameter, leaving it to the researcher and the domain expert to determine if the MAE signals a sufficient quality of fit for their purposes. It is possible that researchers will be unable to find a modeling technique that can produce a fit of sufficiently quality. If so, this provides evidence that continuing with the subsequent steps of our process is likely to not be fruitful; there is little practical value in attempting to perform inference (and eventually generate prescriptions) with highly error-prone estimations.

While we use the standard MAE metric for this purpose, there is nuance (detailed in the Algorithm 2) around how the outcome predictions incorporate the treatment effects for the individuals in their treatment arms. After computing MAE, we can use the user defined cutoff (or parametric threshold) and proceed further if the MAE values are sufficiently small.

Algorithm 2: Validation of HTE models (and Within-Heterogeneity)

input : test data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), potential outcomes function ($\hat{m}$), treatment effects function ($\hat{\tau}_{\text{CATE}}$)

for $i \; in \; 1 \ldots |\mathcal{N}_2|$ **do**

  $\hat{B}_i(0) \leftarrow \hat{m}(x_i, 0, b)$, $\hat{C}_i(0) \leftarrow \hat{m}(x_i, 0, c)$;

  for $T_j \; in \; \mathcal{T} \setminus \{T_0\}$ **do**

    $\hat{B}_i(j) \leftarrow \hat{B}_i(0) + \hat{\tau}_{\text{CATE}}(x_i, j, b), \quad \hat{C}_i(j) \leftarrow \hat{C}_i(0) + \hat{\tau}_{\text{CATE}}(x_i, j, c)$;

  end

end

$$\text{MAE}_b = \frac{\sum_{i=1}^{|\mathcal{N}_2|} |B_i(W_i) - \hat{B}_i(W_i)|}{|\mathcal{N}_2|} \quad \text{MAE}_c = \frac{\sum_{i=1}^{|\mathcal{N}_2|} |C_i(W_i) - \hat{C}_i(W_i)|}{|\mathcal{N}_2|}$$

`/* MAE captures the error in the observed condition` (W_i)`, across all` i `*/`

output: $\hat{B}, \hat{C}, \text{MAE}_b, \text{MAE}_c$

2.4.4 Observed Utility Rank Condition (OUR)

Following from above, we have been able to perform causal inference across the entire response surface, modeling the treatment effect across the space of covariate profiles. We now aim to evaluate if, for individuals, there exists heterogeneity in the effect across treatments. Exploiting such heterogeneity in the effects of the treatments is critical for future prescriptions. Again we turn to the test data subjects, and propose a new metric to compute: **observed utility rank** (OUR). More specifically, for each subject, we compute and rank (within subject) the expected utility under each of the treatment conditions T_j ($j = 1, \ldots, J$). We then

partition the test data into J groups (P_j), each capturing subjects who received their (estimated) j^{th} best treatment condition, as measured by individual utility. The value of this measure is built on the fact that the test data is also a random sample from the population of interest, devoid of selection bias, and that the models have captured the data generating process of the treatment conditions well. Given these conditions, the individual level estimation and rank of utility indicate the degree to which there are differences between the various treatment conditions for a given individual. Furthermore, the observed average utility of each partition P_j is an unbiased estimate of the expected utility from placing a subject in their j^{th} best condition. Therefore, if there exists exploitable heterogeneity in the effect across treatments, and the ability of the models to capture it with sufficient accuracy, we expect that the utility of partitions P_j will be monotonically decreasing with j, and the differences between the partitions will be significantly different. In that case, we conclude that OUR metric is satisfied. If this does not occur–i.e., there appears to be no significant improvement in utility (on average) by assigning a subject to an individually preferable condition–this likely indicates one or both of the two possibilities. First, the accuracy of models that capture the outcomes of interest is poor. Second, there is not sufficient heterogeneity to exploit for prescriptions. In this case, the practitioner should strive to build models that best capture the outcome distribution. If the model with best MAE still fails to satisfy the OUR metric, it implies one or both of the conditions have likely occurred and progress should not continue.

Essentially, OUR captures the relative strength of heterogeneity present in the data and the model's ability to capture it. If the heterogeneity is small but the models are accurate enough to capture it, OUR is satisfied. If the heterogeneity is sufficiently large, even if the models are less accurate, OUR is still satisfied

because the models can still provide a proper ordering of the treatment.

This process of measuring OUR is described in Algorithm 3.

Algorithm 3: Measuring the Observed Utility Rank (OUR)

input : test data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), benefit values ($\hat{B}$), cost values ($\hat{C}$), treatment condition (W_i)

$\mathrm{P} \leftarrow (\{\emptyset\}_1, \{\emptyset\}_2, \ldots, \{\emptyset\}_{|\mathcal{T}|})$; `// vector of` $|\mathcal{T}|$`-many empty set placeholders`

for i *in* $1 \ldots |\mathcal{N}_2|$ **do**

  $R_i \leftarrow |\mathcal{T}|$; `// default to the lowest observed rank`

  for T_j *in* $\mathcal{T}$ **do**

    $\hat{U}_i(j) \leftarrow \hat{B}_i(j) - \hat{C}_i(j)$; `// compute estimate of utility`

    $R_i \leftarrow R_i - \mathbb{1}\{\hat{U}_i(W_i) \geq \hat{U}_i(j)\}$; `// update rank of observed condition`

  end

  $P_{R_i} \leftarrow P_{R_i} \cup \{i\}$; `//` add i `to index` R_i `of P`

end

for j *in* $1 \ldots |\mathcal{T}|$ **do**

  $\mathrm{OUR}_j \leftarrow \frac{1}{|P_j|} \sum_{i \in P_j} \hat{U}_i(j)$; `//utility if observed condition is rank` j

end

output: OUR

We proceed to empirically validate our framework using two different real-world decision making scenarios, one successful in that it succeeds at each step, and the other that fails the OUR stage. We consider it informative to demonstrate and discuss both scenarios, as we believe there is value in learning from

experiments that both succeed and fail to be exploitable by our framework.

In the Discussion and Conclusion section, we acknowledge that there are other procedures for model validation like the one mentioned in (Hitsch & Misra 2018), in which a follow-up experiment is conducted to validate the performance of existing models.

2.5 Empirical analysis

In this section we empirically demonstrate the utility of our proposed prescriptive analytics framework for determining which subjects to target with which interventions, in order to optimize organizational goals. We use data from two different randomized field experiments–one from public policy in the context of stimulating blood donations, and one from marketing domain in the context of referral marketing–to provide a sense of the generality of our framework. These experiments serve as case studies of our frameworks' ability to identify the existence of exploitable heterogeneity, and demonstrate its performance in real-world decision making.

The first experiment is conducted in collaboration with a major blood bank in China, investigating the impact of different mobile messaging interventions in motivating blood donations (Sun et al. 2019); while the second experiment is conducted on the online platform Collage.com, investigating the impact of various call-to-actions on activating referrals and subsequent purchases (Jung et al. 2019). In both experiments the organizations are budget-constrained in their attempt to maximize utility, namely social welfare and profit respectively. Furthermore, each organizational decision maker is facing ex-ante unknown costs and

	Benefit Matrix				Cost Matrix				Utility Matrix			
Subject	B_0	B_1	B_2	B_3	C_0	C_1	C_2	C_3	U_0	U_1	U_2	U_3
1	26	120	1	100	3	92	0	90	23	28	1	10
2	16	24	0	15	10	12	0	20	6	12	0	-5
3	5	45	15	16	5	25	11	1	0	20	4	15
4	0	15	100	55	9	5	50	10	-9	10	50	45

Table 2.1: : An example set of Benefit, Cost, and Utility matrices, where the rows are future (or test data) subjects, and the columns correspond to four different treatment conditions.

benefits from the multiple policies they can deploy. Therefore, these experimental settings present realistic case studies in which we can explore and evaluate the ultimate effectiveness of prescriptive strategies. To empirically validate our framework, we benchmark its HTE-EST procedure from (2.11) against two other prescriptive approaches within the causal inference paradigm that currently dominate practice. These are the Average Treatment Effect (ATE) which assigns all future subject to the policy that has the highest estimated population average treatment effect (increase in utility), and uplift modeling (UM) which assigns each future unit to the condition estimated to provide the largest individual increase in utility. We show below that, under the exploitable conditions of within- and between-treatment heterogeneity of costs and benefit, the prescriptions provided by traditional approaches fall short. They are either too nonspecific or too myopic, and therefore fail to capture attainable utility. In contrast, our analytics framework amalgamates randomized experiments, causal inference, machine learning, and optimization, in order to overcome these limitations.

2.5.1 Comparing prescriptions

In order to compare the prescriptions generated by HTE-EST–built atop integer linear programming (ILP)–to the traditional approaches–i.e., ATE, UM–we can again turn back to our test data, in conjunction with the $\hat{\tau}_{\text{CATE}}$ and $\hat{m}$. More specifically, we will construct matrices–where rows are the test data subjects and columns are treatment conditions–of expected benefit and cost (see Table 2.1). We note that although we utilize estimates of the expected value in our matrices, the error in these estimates is not taken into account. Therefore, in the Appendix we compare the prescriptive results under various procedures for propagating the error. Each of the prescriptive procedures will be given these matrices, and subsequently will provide a set of subject assignments given organizational (budget) constraints. We will then compare the amount of utility each of the methods' prescriptions are able to obtain subject to the constraints. For reference, in our empirical evaluations on the datasets described below, we depict the mean and variance of the utility each method obtains across thirty random partitions of training and test data.

It is also important to note that since each prescriptive technique is provided with the same estimates of individual cost and benefit, their comparative performance is still informative concerning their relative ability, even if the true values are not perfectly estimated. Notwithstanding, we expect our framework's HTE-EST to yield better prescriptions given a set of matrices. Let us take Table 2.1 as an illustrative example, with a budget constraint of 50. We can see that ATE would prescribe putting each subject into condition 1, as it has the highest average utility, leading to 42 in utility (84 in benefit) at a cost 42; ATE is unable to service subject 1 as its cost for treatment 1 would exceed the budget constraint. UM, provides additional flexibility as compared to ATE by selecting the best individual

condition for each subject. Therefore it would prescribe conditions $(1, 1, 1, 2)$ for the subjects respectively, leading to 50 in utility (100 in benefit) at a cost 50, by serving only subject 4. HTE-EST, is the most flexible, as it can assign any condition to any subject, in an attempt to maximize total utility. Therefore it would prescribe conditions $(0, 1, 1, 3)$ for the subjects respectively, leading to 100 in utility (150 in benefit) at a cost 50. Essentially, ATE's prescription are based on which column would have the highest average utility, while UM's prescriptions are based on which column, for each row separately, would have the highest utility.

From our simple example, it becomes evident that ATE and UM are sub-optimal, with ATE being overly general and UM being overly myopic. These methods fail to consider the cost required to achieve the utilities that drive their prescriptions. HTE-EST, however, offers prescriptions that can select any treatment for any subject, taking into consideration the budget constraint and cost associated with the utility being gained. As a result, HTE-EST is able to avoid ATE's overly general prescriptions–with individual prescriptions–and UM's overly myopic prescriptions–by realizing that subjects 1's and 2's individually high utility conditions come at too high a cost. Unlike UM, HTE-EST can utilize the budget saved by placing subject 4 and subject 1 into an individually sub-optimal condition, to obtain additional benefit from being able to serve other subjects, leading to more globally optimal prescriptions.

2.5.2 Blood donation case study

The first field experiment we use to illustrate our framework for optimal utilization of heterogeneous treatment effects is concerned with motivating blood donation. We use the experiment to illustrate how our methods may be extended to non-profit applications and accommodate considerations of policy makers on

Outcome	MAE	St. Dev
Benefit	8.22	0.07
Cost	0.10	0.01
Utility	7.26	0.07

Table 2.2: This table reports the mean absolute error (MAE) of the Random Forests models that estimate the true data generating process of the treatment conditions, in the blood donation experiment, across subjects.

non-monetary utility of individuals. The research context, experiment design and summary statistics are detailed in Section 4 and 5 of Sun et al. (2019). In collaboration with a centralized blood bank in a major city in China, the researchers conduct a large randomized field experiment to test the effectiveness of different interventions in motivating blood donations from subjects and their friends. Specifically, 80,000 eligible subjects are chosen, from the pool of past donors to the blood bank, and randomly assigned into several treatment conditions. The first condition is a control group with 14,000 subjects. For the remaining treatment groups, the researchers send a mobile message and vary its content across groups. The message content explores two treatments to overcome the hurdle of blood donation. The first message condition is a behavioral intervention (with 22,000 subjects), in the form of a message that reminds a potential donor to come to donate or to donate together with friend(s). The second treatment (with 44,000 subjects) informs the potential donor they will receive an economic reward for donation.[4] The details of the mobile messages for the test groups using the behavioral or economic interventions, the choice of sample, the time horizon, the variable collected; as well as randomization check, the summary statistics, and the main results for the experiment can be found in Sun et al. (2019) Section 4

[4]While only one condition contained information of the economic reward, all eventual donors, regardless of treatment condition, receive the same economic reward for donation.

and 5. This original study of the data is concerned with the ATE, i.e., identifying the treatment condition that would motivate the most donations from the subjects. For the purpose of our study, we directly use this field experiment data and further augment it with rich archival data, including demographics (age, gender, education, occupation, marriage status, resident status, and health indicators) and donation history (across 10 years). We follow the practice of the blood bank to define the benefits and costs associated with each donation. Specifically, the benefit is determined by the volume of the each donation (1, 1.5, or 2 units * 220RMB/unit). Similarly, the cost is associated with the reward given out to each donor (30RMB gift for 1 unit, 40RMB for 1.5 and 50RMB for 2 units). An additional cost of 1RMB is also incurred to send each mobile message, which is therefore not applicable to control treatment group. We focus on whether and how the blood bank may leverage the heterogeneity in the treatment effect and design optimal policy at an individual level, as well as compare the performance across different prescriptive policies. The setting features a low response rate in the experiment (about 1%) and the potential for rich heterogeneity across subjects given the large number of covariates available. Such a setting mimics the characteristics of many non-profit and for-profit practical applications such as charitable donation solicitation and digital advertising.

Table 2.2 shows the mean absolute error (MAE) from the application of Algorithms 1 and then 2 of our framework for each treatment condition and outcome of interest. As desired, we observe that the error from our models are quite small on average, and tend to vary closely around their small error values. Figure 2.2 depicts the observed utility rank (OUR) from Algorithm 3, for our models. The x-axis captures the (estimated) rank of the treatment a subject is random allocated into and the y-axis captures the average utility actually experienced by the blood

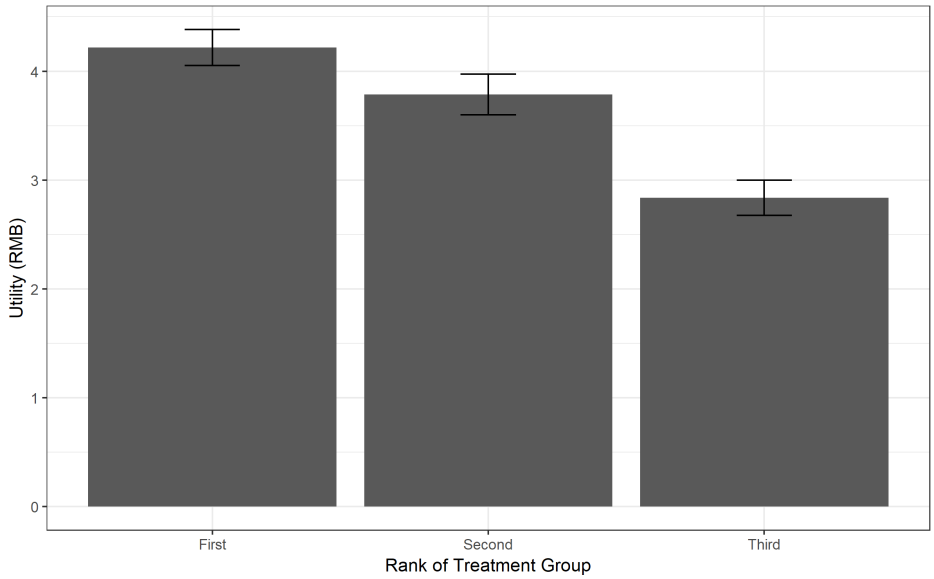

Figure 2.2: This figure captures the observed utility ranking (OUR) graphically, demonstrating the existence of heterogeneity across treatment conditions for individuals (or subpopulations). Subjects generate significantly higher utility on average when placed in more optimal conditions as estimated by our models.

Rank of Treatment allocation	First	Second
Second	0.0025	-
Third	<2e-16	8e-11

Table 2.3: This table shows the p-values for pairwise comparison of average Utility actually experienced by the blood bank for different ranks of treatment allocations.

Allocation	Control	Treatment 1	Treatment 2
Actual (Random)	2484	3888	8022
Individually Optimal	4285	2553	7556
Average Treatment Effect	0	0	14394

Table 2.4: This table servers as a comparison between the allocation of subjects under their actual (random) assignment, and the allocation under the (estimated) individually optimal assignment and ATE.

bank. As desired, we observe that there is a significant negative and monotonic relationship between the (estimated) quality of the treatment a subject receives and the expected utility achieved by the blood bank. We conducted pairwise t-test with Bonferroni adjustment between different ranks of treatment allocation and find that they are significantly different. We report the p-values in Table 2.3.

Given that our models appear to accurately capture the expected utility of each treatment for an individual, and the OUR metrics signals the existence of exploitable heterogeneity between treatments, we proceed to construct matrices of our outcomes of interest (as in Table 2.1). These matrices may assist in a deeper exploration of the experiment, as well as help evaluate the effectiveness of the prescriptive techniques. For example, Table 2.4 presents the number of subjects for whom each condition is their individually optimal allocation, based on the estimated utility, and compares the results with the actually observed random allocation and ATE. When there are no constraints, the optimal policy is simply

to put each subject into their individual best conditions–i.e., diverging from the treatment allocations defined in row 2 of Table 2.4 will lead to sub-optimal utility. Hence, although UM and HTE-EST will provide different prescriptions when facing an organizational constraint, when constraint free, both will output the same (individually optimal) set of prescriptions. However, recall that ATE allocates all subjects to the treatment condition that is optimal at an aggregate level; which, in this experiment, is treatment condition 2, corresponding to the monetary incentives. Furthermore, Table 2.4 shows that for a large number of subjects the control group is ideal; the implications are that the prescriptions from ATE lead to an individually sub-optimal allocation for many subjects, and the total utility obtained from ATE's prescriptions will be strictly less than HTE-EST (and UM). Therefore, the organization should prefer the allocation suggested by our framework, even with an infinite budget, because ATE will place many subjects in the more costly monetary incentive condition incurring large and unnecessary costs. Such costs originate from two sources: moving some subjects from control to the costly treatment group without enough gain, and moving some subjects from a treatment group to control (or the other treatment group) without properly motivating the subject.

Interestingly, our framework proves even more valuable when the decision maker has certain organizational constraints, such as a budget constraint. In Figure 2.3 and Figure 2.4 respectively, we compare two outcomes–the total expected utility realized and number of users served – by each prescriptive targeting strategy, for different levels of a budget constraint. With respect to expected utility, our framework's HTE-EST performs significantly better than the other strategies, across the range of budget constraints, followed by UM and then ATE.

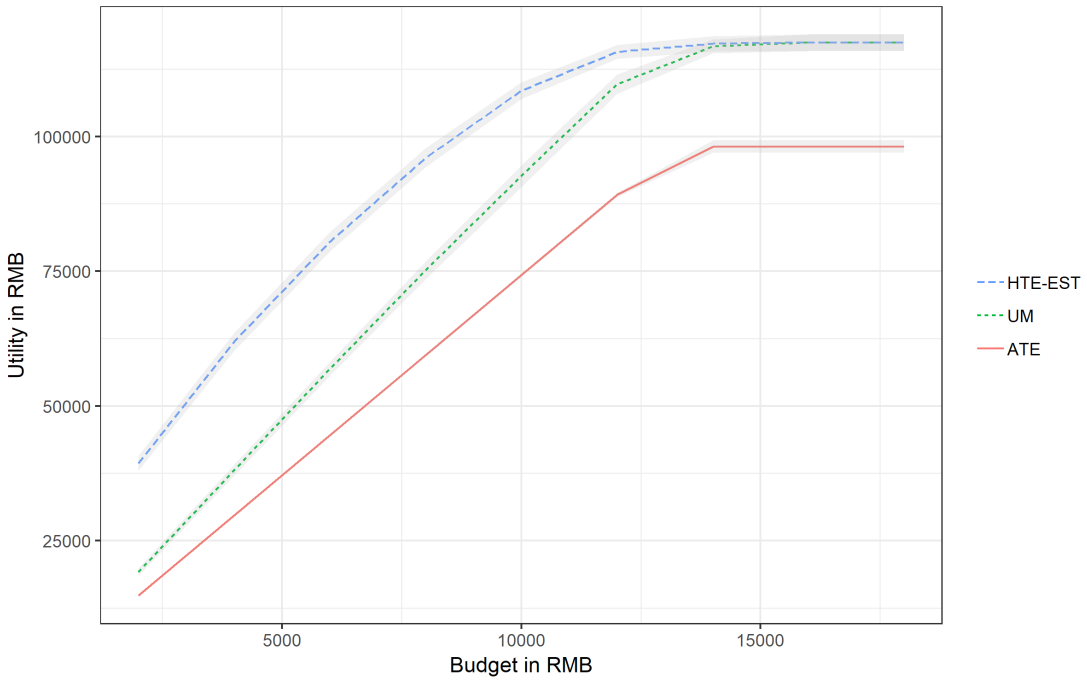

Figure 2.3: This plot shows the expected total utility generated by each prescriptive method, over a range of budget constraints.

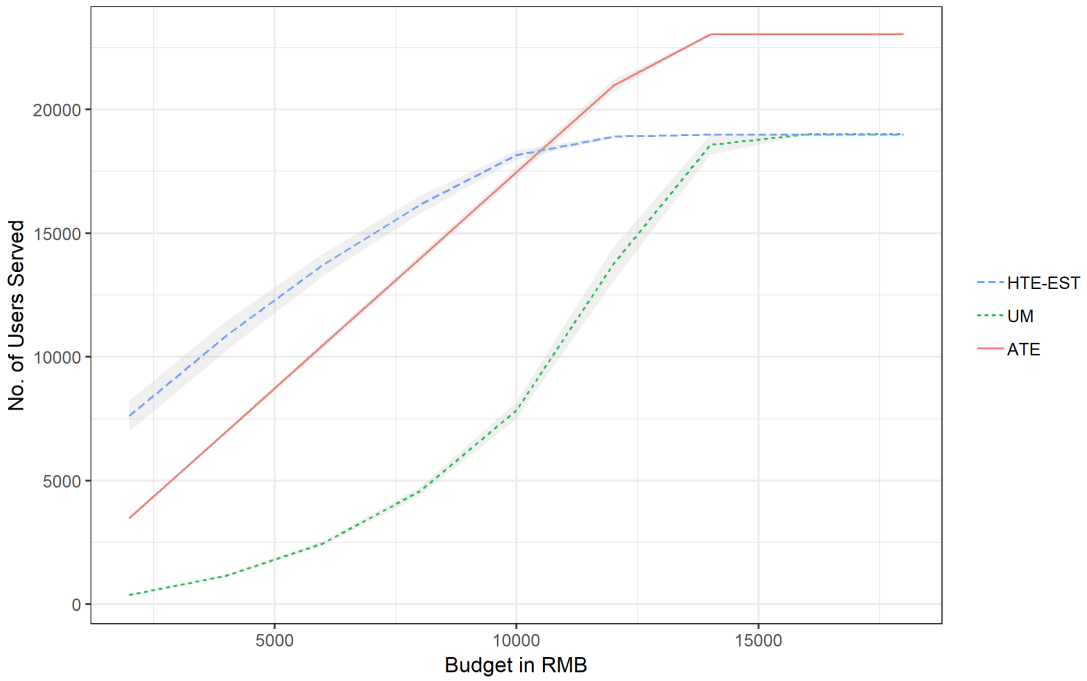

Figure 2.4: This plot shows the expected number of subjects targeted with a treatment, by each prescriptive method, over a range of budget constraints.

This behavior is rather intuitive, as UM is a greedy approach which simply allocates as many subjects as possible to their individual optimal condition. As a consequence, it will first serve subjects with high individual utility; therefore when the budget it small, it will be exhausted after a few subjects are treated (Figure 2.4). Conversely, HTE-EST may elect to place subjects in less optimal individual conditions, to preserve its precious budget in order to obtain additional total utility. As a result, we find that prescriptions from HTE-EST result in up to 240% and 340% higher utility than UM and ATE respectively. With a sufficiently large budget, the constraint is no longer relevant, and as described above, HTE-EST and UM would obtain the same (individually optimal) set of prescriptions. However, even at a large budget, HTE-EST still leads to a 19% increase in utility over ATE, suggesting how much utility is sacrificed by following the simplified guideline from ATE (row 3, instead of row 2 of Table 2.4).

HTE-EST's ability to strategically place a (specific) set of subjects in a secondary condition, enables it to capture a significantly larger number of subjects that can obtain additional utility from the treatment. Moreover, HTE-EST is also able to identify subjects for whom the control condition is optimal–i.e., no treatment will affect their ultimate decision to (not) donate. Figure 2.4 shows that these two facets together make HTE-EST fairly conservative with (and effective at) how many users to target with treatments. Such selectivity can be beneficial, as we know that following a prescriptive strategy like ATE and targeting indiscriminately users (with uninteresting treatments) can have negative consequences (Ghose et al. 2017).

In summary, our framework presents an ideal prescriptive strategy by combining machine learning, causal inference, and optimization to exploit heterogeneity

and parsimoniously identify precisely the right subjects to target in order to maximize the amount of blood donations received, at any level of budget constraint.

2.5.3 Referral marketing case study

The second field experiment we use to illustrate our method on optimal utilization of heterogeneous treatment effects is testing the optimal framing (altruistic, selfish, equity) that a firm can use to motivate product referrals. We use the example to illustrate how a profit maximizing firm may benefit from our method. The research context and experiment design are detailed in (Jung et al. 2019)). Specifically, the randomized field experiments is targeting existing customers of a large US based online platform called Collage.com. On this platform, users can design a collage by uploading photos and customizing the layout with the proprietary software tools. Once a user creates the layout, she can purchase various types of customized printed products, such as blankets, photo-books, canvases, etc. A large number of customers purchase a variety of products from the platform every day (with $22 million in revenue for 2015). The treatments in the experiment manipulate the framing, i.e. solely focus on varying the words and the emphasis of certain phrases, of the call-to-action while keeping all other aspects of the incentive and messaging constant across groups. The experiment offers both, the sender of the referral and the recipient, a free product voucher that comes with free shipping and no expiration date ($25 worth in listing value), which should appeal to many types of users. The randomized field experiment design allows clean identification of the causal effect of the framing of the calls-to-action on customers' referral decision, whether they share, to what extent they share, as well as on their induced referral outcomes as measured by the number of successful referrals. Specifically, in the experiment, the researchers randomly assign 100,000

customers, who have made purchases on the platform in the past, into four test groups (10,000 in control and 30,000 in each of the three treatment groups), and email each group with different calls-to-action. The data on customers' referral behaviors and outcomes are collected within a 5-week window after the experiment. Based on extensive discussions with the CEO and the marketing team of the partner company, we arrived at the following benefits and costs associated with each referral and voucher redemption. Specifically, the benefits of a referral comes from three parts: the impression value to the recipient(s) at \$0.01/referral, the user acquisition value if the recipient registered with Collage.com; at \$3/registration; and the transaction value if the recipient(s) actually made purchases, calculated based on the actual transaction amount. Conversely, the cost of a referral is associated with the redemption of the free product voucher by the sender, recipient(s), or both. Each redemption incurs an average cost of \$6. It is critical to note that relatively few customers who sent out or received a referral redeemed the free voucher. This serves as the source of heterogeneity on the cost side, with benefit side heterogeneity being resulting from the differential response to the promotion. For the purpose of our study, we directly use the field experiment data and augment it with rich archival data, including product characteristics, individual characteristics, customers' past purchases and their Net Promoter scores (NPS) [5]. The data from the large randomized experiments and the archival data allows us to identify the causal effect of different calls-to-action as well as to explore the heterogeneity underlying the treatment effect. As mentioned earlier, the key challenge is that *ex-ante* the marketer does not know whether any of these call-to-actions have a causal impact on profitability, let alone know whether the impact of the three different arms may be heterogeneous, as a function of existing

[5]The Net Promoter Score is an index used to measure loyalty and customer satisfaction, as the willingness to recommend a company's products or services to others.

Outcome	MAE	St. Dev
Benefit	7.72	0.006
Cost	7.81	0.023
Utility	0.38	0.044

Table 2.5: This table reports the mean absolute error (MAE) of the Random Forest models that estimate the true data generating process of the treatment conditions, in the referral marketing experiment, across subjects.

customer 'data' – characteristics, and behaviors.

We begin with Table 2.5 which shows the MAE of our Random Forest models, for each treatment condition and outcome of interest. As desired, we observe across all combinations that the error from our models are quite small on average, and tend to vary closely around their small error values. Furthermore, Figure 2.5 depicts the observed utility rank (OUR) for our models, where the x-axis captures the (estimated) rank of the treatment a subject was random allocated into and the y-axis captures the average utility actually experienced by subjects. In this case, we fail to observe a negative or monotonic relationship between the (estimated) quality of the treatment a subject receive and the expected utility achieved by the subject. More specifically, there is no observable significant difference in the expected utility across the treatment conditions. Given that Table 2.5 demonstrates a rather small amount of prediction error in our models, this leads to the conclusion that there is not a sufficient amount of individual-level heterogeneity across treatment conditions present, to exploit for optimization. As a result, we see no benefit in continuing to the optimization stage. We do speculate that the lack of between-treatment heterogeneity may be driven by three factors: 1) the drastic difference between the magnitude for the benefit (typically on the order

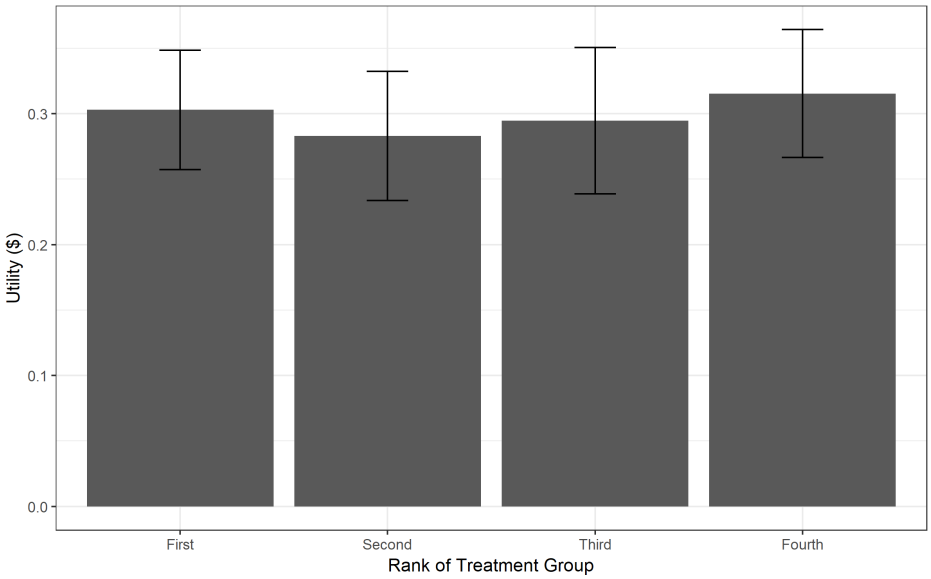

Figure 2.5: This figure captures the observed utility ranking (OUR) graphically. It demonstrates the absence of heterogeneity across treatment conditions for the same individuals. Subjects on average do not gain a significantly higher utility when placed in the optimal conditions as estimated by the models.

of cents) and that of cost (on the order of dollars);[6] 2) the inherent challenge in predicting the benefit measure, as revealed from the large variance in the summary statistics associated with the total spend and total discount measures in the Appendix; and possibly 3) the lack of useful and relevant individual-level features. We note that the lack of significant individual heterogeneity across treatments, provides no commentary on the existence of a (heterogeneous) treatment effect across the whole sample. Our three-stage prescriptive analytic framework allows us to identify the potential challenges at an early stage in the analytic cycle and save the efforts and expense on further optimization. It can be useful to the organization for future data collection efforts on customer features, which could be added to the framework and the OUR condition checked for between-treatment heterogeneity.

2.6 Robustness checks

2.6.1 Error simulation study

In the results above, we use our models' estimates of cost and benefit to construct the matrices necessary to carry out the prescription generation. This procedure takes these estimates as correct; although, we know the estimates from our models are consistent (under appropriate regularity conditions), they (as all estimates) have a degree of error. Therefore, we conduct a simulation analysis of the prescriptions generated from our process (Figure 2.6), attempting to (more properly) propagate the error into our prescriptions. Specifically, we take our final tree

[6] To no avail, we attempted a variety of different Machine Learning models, as well as up and down sampling to determine if the lack of heterogeneity a across conditions was a consequence of implicit modeling assumptions or imbalance.

models, and recognize that when providing the estimate for cost and benefit for a subject, it is common to use the mean of the tree's leaf the subject falls into. We now consider two additional ways to model the values in the leaves of our trees, which are then used to generate these estimates in our matrix of benefits and costs. First, "Gauss", where a Gaussian distribution is fit to the (training) data points that reside in a tree's leaves and 2) "Emp", where we utilize the empirical distribution of the (training) data points that reside in the leaves of the tree. Therefore, when we have a subject for whom we want to estimate benefit and cost values, we can now locate their leaf in the tree, and draw a value from the Gaussian distribution (Gauss) or empirical distribution (Emp) at this leaf. This allows the degree of uncertainty that exist to propagate into our matrices and eventual optimization[7].

Figure 2.6 shows the results for each of the three prescriptive analytics allocation approaches (ATE, HTE-EST, and UM) for the two error propagation approaches (Emp, Gauss), along with our original mean based approach for reference. We calculate confidence intervals by splitting entire data into training and test using a different seed ten times. From this graph we can make three observations about the final utility derived. First, regardless of which error-propagation procedure we use, the ordering of the prescriptive methods remain constant: HTE-EST, UM, ATE. Intuitively, if we consider the set of potential allocations under each of the methods, ATE is a subset of UM, which is itself a subset of HTE-EST; therefore, HTE-EST should always dominate in utility, followed by UM, and ATE. Secondly, within a given allocation approach, the ordering of propagation approaches, remains constant: Gauss, Emp, Mean. Intuitively, by using

[7]There are other ways of introducing error through global perturbation of nodes. We believe the qualitative pattern of the results is likely similar to our simulation below

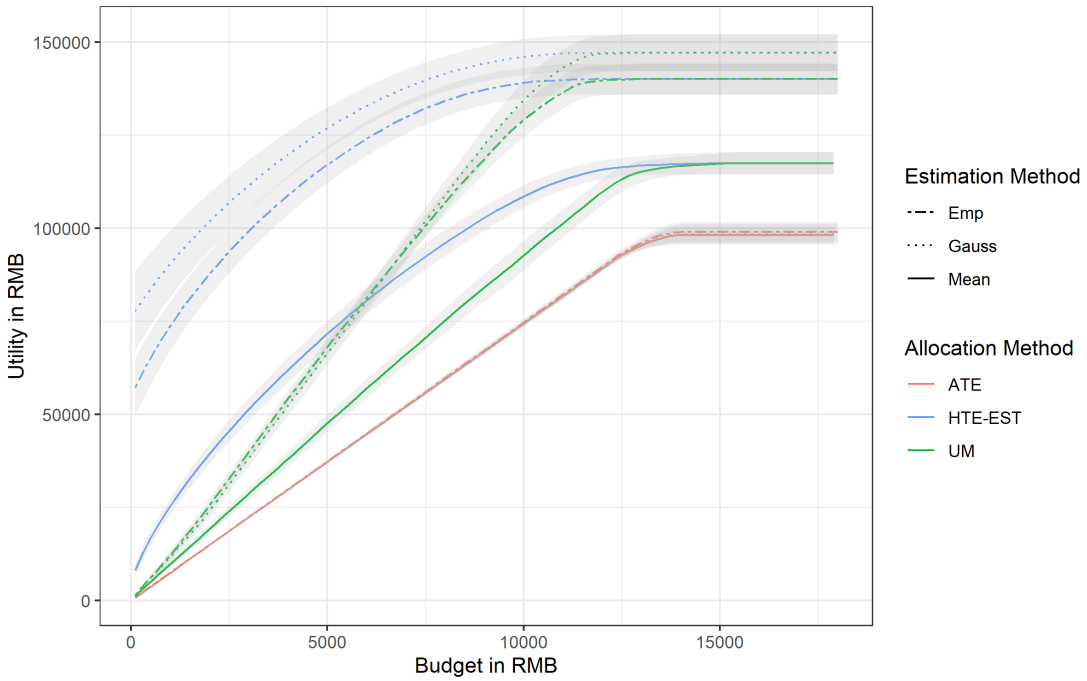

Figure 2.6: This plot shows the expected total utility generated by each allocation method for different estimation procedures, over a range of budget constraints.

the mean, as we did originally, when the prescriptive algorithms attempt to optimize allocations, they are forced to consider all customers who fall into the same leaf as providing the same cost and benefit. Essentially, units in the same leaf are exchangeable deterministically, from a estimated utility point of view, as their estimated values of interest have no sampling variation. Therefore, even though (technically) the optimization of allocations occur at the unit level, the values used to determine these allocations are solely determined at the leaf level. When we introduce estimation error into these values, we now introduce sampling variation in the utility measures at the unit level, and therefore units are now only exchangeable stochastically. This increased flexibility allows the allocation methods to optimize even more, ensuring that the units that happen to render large(r) amounts of utility are prioritized, even if others in their leaves render small amounts of utility. Conversely, those that render small(er) amounts of utility can even better be separated out and placed in the (low-cost) control condition, even if others in their leaves render high amounts of utility. Using the (parametric) Gaussian distribution at each leaf, as opposed to the less restrictive (nonparametric) empirical distribution tempers this a flexibility a bit, but the logic still carries over, and thus the observed ordering.

Finally, the confidence intervals around the utility estimates are larger when a form of error propagation is used. Intuitively, though the estimated mean utility for these error propagating regimes are higher, because of the flexibility in optimization, there is more uncertainty around these values, given the additional variation introduced at the unit level. We also note that this process is data-driven and replicable by any practitioner, to get a sense of how different their allocations and utilities would be under any of the allocation and error propagation regimes. Our recommendation for practitioners is to conduct an analysis similar to ours,

	MAE		MSE		ME	
Outcome	Causal Forest	Berry 2S	Causal Forest	Berry 2S	Causal Forest	Berry 2S
Benefit	7.77	8.22	1689.65	1687.14	0.43	-0.12
Cost	0.92	0.97	23.34	23.28	0.05	-0.01
Utility	6.86	7.26	1317.16	1315.55	0.39	-0.11

Table 2.6: This table reports the accuracy metrics (MAE, MSE, ME) of the models that estimate the true data generating process of the treatment conditions, by two different learners, across subjects.

comparing the results from the various allocation methods, but to follow the conservative mean prediction approach for their planning, recognizing the potential for further upside.

2.6.2 Comparing HTE learners

In Section 2.4.2, we operationalize heterogeneous treatment effects to understand the expected effect of placing a subject into a given treatment condition on the outcome of the subject. We use a two-stage procedure proposed by Berry et al. (2016) ($\hat{B}_i(j) = \hat{\tau}^B_{\text{CATE}}(X_i, j) + \hat{B}_i(0)|X_i$), which is demonstrated to have better performance than Causal Trees (Athey & Imbens 2016), a one-stage, single-tree learning procedure. The argument is that utilizing single tree methods–e.g., Causal Tree or Transformed Outcome Tree (Athey & Imbens 2016)–to capture treatment effects can lead to estimations that are partly fit on the error terms, which eventually find treatment effects even when they are not present.

We estimate treatment effects using Causal Forest (Wager & Athey 2018) (an ensemble method, with a Causal Trees base learner) and predict outcomes by adding control group predictions to treatment effect estimates. Table 2.6 shows the comparison of accuracy metrics–Mean Absolute Error (MAE), Mean Squared

Error (MSE), and Mean Error(ME)–on the out of sample, test data for Causal Forest and Berry-2S (the two-stage method proposed by Berry et al. (2016)). We first note that MAE and MSE for both the learners are comparable, paying attention to $MSE = Bias^2 + Variance$, i.e., mean squared error can be decomposed into the square of bias plus variance. Moreover, we note that ME is essentially an empirical measure of bias; and therefore we see that Causal Forest has higher bias (ME) than Berry-2S. This shows that Causal Forest may trade off additional bias for lower variance given their very similar MSE. The presence of this additional bias introduced by Casual Forest, in addition to its base learner's potential to partly fit on the error term, might result in inaccurate predictions of rank of treatment groups for each subject. Therefore, we perform the OUR analysis for Causal Forest predictions and see that the condition fails (Figure 2.7), recalling that the OUR metric evaluates a model performance, by obtaining an unbiased estimate of the utility it would provide across treatment conditions. It therefore appears that Causal Forest's additional bias undermines it's ability to properly rank treatment conditions and, following the logic presented in Section 2.4.4, we conclude that the Causal Forest learner fails to capture the exploitable individual heterogeneity between treatment group. Moreover, we have evidence that this between heterogeneity exists, given the two-stage procedure passed the OUR condition, and thus the outcome predictions from Causal Forest are not reliable for further budget constrained optimization.

2.6.3 Extension to the contextual bandits

In this chapter, we focus on leveraging data from existing large-scale randomized field experiments. An organization has often accumulated a large number of historical experiments and faces the objective of maximizing the utility in the

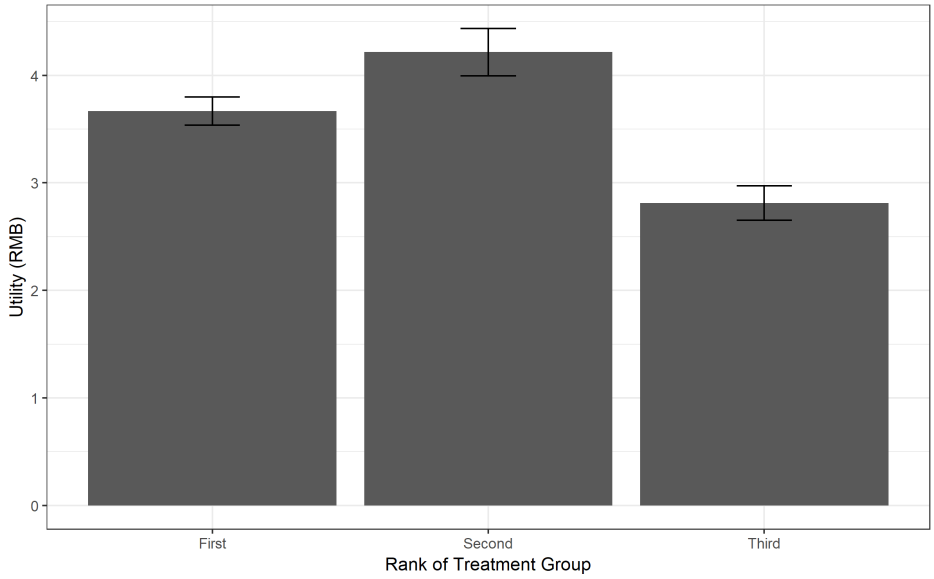

Figure 2.7: This figure captures the observed utility ranking (OUR) for blood donation experiment using Causal Forest learner

presence of a budget constraint when designing new policy interventions. With the use of Multi-armed bandit algorithms, our framework can be extended to situations when no historical experiment data is available. Multi-armed bandits are sequential experimentation procedures that use a combination of exploration and exploitation strategy with the goal of maximizing overall utility. In these procedures (Agrawal & Goyal 2012), the treatment allocation is determined sequentially every time an outcome or a batch of outcomes are observed (Zhou et al. 2019). Contextual Bandits (Langford & Zhang 2007) are an extension to the bandit approach where treatment allocations are done not just based on the values of outcomes but also on the context at hand. For example, a context can be the known characteristics of an individual being allocated. There are parallels between our framework and the Contextual Bandits approach in that the aim is to improve the overall utility of the allocation process and the usage of context in

order to make treatment allocations.

While contextual Bandits are continuous learning procedures with exploration and exploitation trade-offs, our framework is a three-stage discrete process which folds together exploration via randomized experiments, learning via Heterogeneous Treatment Effect estimation, and exploitation via optimal allocation of subjects in a step by step process. When there is no pre-existing experiment data, contextual bandit algorithms like LinUCB (Chu et al. 2011) may offer a nice way to jump start and optimize the utility while exploring the variation. When there is pre-existing data from randomized experiments, the approach we have presented may be readily used. In addition, we developed a modified version of the LinUCB algorithm that uses experiment data as an extension of our method[8]. The modified LinUCB also allows us to make a comparison between HTE-EST and LinUCB. The main idea in the modified LinUCB algorithm is to apply ridge regression on the randomized experiment data to calculate initial learning parameters and use them as inputs to LinUCB algorithm. After this stage, sequential allocations are done based on exploration-exploitation trade-off instead of pure exploitation done in HTE-EST. The detailed algorithm can be found at the end of the Appendix.

Figure 2.8 compares the total expected utility realized by HTE-EST and modified LinUCB for different budget constraints. We see that HTE-EST still performs better compared to the modified LinUCB. We believe that given the matrices of Benefit, Cost and Utility (as in 2.1), the expected realized utility generated by the prescriptions of HTE-EST will always be higher than any other prescription method because, for a matrix of non-changing predicted values, the solution (prescriptions) generated by the ILP optimization used by HTE-EST is theoretically

[8]Again, LinUCB does not need any pre-existing experiment data and offers an advantage in that scenario

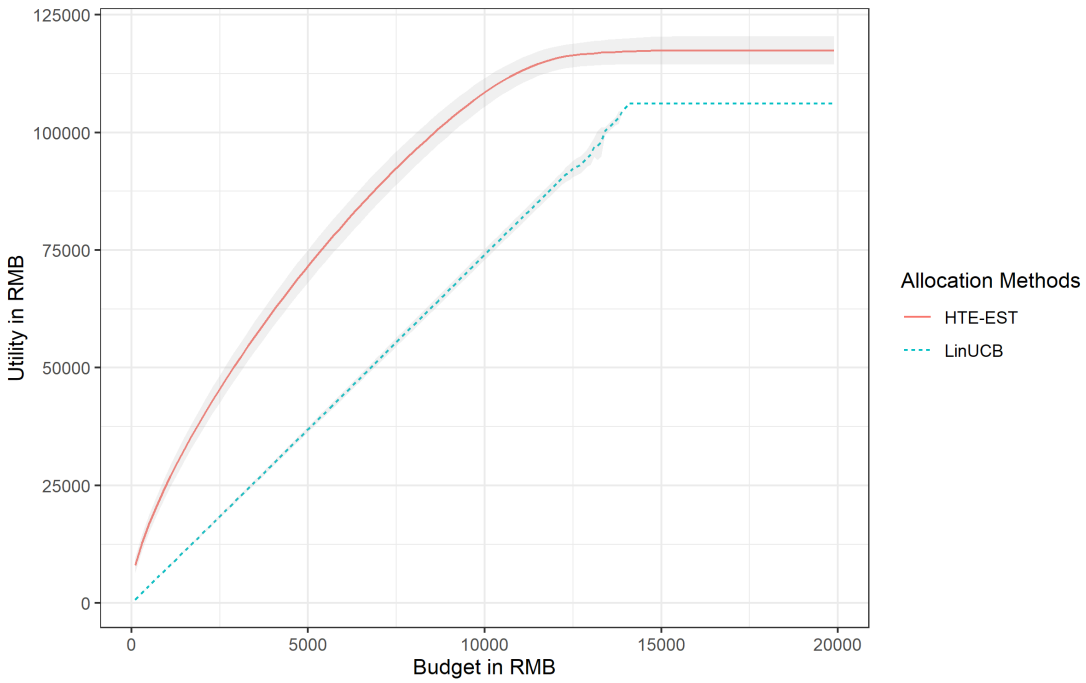

Figure 2.8: This plot shows the expected total utility generated by modified LinUCB algorithm and HTE-EST, over a range of budget constraints.

optimal. However, in real world setting, where LinUCB uses realized outcomes instead of predicted outcomes, we believe there is a possibility for realized utility to be higher than that of HTE-EST. Finally, from an implementation perspective, our framework requires less supporting infrastructure compared to bandit approaches and also is readily accessible to less sophisticated organizations.

2.7 Discussion and concluding remarks

In this study, we address the general problem of a budget-constrained decision maker facing, ex-ante, unknown costs and benefits from multiple policy levers that she can potentially deploy to optimize an organizational goal. We define and deploy a three-stage, prescriptive analytics approach as one that folds together the use of a) randomized field experiments and causal inference, b) machine learning to identify heterogeneity in treatment effects, and c) constrained optimization to optimally decide which subpopulations to treat with which policy levers to maximize profit.

We show that there is potential in combining the four pillars of analytics, as they are different components of a symbiotic process: **randomized experiments** enable the *exploration* of policy outcomes, **causal inference** enable the *estimation* of these policies' impact, **machine learning** enables the *analysis* and *prediction* of the impact variation across subpopulations, and **optimization** enables the *prescription* of a future optimal strategy under constraints.

We use data from two large-scale, randomized field experiments–one from public policy in the context of stimulating blood donations, and one from referral marketing–to illustrate the generality of our framework and demonstrate its performance in real-world decision making as compared to traditional approaches of

using the average treatment effect and uplift modeling. From the first study on donor recruitment, we find that that the prescriptions from our framework may result in up to an 340% and 240% increase in overall utility as compared to the prescriptions provided by ATE and UM respectively. We show that ATE and UM suffer from being either too nonspecific or too myopic, and therefore fail to attain the best utility. Our approach is aware of treatment heterogeneity and the budget constraint thus can optimally allocate the treatments across individuals. In addition, as demonstrated in the second study on referral marketing, we propose a new criteria–observed utility rank (OUR)–to detect heterogeneity in the effect across treatments. Identifying such heterogeneity in the effects of the various treatments are critical for future prescriptions. When between treatment heterogeneity is unavailable in the data, the organization can save the efforts and costs of further (likely unsuccessful) exploitation. Instead they may rather collect more data to explore, aided by exploratory analytics methods, for other sources of heterogeneity[9].

Our work is motivated by the observation that a vast majority of F500 companies have a *culture and capabilities deficit* to systematically make sense of the data they already have to create value (Hosanagar & Saxena 2017). Most approaches in the extant literature to solving the proposed decision-making problem of this chapter do not combine all the pillars of analytics. While many companies are starting to use predictions from historical (observational) data, such an approach does not address the (often critical) causal questions. Even, of the organizations

[9]Specifically, OUR condition is not satisfied when the between heterogeneity is too small to create a statistical difference or when the chosen models are too inaccurate to find the heterogeneity present. The latter is why we recommend searching for and selecting the model that minimizes MAE, using whatever means at the company's disposal. After minimizing MAE with present data, if either case still exists, the recommendation for the company would be to collect more data and re-select models to reduce the MAE. Both these steps will increase the probability of OUR being satisfied.

(and policy makers broadly) that subscribe to casual inference, few actually go beyond the average treatment effect, which effectively assumes the population has a homogeneous response to treatment. We believe that our proposed full spectrum approach to business analytics can make a significant contribution in reducing this deficit.

Future research can further extend our framework in a few ways. Firstly, we have studied two representative real-world decision scenarios, one for non-profit (blood donation) and one for for-profit (referral marketing). Future studies may explore decision scenarios in other contexts, such as pricing, new product development and testing, and user-interface design. We expect the importance of treatment heterogeneity and organizational constraints might differ across contexts and thus influence the improvement that the prescriptive analytics framework can help achieve. Secondly, in our current study, we treat the data acquisition process as exogenous and given. In other words, we assume an organization may utilize all existing data within the organization in the prescriptive analytic approach. However, with the emergence of a large number of third-party data vendors and data exchange (e.g. TowerData, BlueKai), organizations can increasingly acquire new customer data to improve advertising campaigns and expand customer base. When acquiring consumer data from external sources, it is important to decide which features of the customer data is of value and should be acquired. Our framework, especially the development of OUR criteria, may help organizations systematically evaluate the value of external data. Such data acquisition process would also help organizations better understand whether the lack of treatment heterogeneity (as in our second study) is due to limited customer features available or intrinsic (unpredictable) nature of the decision scenario. In the same vein,

firms can (and should) take into account the potential heterogeneity when thinking of the size of the experiment. A larger experiment, with more sample, may not only boost the power to detect the main effect, but also create variation for the estimation of heterogeneous treatment and further optimization. Future research may take the sample size as an experiment design choice and investigate its implication to the estimation of heterogeneity and optimal policy. Finally, our study validate the performance of the proposed approach using the data from the two large-scale experiments. Future research can directly compare the performance of our suggested optimal policy with other personalized policy using a field test.

We envision that the integrated prescriptive analytic framework can be further enriched, customized and deployed in a wide range of non-profit or for-profit, digital native or traditional, and established or emerging organizations. We hope that our study serves as a valuable first step for such future efforts.

Appendix

Summary statistics of the two case studies

	N	Mean	St. Dev.	Min	Max
No. of past orders	99,881	2.68	4.46	1	499
Total spend (USD)	99,881	77.87	180.13	0	37,794.07
Total discount (USD)	99,881	206.58	505.43	−0.51	102,011.1
Total refunded (USD)	99,881	2.65	22.91	0	2,383.78
Last purchase (days)	99,881	368.01	149.49	124	559
NPS	99,881	1.84	3.75	0	10
No. of NPS comments	99,881	0.28	0.7	0	37
Firm cost (USD)	99,881	0.02	0.73	0	60
Firm gain (USD)	99,881	0.32	8.181	0	781.8
Firm utility (USD)	99,881	0.29	8.17	−47.95	781.8

Table 2.7: This table reports the summary statistics of variables and outcomes in the referral marketing experiment.

	N	Mean	St. Dev.	Min	Max
Weight (kgs)	57,575	65.06	11.08	45	115
Age	57,575	28.26	8.88	18	61
Total past donation (ml)	57,575	483.75	407.64	0	6,400
Most recent donation (months)	57,575	15.04	6.75	0	116
Num voluntary donation	57,575	1.02	1.38	0	69
Num group donation	57,575	0.35	0.68	0	9
Num mutual donation	57,575	0.09	0.29	0	2
Num plasma donation	57,575	0.02	0.64	0	68
Firm cost (RMB)	57,575	1.32	4.91	0	52
Firm gain (RMB)	57,575	4.22	41.66	0	440
Firm utility (RMB)	57,575	3.72	36.77	0	389

Table 2.8: This table reports the summary statistics of variables and outcomes in the blood donation experiment.

Pseudo code for modified LinUCB algorithm

Algorithm 4: LinUCB Implementation

input : training data ($\mathcal{N}_1$), testing data ($\mathcal{N}_2$), treatment condition set ($\mathcal{T}$), $\alpha \in \mathrm{R}^+$ (Hyper-parameter for LinUCB algorithm), Budget (M)

/* $\mathcal{N}_1$ is composed of benefit outcomes ($\vec{B}$), cost outcome ($\vec{C}$), Utility outcome ($\vec{U} = \vec{B} - \vec{C}$), covariate matrix ($\mathbb{X}$), and treatment condition ($\vec{W}$) */

/* $\mathcal{N}_2$ is composed of covariate matrix ($\mathbb{X}$), predicted outcomes ($\hat{\vec{B}}, \hat{\vec{C}}, \hat{\vec{U}} = \hat{\vec{B}} - \hat{\vec{C}}$) for each treatment group (Output from Algorithm 2) */

/* We use i to index variables to individual units, and j to index variables related to treatment conditions */

// Training

for $T_j \; in \; \mathcal{T}$ **do**

  $\vec{U_j} \leftarrow \{U_i \mid W_i = j \; \forall \; i\}$

  $\mathbb{X}_j \leftarrow \{\vec{X_i} \mid W_i = j \; \forall \; i\}$; // separate training data by treatment condition

end

for $T_j \; in \; \mathcal{T}$ **do**

  $\mathcal{A}_j \leftarrow \mathbb{X}_j^T \mathbb{X}_j + I$

  $\vec{b_j} \leftarrow \mathbb{X}_j^T \vec{U_j}$

end

// Testing (Allocation procedure with budget constraints)

for $i \; in \; 1 \ldots |\mathcal{N}_2|$ **do**

  $m_r \leftarrow M$

  do

    for $T_j \; in \; \mathcal{T}$ **do**

      $\hat{\vec{\theta}_j} \leftarrow \mathcal{A}_j^{-1} \vec{b_j}$

      $\vec{p}_{i,j} \leftarrow \hat{\vec{\theta}}_j^T \vec{X}_i + \alpha \sqrt{\vec{X_i}^T \mathcal{A}_j^{-1} \vec{X_i}}$

    end

    Choose treatment $T_j =_j p_{i,j}$ with ties broken arbitrarily and observe predicted utility $\hat{U}_i(j)$ and cost $\hat{C}_i(j)$ for that treatment.

    $\mathcal{A}_j \leftarrow \mathcal{A}_j + \vec{X}_i \vec{X_i}^T$

    $\vec{b_j} \leftarrow \vec{b_j} + \hat{U}_i(j) \vec{X}_i$

    $m_r \leftarrow m_r - \hat{C}_i(j)$; // Remaining Budget

  while $m_r > 0$;

end

output: Predicted Utility $\hat{\vec{U}}$ and Predicted Cost $\hat{\vec{C}}$

Chapter 3

Turning referral data into assets: Evidence from an RCT on algorithmic vs. social targeting

3.1 Introduction

Digital transformation is defined as the use of digital technologies to transform every area of an existing business. A 2019 McKinsey quarterly report (Bughin et al. 2020) found that the top 10% of the digitized incumbents earned as much as 80% of the revenue in their industries. One of the key strategic themes of digital transformation is to turn the data the company collects into assets that help the company grow (Rogers 2016). For example, Netflix transformed from an online DVD rental company to the streaming giant it is today. It collects granular data on user watch behavior and preferences, then the data are used to make decisions on what TV shows to produce and which celebrities to cast. In 2017, Netflix renewed 93% of original TV shows after the first season, compared

to 35% for cable television. There is no question that companies can use raw data to optimize processes, drive innovations, and create new products; however, there is a dissonance between the pace at which the data are created and the pace of development of methods used to make sense of the data and turn them into actionable insights (Economist 2010). Historically, algorithmic development and managerial decision-making research have been disconnected. Addressing this gap and using algorithmic development to solve critical managerial challenges will help turn data into assets.

Companies test offers, packages, and incentives to better understand customers. These campaigns generate data that may contain value beyond their initial purpose. In this chapter, we examine digital transformation through the context of referral marketing. With the rise of the Internet, social networks, and social media, electronic word of mouth plays an important role in driving purchase decisions (Chevalier & Mayzlin 2006). Referral-based programs that leverage existing customers' social networks to reach potential new customers are particularly effective (Bapna & Umyarov 2015, Jung et al. 2020). Schmitt et al. (2011) show that customers acquired through referral programs are 16% more valuable than those from other channels. As a result, referral programs that promise incentives to existing customers and to new customers are ubiquitous. You can refer your friends to a credit card you own (Discover), a mobile network you use (Mint), or to a concert or event you plan to attend (TicketSpice). Even the car company Tesla promises incentives if you successfully refer a friend. There are some problems with referral marketing, however. First, referral campaigns soliciting information from existing customers place the burden on those customers to take action, which increases email overload and annoyance (Grevet et al. 2014) and leads to decreased loyalty (e.g., unsubscribes). Second, the incentive mechanism

for new customers can prove costly for companies if those customers do not return without the incentive. This is a problem, especially for those companies with low-profit margins and those that sell high-ticket items. Third, the conversion rates in referral campaigns are usually low, making them cost-ineffective and non-scalable.

A line of literature investigates how companies can leverage big data and machine learning to generate value from the data they collect. For example, McFowland III et al. (2021) provides a prescriptive analytics framework that uses the company's existing randomized control trial (RCT) data to create individual-level targeted policy interventions to maximize overall profits. Simester et al. (2020) show that machine-learning methods are effective in targeting promotions when prospecting for new customers. To solve the above problems in referral marketing, we first want to investigate how referral targeting compares to algorithmic (machine-learning, matched) targeting in its effectiveness. Second, we want to see if we can use existing machine-learning or econometric methods to turn data stored on previous referral marketing campaigns into assets we call matched targeting to compare the assets' effectiveness to referral targeting. Third, we want to understand the mechanisms behind a referral and why they are valuable. We will also unpack the effects of 'information' and 'influence' that play a role in the purchase decision of a referred customer.

To answer the above questions, we conducted two field experiments with the University of Minnesota Alumni Association (UMAA). The UMAA offers a paid membership program that helps alumni stay connected to the university and gives them access to member-specific events, discounts on athletic events, and university-related merchandise. We designed a referral-marketing program to solicit referrals from its existing members. The cost of this method is the infrastructural challenges of solicitation, tracking of referrals, and also any incentives

that would be given to the existing members. The benefits are that this would unlock a new stream of customer discovery and a predictive cost of customer acquisition. We also built a machine-learning (ML) model using the data on previous membership purchases to target alumni with a high propensity to purchase a membership. The cost of this method is the infrastructure cost of storing the data and the technical challenge of building and maintaining a machine-learning targeting model. The benefit is the ability to generate recurring leads seasonally without the dependence of humans in the loop. We find that referral targeting has a better uptake rate than machine-learning targeting for the same offer. We also designed a matched targeting method using coarsened exact matching (CEM) on existing referral data and targeted people with similar characteristics (information motivation) to that of the referred members. We find that both referral and machine-learning targeting have better uptake rates than matched targeting. This also shows that social influence plays a significant role in a referral mechanism compared to just straight information, though we do not claim any causality.

We make four contributions through the studies in this chapter. First, we assess the new member conversions through two different mechanisms, referral targeting, and algorithmic (or machine-learning) targeting. Referral targeting depends on factors outside the company's control, i.e., dependence on existing members to refer. By contrast, algorithmic targeting depends on the company's ability to both use and turn its data into assets. Second, we create a matched targeting algorithm to turn the referral data into assets and then compare the assets' effectiveness to other targeting mechanisms. Third, we show the causal impact of using machine-learning targeting using a machine-learning control group; more on this in the following sections. Fourth, we attempt to unpack the mechanism behind a successful referral.

The remainder of this article is organized as follows. In section 3.2, we develop theory and pose hypotheses. In section 3.3, we talk about the field setting we use to test our hypothesis and design the experiment. This is followed by sections 3.4 and 3.5, where we present the details of two studies along with analysis, results, and discussion. We conclude in section 3.6, where we summarize the findings, discuss limitations, and end with managerial implications and future research scope.

3.2 Related literature and hypothesis development

Customers use Word of Mouth (WOM) to reduce uncertainty around the quality and fit of a product or service (Chen et al. 2021), so for companies, it is an effective marketing tool (Godes & Mayzlin 2004). The Internet makes it easy for customers to find WOM advice through online blogs and review websites (Chevalier & Mayzlin 2006)—as well as giving them the ability to ask their friends and extended networks using social media —and for companies to disseminate and manage WOM through channels such as referral marketing (Dellarocas 2003). A line of existing literature shows the effectiveness of electronic WOM (eWOM) (Chen & Xie 2008, Duan et al. 2008, Goh et al. 2013, Li & Hitt 2008). Chevalier & Mayzlin (2006) find that the addition of a favorable review on Amazon increases a book's sales, and Duan et al. (2008) find that WOM chatter about a movie has a direct impact on its box office performance either positively or negatively.

While most eWOM is user-generated content that sits on a public website, social referral mechanisms use a person's social ties as a delivery paradigm for recommending a product or service (Brown & Reingen 1987). Unlike other eWOM,

referral programs are company-initiated campaigns that incentivize existing customers to refer their friends and family members to become new customers (Kumar et al. 2010). Companies can use incentives to stimulate referrals and have greater control over messaging about their products, unlike in traditional user-generated eWOM (Berman 2016).

Another line of literature shows the effectiveness of referral marketing (Chen n.d., Gershon et al. 2020, Gains 2017). Referral targeting is composed of two aspects, information and influence (Van den Bulte et al. 2018)). Customers refer friends who they believe will be better off if they use the product or service (Berman 2016). They may use latent data about themselves and others in their social network to identify these potential new customers (information). Then, once they recommend a product, the presence of social proof increases trust in the product for these potential new customers (influence) (Bapna & Umyarov 2015). This is augmented by a call to action (e.g., asking them to purchase or sign up) from a social tie rather than one from the firm itself (Berman 2016); however, if we use a technique like matching to mimic this process and find customers similar to those referred by existing members, we might be able to capture the information aspect sans influence. The efficacy of matched targeting will depend on how effectively the matching algorithm can capture this information, while the effectiveness of algorithmic targeting will depend on the richness of data the firm possesses on potential customers and the learning algorithm used for finding the decision boundaries between those who are likely to enroll in a paid membership and those who are not. In this chapter, we investigate which targeting strategy is more effective for the firm: social targeting or algorithmic targeting. This helps us understand whether existing members truly have better (latent) information on potential new members compared to the firm. In addition, though matching

methods capture the information contained in the referral process, the call to action is still made by the firm, resulting in a lack of direct social influence. We want to understand whether the presence of social influence plays a significant role in the conversion of a referral.

Referral targeting not only captures the information that existing customer provides on their friends who would benefit from a product (Van den Bulte et al. 2018), but the presence of social proof in the targeting emails exerts social influence in the uptake decision (product purchase) that the new potential customer will make (Burtch et al. 2018). On the surface, this stacking of effects would seem to have a better effect on the conversion rate, but the effectiveness of ML targeting also depends on the richness of the company's data (Gupta et al. 2021) and the training algorithm used to learn the uptake propensity from past data (Raschka 2018). ML targeting could capture complex patterns in the feature set that could indicate the propensity of product purchase. For this reason, we believe that the effectiveness of referral targeting would be different from that of machine-learning targeting. Also, because the referral process uniquely captures the information surrounding the need for a product and exerts social influence, we believe it will be more effective than randomly targeting potential members with the same offer. We also believe that the referral process will have a better uptake rate than the baseline propensity of a potential customer to become a new customer in the absence of any offer.

Machine-learning targeting uses past data to understand which combination of features and what values are indicative of a purchase. The main goal of model training is to learn feature combinations to predict purchases better than randomly picking a sample of potential members from the population. Given a rich set of data that is predictive of product purchase and sound learning algorithms, we

believe that ML targeting will outperform random targeting of the same product offers. Similarly, we believe it will outperform a control group that does not receive an offer. While building an ML model, we usually train the data on the previous product purchases. Targeting members who the ML model shows to have high propensity to purchase or take action can still have blind spots, as some of these members would have purchased membership anyway, even in the absence of any offer targeting. We believe that targeting them with an offer nudges them and increases their propensity to purchase compared to the absence of any offer.

In conclusion,

3.3 Institutional details and study design

The University of Minnesota Alumni Association (UMAA) has a mission to connect alumni, support student success, and advance the University of Minnesota (UMAA 2019). UMAA primarily offers two types of paid memberships—lifetime and annual — so alumni can be involved with the university while also taking advantage of benefits such as discounted university game tickets and invitations to exclusive alumni events. Lifetime membership consists of a one-time payment of $750 for a single membership and $950 for joint membership with a spouse. Annual membership costs $50 for a single membership and $60 for joint membership with a spouse each year. During the timeline of the experiment, UMAA had 17,984 registered lifetime members and y registered annual members. UMAA stores rich demographic and behavioral data on alumni, including birth year, gender, marital status, ZIP code, household income, graduation year, graduation degree, event attendance, email click behavior, volunteer activity, and previous membership. This dataset is unique because it contains the entire population of

interest—in this case, the population of all potential new members—since UMAA stores data on all alumni.

3.3.1 Experiment design

We use the terminology below to describe different types of members in UMAA and the experiment for clarity and consistency. We borrow some terminology used in Gershon et al. (2020) from a similar referral-based research context. We use terminology as described in the table 3.1.

Term	**Definition**
Alumni	Those who graduated from the University of Minnesota with any degree
Members	Alumni who are current lifetime or annual members of UMAA
Non-members	Alumni who are not current members of UMAA
Solicitation stage	Process of UMAA sending emails to ask its members to refer potential future members
Referral stage	Process of members referring their friends and giving the friends' information to UMAA
Uptake stage	Process where recipients choose whether to follow through with the referral, take the offer, and become a member or not
Senders	Members who provide referral information
Recipients	Those whose information senders give to UMAA via the referral solicitation portal
Uptakers	Recipients who successfully take up the offer in the experiment and buy a membership

Referral rate	Number of referrals made divided by the number of members asked to make referral
Uptake rate	Number of recipients who follow through and take up the offer divided by the number of referrals made by the members

Table 3.1: Terminology used in the studies

The experiment design consists of two studies: one with lifetime members and one with annual members. We contact only lifetime members as part of the solicitation stage in the lifetime member study and only annual members in the annual member study.

3.4 Lifetime member study

This study seeks to understand the effectiveness of machine-learning (algorithmic) targeting relative to the referral targeting strategy.

We operationalize referral targeting in the solicitation stage. In this study, we solicit referrals only from lifetime members. Even though there are 17,984 lifetime members, only 12,183 have valid email addresses on record without any opt-out preferences. We sent an email to these members calling for referrals and asked them to refer their friends. We promised a University of Minnesota-branded face mask as an incentive (3.1). The link in the email redirects them to a portal where they can enter friends' information (first name, last name, email, college affiliation). The left side of the figure 3.1 shows the portal. We collect this information to match the referral with a record in the alumni database, and we send a membership offer via email. We also sent two follow-up reminder emails to remind them to refer their friends, with one week between each email.

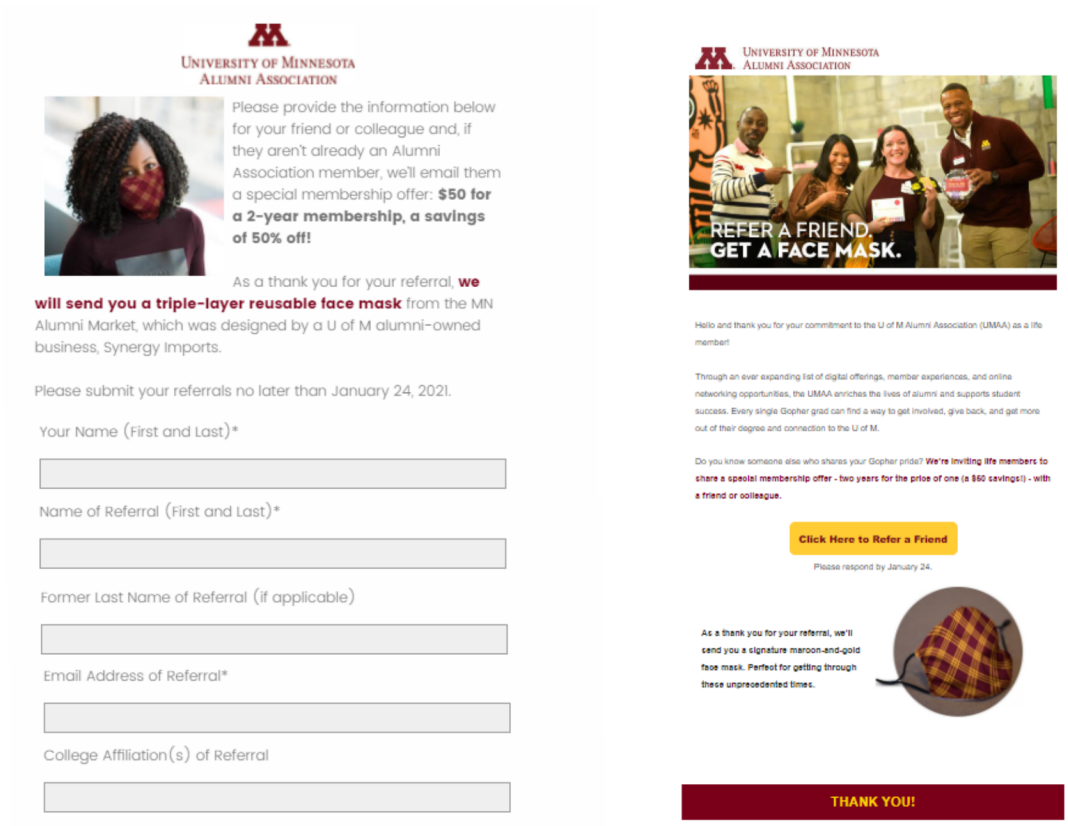

Figure 3.1: The left side shows the portal where members can enter their friends' information. The right side shows the email sent to lifetime members to solicit referrals.

This study is a between-subject design. We describe each group in detail below.

Referral group: This group comprises alumni referred by existing lifetime members of UMAA.

ML targeting group: We use data from historic membership purchases and build a machine-learning (ML) model to predict which alumni are most likely to purchase a membership. We use demographic information, email clicking behavior, and engagement information as input data for the ML model. We use data from 2016 to 2019 as features and to memberships purchased in 2020 as target for building the model. We use XGBoost, a popular supervised machine-learning model, to learn whether an alumnus purchased a membership. More details about the model are in section 3.4.1. We use this model to predict the probability of a non-member purchasing a membership in 2021. We then send a membership offer to the top n number of people with the highest propensity as predicted by the model. The problem here is that some non-members might purchase the membership even without receiving the email offer. We created another group called the ML control group to control for this and help us establish the causal effect of ML-based targeting.

ML control group: We rank order the non-members based on the descending order of propensity to purchase membership based on the ML model. We allocate every even number ranked non-member into the ML control group until we reach the sample size requirement specified in the power analysis.

Random group: The membership offer we send to current referred members and those identified by the machine-learning model might have baseline effectiveness, meaning there is a baseline effect of the offer irrespective of the targeting method. To control this, we create a 'random group' where we send the offer to

randomly selected alumni to understand the baseline effectiveness.

Control group: Some alumni might buy a membership even without any offer presented to them. We create a 'control group' with randomly selected alumni who are not solicited with any offer to understand this baseline.

We randomly allocate any members selected to be part of more than one group to just one of those groups. For example, if 'a' was randomly chosen in the random group but also had a high enough propensity to be in the ML group, we randomly allocate 'a' to either of the groups.

3.4.1 Machine learning modeling

We use historical membership-purchase data to learn the characteristics of alumni who are most likely to purchase a membership. We use demographic, engagement, email, event, and membership-related attributes from 2016 to 2019 that UMAA collected on all alumni as features in the machine-learning model. Table 3.2 lists all the features used; most of them are self-explanatory. We used the binary indicator of whether an alumnus purchased a membership in 2020 as a target variable that the machine-learning model will learn.

We split the entire dataset into training and testing sub-datasets using a 90%-10% split stratified on the target variable to ensure enough minority class data points in the test set. The training dataset has 430,739 data points, while the test dataset has 47,860 data points.

One problem with the training data is that it has an extreme imbalance since the number of data points without a purchase is significantly more than those with a purchase, so the training model will spend most of its time on non-purchase data points and not learn enough from purchase ones. We use SMOTE (Synthetic Minority Oversampling Technique) to upsample the minority class and create

synthetic data points that mimic membership purchase and solve the imbalance problem.

We then use z-scaling to normalize all the numerical features and bring them to a similar scale. We use one-hot encoding to create dummy variables for all categorical features. As the feature set is large, we use PCA (principal component analysis) to treat the curse of dimensionality and pick the components that capture the most variance.

We choose Area under the ROC Curve (AUC) as the metric for evaluating the performance of the ML model, especially as the data are extremely imbalanced and the downstream usage of the ML model is to rank-order all alumni according to the propensity to purchase a membership. We trained multiple algorithms on the data and eventually chose the XGBoost algorithm as it has the best AUC. We further do hyperparameter tuning using grid search and 5-fold cross-validation to tune all the parameters of XGBoost algorithm and converge on the parameters (n_estimators=150, max_depth=2, min_child_weight=5, gamma=0.0, subsample=0.6, colsample_bytree=0.8, seed=27, reg_alpha=0.00002) that give best cross-validated AUC . The AUC on the training data is 0.874 and 0.90 on the test data. We present the confusion matrix for predictions on test data in table 3.3.

Features such as affinity network interest, marital status (married people are more likely to convert), age (older people are more likely to convert), athletic interest, learning events, in_tc_metro_area, and gender (males are more likely to convert) have the highest importance in the ML model.

3.4.2 Power analysis

We conducted an a priori power analysis to determine the sample size needed in the experiment. We set the significance level (Type I error rate) at 0.05 and the

Feature Type	Features
Demographic attributes	marital status, gender, age, in tc metro area, is current tc employee, athletic interest, travel interest, affinity network interest, web topic opt-ins
Engagement attributes	umn event, umn member, umn donor, umn volun, umn inform, umn loyalty, umn avg annual score 5 years
Email related attributes	learning emails, legislature emails, social emails, sports emails, general ctr emails, learning events, legislature events
Event-related attributes	networking events, other events, social events, sports events, total type person events
Membership-related attributes	member before, years before

Table 3.2: Features used in the machine-learning model

	Predicted non-purchase	Predicted purchase
Actual non-purchase	39856	3316
Actual purchase	67	221

Table 3.3: Confusion matrix for ML model predictions on test data

power (1- Type II error rate) at 0.8. As the outcomes (membership purchase) of the experiment are Bernoulli random variables, we need to specify the parameters (P1, P2) of the outcome distributions of both groups to determine the sample size required to detect the difference. We use prior UMAA data from similar membership solicitation campaigns to specify the parameters. We use results from two such campaigns to guide the decision of setting P1 and P2. In the first such campaign, UMAA sent an email offering a set of university-branded mittens in exchange for purchasing an annual membership to 200,000 non-member alumni. Twenty alumni purchased the membership, making the uptake rate 0.01% or P=0.0001. This is a good prior for the uptake rate of the random group. In the second such campaign, UMAA sent an email with the same offer as in this study (see below: two-year annual membership for the price of one) to 2,000 non-member alumni whom UMAA considered to have a high propensity to purchase a membership. The uptake rate in this campaign was 0.4% or P=0.004. This uptake rate is a good prior for the uptake rate of ML and referral groups. Based on the above parameters (alpha = 0.05, power = 0.8, P1=0.004, P2=0.0001), each group must have at least 1,663 members to detect the effect (difference between P1 and P2).

However, the uptake rates of the referral group and ML group would be much closer than that of the ML and the random groups. To detect this difference, the sample size in each group must be much larger than 1,663. We use a bottom-up approach to determine the sample size. The maximum number of emails sent to lifetime members in the solicitation stage could be 12,183 (the number of eligible lifetime members). Assuming the uptake rate of 0.8% in referral group (P=0.008) and setting the other parameters (alpha = 0.05, power = 0.8, p1=0.004), the sample size of ML group should be at least 21,438. These numbers give us a good

idea of the sample size allocations to different groups.

Based on the above power analysis and the constraints of UMAA, we allocated around 20,000 members each to both the machine-learning group and machine-learning control groups and 5,000 members each to the random and control group. More on this in 3.5

Everyone in the experimental groups other than the control group and ML control group is sent an email with a discount membership offer (two-year annual membership for the price of one). The main difference between the email sent to those in the referral group vs. other groups is the presence of social proof in the email text. The email sent to referrals explicitly uses the name of the sender who referred the recipient, as shown on the left side of the figure 3.2. In contrast, emails sent to alumni in other groups mention that they were selected to receive a limited-time membership offer but without the personalization, as shown on the right side of the figure 3.2.

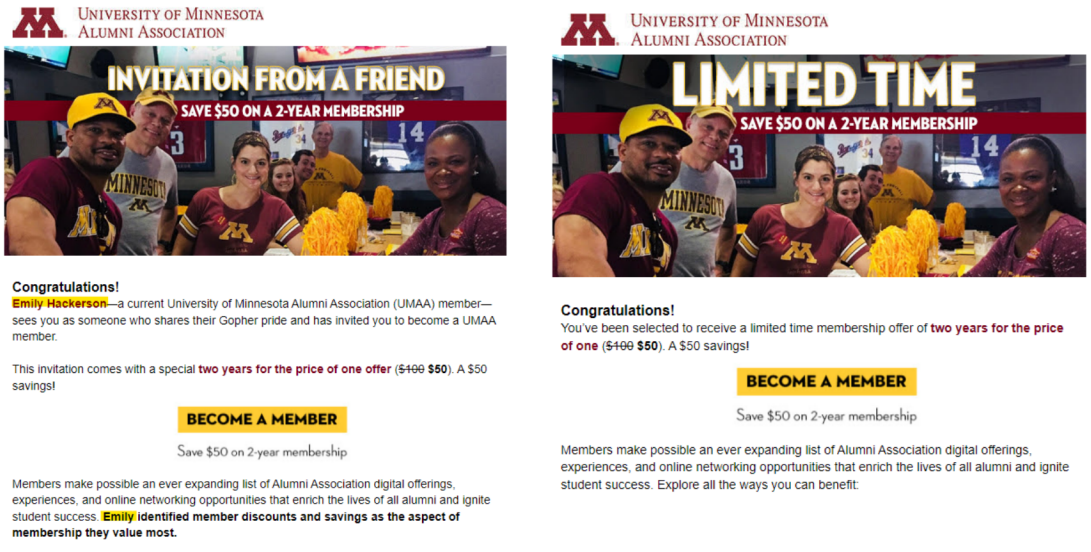

Figure 3.2: An example email sent to recipients during the uptake stage

3.4.3 Empirical strategy and results

We received 101 referrals from lifetime members; however, only 84 were actionable. We matched the details contained in the referrals with existing records and removed referred alumni who were already current members, opt-outs, or who had invalid email addresses. Only 1 out of 84 recipients in the referral group converted and became an annual member in the uptake stage. There were 59, 32, and 3 conversions in the ML group, machine-learning control group, and random group respectively. None of the 4,799 control-group members who did not receive a membership offer became a member. Table 3.5 shows the sample sizes, number of uptakes, uptake proportion, and conversion rates for all groups.

We use a two-sample chi-squared test for equality of proportions to test if the difference between the uptake proportions in the two groups is statistically significant. The qualitative variable in the test is the uptake rate, and the comparison groups could be any of the experimental groups of interest. We use the chi-squared test statistic as defined in the equation 3.1 with one degree of freedom. When the uptakes are too small, and the chi-squared approximations fail, we use Fisher's exact test, which assesses the null hypothesis using hypergeometric distribution (sampling without replacement). Table 3.4 shows an example contingency table used to calculate the test statistic.

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{3.1}$$

Table 3.6 shows the difference in uptake rates (effects) between all groups and the p-values of the effects. The main effect of interest is the difference between the uptake rates of the referral and ML groups. We find that the referral group has a higher conversion rate, although it is not significantly different from the

	Uptake	No Uptake	Total
Referral Group	1	83	84
ML Group	59	19926	19985
Total	60	20009	20069

Table 3.4: Example contingency table for uptake rates of the referral and ML groups

Group	N	No. of uptakes	Uptake Proportion
Referral Group	84	1	0.0119
ML Group	19985	59	0.0029
ML Control Group	20015	32	0.0016
Random Group	4760	3	0.0006
Control Group	4766	0	0

Table 3.5: Experimental group sample sizes and uptake proportions

ML group. The ML group has a significantly better conversion rate than the ML control group signifying the effectiveness of the ML-based targeting model. The ML group also has a significantly higher uptake rate than the random group. The table has no entries for the control group as there were no uptakes in that group.

3.4.4 Discussion

Based on the outcomes of the lifetime study, we have four main discussion points.

Under-powered referral group

The referral group has a better uptake rate than the ML group; however, the null hypothesis cannot be rejected as the difference is not statistically significant. We believe this is because the study is underpowered, and the number of referrals from lifetime members in the referral stage is insufficient. As only one person

	Referral Group	ML Group	ML Control Group	Random Group
Referral Group	-	0.009 (0.610)	0.010 (0.328)	0.011** (0.049)
ML Group	-	-	0.0013*** (0.003)	0.0023*** (0.003)
ML Control Group	-	-	-	0.001* (0.080)
Random Group	-	-	-	-

Table 3.6: The effect sizes and p-values of the difference between various groups in the study

Probability range	% of members in the range	% of conversions in the range
0.9-1.0	5%	15%
0.8-0.9	13%	32%
0.7-0.8	25%	23%
0.6-0.7	45%	27%
0.58-0.6	12%	3%

Table 3.7: Predicted probability ranges of alumni in the lifetime study and the distributions of conversions

purchased the membership in the referral group, it poses a challenge to infer the difference between the machine-learning and referral groups.

Causal impact of ML model for targeting

The machine-learning group outperformed both the machine-learning control group and the random group, showing that the machine-learning (algorithmic) model effectively learned the patterns in the data and the targeting method works. This also shows the effectiveness of the email nudge and the offer inside. Alumni allocated to both the machine learning and machine-learning control groups were predicted to have a high propensity for membership purchase by the ML model; however, the outperformance of the machine-learning group causally shows the effectiveness of machine-learning targeting. It removes any doubt that the members in the ML group would have converted even in the absence of any targeted offer.

Effectiveness of targeting using existing data

Both the referral group and ML group outperformed the random group. This shows the value of using data that the company has to target specific alumni for membership solicitations rather than randomly targeting a subset of the population.

Limitations

Although the lifetime member study has strong internal validity, one limitation is that the uptake rate of the ML group depends on the threshold cutoff used to determine the eligibility of alumni selected for the group. ML group assignments occur based on the rank order of the probability of uptake for all alumni. The

sample size estimated from the power calculations decides the probability threshold (0.58 in this study) above which a person is allocated to the ML group. This threshold will determine the effectiveness of the ML group. If this cutoff were 0.9, the uptake rate would be much higher. As shown in the table 3.7, the alumni who are predicted to have higher uptake rates also actually have higher uptake rates in practice. This shows that the ML model works; it also shows that the probability cutoff will determine the overall uptake rate of the ML group and its difference from the referral group.

Even though we find that the referral group had a better uptake rate than the ML group—although not statistically significant—we believe results may vary depending on the company's context. The effectiveness of the ML model will depend on the company's specific data and its relevance in predicting if a customer has a propensity to purchase a product. Similarly, the effectiveness of the referral mechanism will depend on the incentive used in the solicitation stage and the customers' love for the product.

3.4.5 Ideal design

In order to test the hypothesis at hand, we would need the baseline (absence of any nudge) membership conversion rate for all the groups - referral, machine learning, and random. All these groups do not get any email offers. This will give us the conversion rate of those who would buy membership anyway. Second, we need the conversion rate when UMMA sends the email offers for the random group, the ML group, and the referral group. This will help us understand the effect of the nudge on membership purchases. Third, we need a conversion rate for the social solicitation of membership. In this case, the email would be sent from existing members to potential new members.

In conclusion referral group will be randomly split into three subgroups, the ML group will be split into two subgroups and the random group will be split into two subgroups.

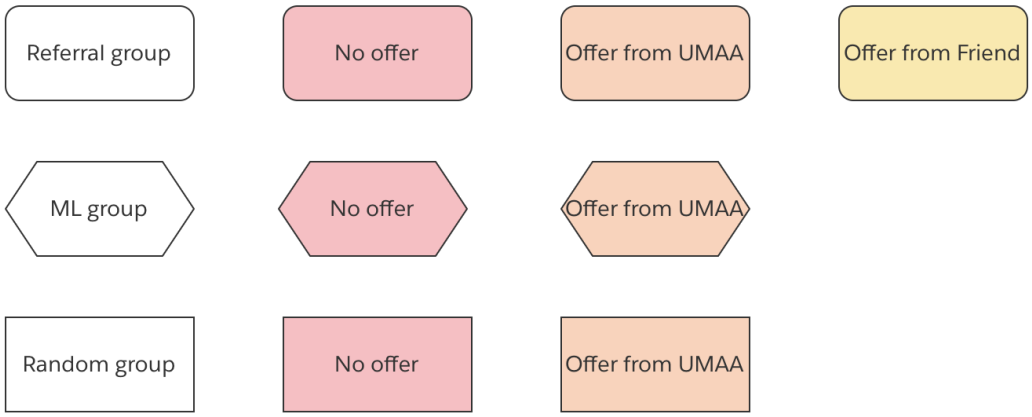

Figure 3.3: An example email sent to recipients during the uptake stage

3.5 Annual member study

The lifetime member study falls short in some respects. First, the uptake rate in the referral group is not high. This could be because the study was underpowered and the number of lifetime members available to refer their friends might be less, or their willingness to refer was low. We wanted to see if we could resolve this by soliciting referrals from annual members. Second, we could not test the mechanism behind uptake decisions in the referral process. We posit that the combination of information and influence drives the purchase decision, but the low number of referrals from lifetime members prevented us from testing it. Suppose there were a large number of referrals. In that case, we could conduct a study where 50% of the referrals were sent an email without any social proof and 50% with

social proof to untangle the effect of influence on uptake decision. To overcome this, we created another group called the 'matched group.' We use coarsened exact matching (CEM) to find alumni with the same characteristics as the referrals in the lifetime study. Even though the members in the matched group were not actual referrals, we believe this will help us better understand the mechanism behind the referral uptake. Third, the machine-learning targeting performed significantly better than the random and control groups in the lifetime study. We were able to causally attribute the effectiveness of the ML-based targeting by comparing it with the machine-learning control group. We want to see if we can replicate the ML model performance in this study and robustly claim the causal impact of ML-based targeting.

For the annual study, we solicit referrals from annual members instead of lifetime members during the solicitation stage. The process of solicitation, referral, and uptake for the referral group is similar to that of the lifetime member study. We target alumni with a high propensity of conversion predicted using the same ML model used in the lifetime member study for the ML group. The ML control group, random group, and control group follow the same allocation rules as the lifetime member study. This study also has a matched group. The matched group contains alumni who have similar characteristics to those referred by lifetime members in the lifetime member study. We use coarsened exact matching (CEM) and match on all features used in creating the ML model described in table 3.2. We obtained 9,818 matches from the existing alumni database with similar characteristics to those referred by lifetime members. Similar to lifetime members, we sent an email (figure 3.4) to annual members, but in this study, we gave them a choice of rewards to encourage more referrals instead of just a blanket reward like we used in the lifetime member study.

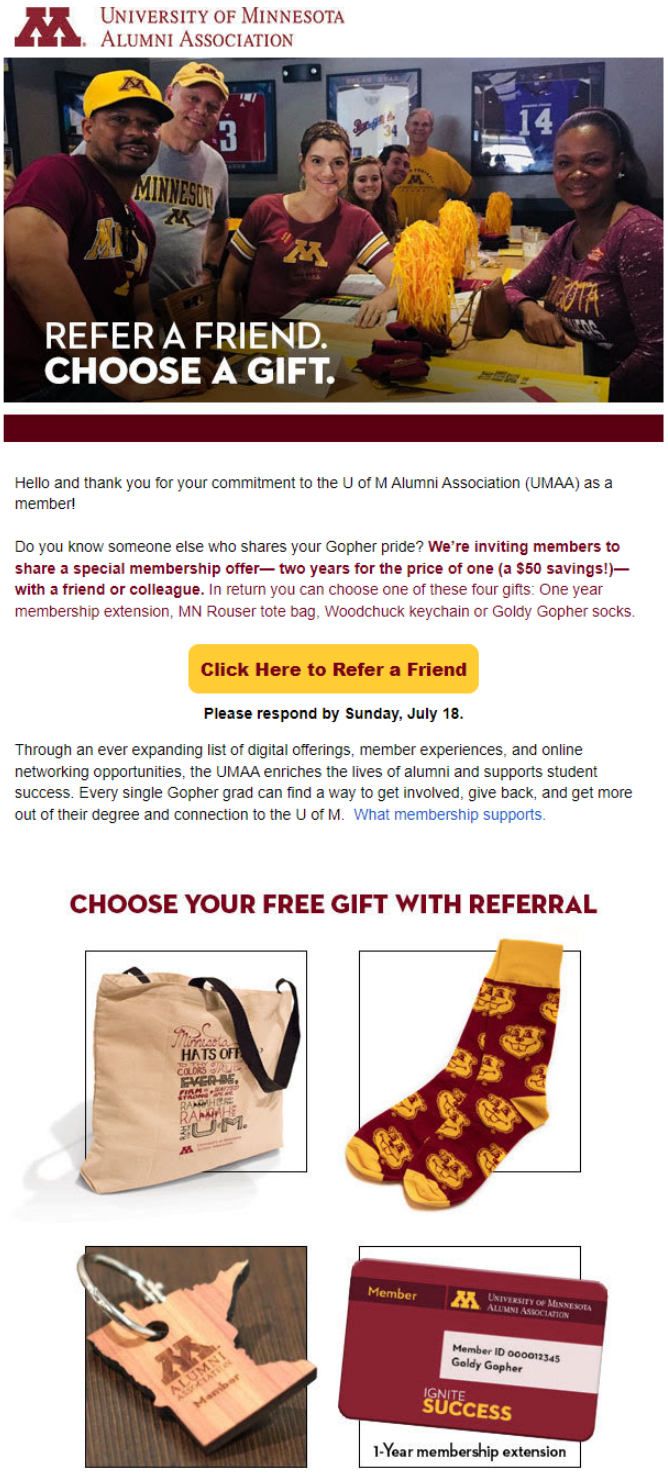

Figure 3.4: Annual members referral solicitation email

3.5.1 Coarsened exact matching

We use coarsened exact matching (Iacus et al. 2012) to find alumni with similar features to those referred (recipients) by lifetime members from the lifetime member study. In coarsened exact matching, every feature of interest is binned into intervals. Each of the 100 referrals from the lifetime study is given a bin signature based on the values of the features. Alumni with the same bin signatures are matched with the referrals. For example, if a referral is female (F), with high-income status (Hi) and college (C) education, the bin signature would be F_Hi_C. This person is matched with a person with the same signature in the alumni database.

3.5.2 Results

Table 3.8 shows the sample sizes, number of uptakes, and uptake proportion of all groups in the experiment. We received 15 referrals from annual members, and only 1 of those uptook the membership offer and became a member. There were 63 uptakes in the ML groups showing the robustness of the ML model used for targeting; this number is consistent with the lifetime member study. There were seven uptakes in the ML control group; this number is less than that of both the ML group in this study and the ML control group in the lifetime study. There were just two uptakes from the matched group, and there was only one uptake in the control group. Surprisingly, there were no uptakes in the random group.

We again use Fisher's exact test to test the difference between groups as the observed counts (number of uptakes) are small. This test will assume the underlying distribution to be hypergeometric, providing accurate inference on the difference between proportions of groups. Table 3.9 shows the effect sizes between

Group	N	No. of uptakes	Uptake Proportion
Referral Group	15	1	0.0667
ML Group	19795	63	0.0032
ML Control Group	19792	7	0.0003
Matched Group	9818	2	0.0002
Random Group	4796	0	0
Control Group	4796	1	0.0002

Table 3.8: Experimental group sample sizes and uptake proportions for annual member study

	Referral Group	ML Group	ML Control Group	Matched Group	Control Group
Referral Group	-	0.063** (0.05)	0.067*** (0.006)	0.067*** (0.004)	0.067*** (0.007)
ML Group	-	-	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)
ML Control Group	-	-	-	0.0001 (0.72)	0.0001 (1)
Matched Group	-	-	-	-	0 (1)
Control Group	-	-	-	-	-

Table 3.9: The effect sizes and p-values of the difference between various groups in the annual member study

different groups and their corresponding p-values in brackets. The main effect of interest is the difference between the uptake rate of referral and ML groups. We find that the referral group has a better uptake rate, and the difference is significant; however, it is hard to robustly claim the effect because there is just one uptake in the referral group. The uptake rate in the ML group is significantly higher than in the ML control, matched, control, and random groups. This finding is similar to the one in the lifetime study.

The referral group and ML group significantly outperformed matched group. There is no significant difference between the matched group and the ML control and control groups.

Probability Range	% of members in the range	% of conversions in the range
0.9-1.0	1%	27%
0.8-0.9	24%	24%
0.7-0.8	35%	35%
0.62-0.7	35%	14%

Table 3.10: Predicted probability ranges of alumni and the distributions of conversions in the annual member study

3.5.3 Discussion

Based on the above results, we have three main discussion points.

Performance of the matched group

The uptake rate of the matched group (whose email text does not have social proof) is significantly lower than that of the referral group (whose email text does have social proof), meaning influence plays a crucial role in the uptake decision, not information. The limitation here is that the alumni in the matched group were not actual referrals but had similar characteristics to that of the referrals per coarsened exact matching. The uptake rate could be different in an ideal situation where actual referred members were sent an email without social proof rather than matched members.

Another reason could be that, unlike the ML group, the coarsened exact matching learns referral information, not uptake information. This difference gives a baseline advantage to the ML group. Also, the learning algorithms used in the ML group are more advanced and can capture more nuance than coarsened exact matching, which is similar to the KNN algorithm.

Causal performance of ML group

The performance of the annual study's ML group is consistent with that of the lifetime study. Again, the ML group performs better than both the ML control group and the random group, proving the effectiveness of the data used in training and the learning algorithm used to build the ML model. This helps us establish the causal effect of ML targeting on the uptake decision and gives confidence about the robustness of the model with its consistent performance in both the lifetime and annual members experiments.

Referral group performs better than ML group

Even though the number of referrals is still less in this study, the uptake rate is statistically significantly higher than in the ML group. This shows that referral targeting is an effective way to acquire new customers even though the scale is not large.

3.6 Summary and concluding remarks

In this chapter, we contribute to the idea of digital transformation, specifically to the aspect of how companies can use advanced methods to turn data into assets. We propose two ways to target new customers using a company's existing data. First, we use its referral data to create a new targeting mechanism. Second, we use its existing data on customer purchases to create a machine-learning targeting algorithm that can be used to target customers to purchase products. We designed a new referral campaign asking a company's existing customers to refer their friends. We then discuss the relative effectiveness of referral targeting to that of data-based approaches—matched and machine-learning targeting—and

weigh the trade-offs. We conducted two field experiments with the University of Minnesota Alumni Association to target alumni and send them offers to purchase memberships. For the referral mechanism in each study, we ask first lifetime members and then annual members to refer their friends. In both studies, we build machine-learning models to target existing alumni. We find that referral targeting performs marginally better than ML targeting, although the overall number of referrals is meager. We also see that ML targeting causally performs better than no offer or random targeting, clearly showing the value of using existing data to target people and turn the data into an asset. We also find that both referral targeting and ML targeting outperform matched targeting. Additionally, we wanted to investigate the mechanism behind referral conversion: information or influence. As the performance of the matched-targeting method, where information is the primary driver of conversions, is worse than that of referral targeting, we believe influence plays a huge role in the uptake decision of a referral.

There are some limitations to these studies. First, the studies might lack external validity. The effectiveness of ML targeting will depend on the data that the company stores on the customers and its technical ability to translate said data into a functional machine learning model. Also, the cutoff of the ML group will determine its effectiveness, making the comparison with other groups hard to generalize. Second, the effectiveness of the referral group will depend on the company's incentive to get referrals, the customers' love of the product, and their willingness to refer their friends.

The conclusion about the mechanism behind a referral (information vs. influence) is not clean because we used a matched group instead of actual referrals because of a lack of referrals from existing members. The mechanism used may play a more significant role in a perfect scenario than what we could deduce from

the study.

Companies can use these findings to weigh the tradeoffs between referral marketing and ML (algorithmic) targeting. Referral marketing is a costly affair. Apart from product offers to new customers, existing customers must receive an incentive in order to make referrals; however, the results suggest that the uptake rate for referral targeting would be higher in actual practice. On the other hand, machine-learning targeting could be more cost-effective since it does not require incentivizing any existing customers, but its effectiveness will depend on the quality of data that the company collects and the costs it incurs to acquire, store, and hire the experts needed to build and maintain those models. One item of note is that turning the company's data into assets using ML targeting will consistently outperform random targeting.

We also believe that companies should collect rich data whenever possible because there is enough evidence that digital transformation is fundamental to the growth of a company, and data lie at the center of what is possible for firms. With new advancements in methods research, stored data could be used to gain that competitive edge.

It would be interesting to test the above hypothesis in another setting to see if similar effects hold, and if they do not, why? Would it be a lack of sufficient data to build a machine-learning model, or is it the weakness of referrals? This will help us understand the generalizability of the results.

In the future, we want to cleanly identify the effects of information and influence, i.e., the mechanisms behind a referral. With the small number of referrals, splitting them into two groups with one group having no social proof was infeasible. Also, most companies are wary of sending a referral email without social proof. They do not want to send unexpected emails to potential new customers

without saying that their friends referred them. We would want to power the studies even better (by having a larger sample size) to see the clear differences between referral targeting and ML targeting. This would be possible in a more significant setting with a high number of referrals. In the future, we want to dig deeper into referral marketing design and investigate how we can drive the uptake rate of a referral even further.

Chapter 4

Conclusion

This thesis contributes to the topic of digital transformation, specifically how data can be turned into assets. We look at two different contexts, randomized control trials and referral marketing. We choose RCT's because the culture of using experimentation for data-driven decision making is becoming more common in companies, and there is research that shows that start-ups that adopt AB testing to deploy features have increased performance on several critical dimensions, including page views and new product features (**??**). We choose referral marketing because it provides marketers an entirely new way of acquiring new customers. Research shows that referral-based programs that leverage existing customers' social networks to reach potential new customers are effective and customers acquired through referrals are more valuable than other channels (Schmitt et al. 2011).

In chapter 2, we define a prescriptive analytics framework that addresses the needs of a constrained decision-maker facing, ex-ante, unknown costs and benefits of multiple policy levers. The framework is general in nature and can be deployed in any utility-maximizing context, public or private. It relies on randomized field

experiments for causal inference, machine learning for estimating heterogeneous treatment effects, and the optimization of an integer linear program for converting predictions into decisions. The net result is the discovery of individual-level targeting of policy interventions to maximize overall utility under a budget constraint. The framework is set in the context of the four pillars of analytics and is especially valuable for companies that already have an existing practice of running A/B tests. The key contribution of this work is to develop and operationalize a framework to exploit both within and between-treatment arm heterogeneity in the utility response function to derive benefits from future (optimized) prescriptions. We demonstrate the value of this framework compared to benchmark practices– i.e., the use of the average treatment effect and uplift modeling–in two different settings. Unlike these standard approaches, our framework can recognize, adapt to, and exploit the (potential) presence of different subpopulations that experience varying costs and benefits within a treatment arm while also exhibiting differential costs and benefits across treatment arms. As a result, we find a targeting strategy that produces an order of magnitude improvement in expected total utility for the case where significant within- and between-treatment arm heterogeneity exists.

In chapter 3, we solve some of the challenges associated with referral marketing like email over load and annoyance, costly incentives and low conversion rates by turning the referral data that the company has into an asset. We investigate how referral targeting compares to algorithmic (machine-learning, matched) targeting in its effectiveness. We see if we can use existing machine-learning or econometric methods to turn data stored on previous referral marketing campaigns into assets we call matched targeting and compare the assets' effectiveness to referral targeting. We understand the mechanisms behind a referral and why they are valuable. We also unpack the effects of 'information' and 'influence' that play

a role in the purchase decision of a referred customer. In two field experiments we conduct with University of Minnesota Alumni Association by designing a referral campaign to ask its existing members to refer their friends to the Annual membership program. We find that referral targeting has a better uptake rate than machine-learning targeting for the same offer. We also designed a matched targeting method using coarsened exact matching (CEM) on existing referral data and targeted people with similar characteristics (information motivation) to that of the referred members. We find that both referral and machine-learning targeting have better uptake rates than matched targeting. This also shows that social influence plays a significant role in a referral mechanism compared to just straight information, though we do not claim any causality.

Both these chapters show that critical managerial challenges can be solved using algorithmic development (machine learning and econometrics) and data that company already stores can be turned into an asset. In future, we want to investigate more contexts of information systems and marketing to make impact.

References

Agarwal, R. & Dhar, V. (2014), 'Big data, data science, and analytics: The opportunity and challenge for is research', *Information Systems Research* **25**(3), 443–448.

Agrawal, S. & Goyal, N. (2012), Analysis of thompson sampling for the multi-armed bandit problem, *in* 'Conference on learning theory', pp. 39–1.

Aral, S. & Walker, D. (2011), 'Creating social contagion through viral product design: A randomized trial of peer influence in networks', *Management Science* **57**(9), 1623–1639.

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.

Auer, P. (2002), 'Using confidence bounds for exploitation-exploration trade-offs', *Journal of Machine Learning Research* **3**(Nov), 397–422.

Bapna, R. & Umyarov, A. (2015), 'Do your online friends make you pay? a randomized field experiment on peer influence in online social networks', *Management Science* **61**(8), 1902–1920.

Berman, B. (2016), 'Referral marketing: Harnessing the power of your customers', *Business Horizons* **59**(1), 19–28.

Berry, G., Franco, A., Peysakovich, A. & Taylor, S. (2016), 'Two stage: A simple framework for finding cates', *Conference on Digital Experimentation (CODE)* .

Bertsimas, D. & Kallus, N. (2014), 'From predictive to prescriptive analytics', *arXiv preprint arXiv:1402.5481* .

Biesecker, L. G. (2013), 'Hypothesis-generating research and predictive medicine', *Genome Research* **23**(7), 1051–1053.

Brown, J. J. & Reingen, P. H. (1987), 'Social ties and word-of-mouth referral behavior', *Journal of Consumer research* **14**(3), 350–362.

Bughin, J., Deakin, J. & O'Beirne, B. (2020), 'Digital transformation: Improving the odds of success'.
URL: *https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-transformation-improving-the-odds-of-success*

Burtch, G., Hong, Y., Bapna, R. & Griskevicius, V. (2018), 'Stimulating online reviews by combining financial incentives and social norms', *Management Science* **64**(5), 2065–2082.

Chen, P., Hitt, L. M., Hong, Y. & Wu, S. (2021), 'Measuring product type and purchase uncertainty with online product ratings: A theoretical model and empirical application', *Information Systems Research* **32**(4), 1470–1489.

Chen, Y. (n.d.), 'Enhancing effectiveness of referral programs by promoting better matching: Evidence from field experiments'.

Chen, Y. & Xie, J. (2008), 'Online consumer review: Word-of-mouth as a new element of marketing communication mix', *Management science* **54**(3), 477–491.

Chevalier, J. A. & Mayzlin, D. (2006), 'The effect of word of mouth on sales: Online book reviews', *Journal of marketing research* **43**(3), 345–354.

Chu, W., Li, L., Reyzin, L. & Schapire, R. (2011), Contextual bandits with linear payoff functions, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', pp. 208–214.

Dellarocas, C. (2003), 'The digitization of word of mouth: Promise and challenges of online feedback mechanisms', *Management science* **49**(10), 1407–1424.

Ding, W., Qin, T., Zhang, X.-D. & Liu, T.-Y. (2013), Multi-armed bandit with budget constraint and variable costs, *in* 'Twenty-Seventh AAAI Conference on Artificial Intelligence'.

Domingos, P. (1997), Knowledge acquisition form examples vis multiple models, *in* 'Proceedings of the Fourteenth International Conference on Machine Learning', ICML '97, pp. 98–106.

Dopazo, J. & Aloy, P. (2006), 'Discovery and hypothesis generation through bioinformatics', *Genome biology* **7**(2), 307.

Duan, W., Gu, B. & Whinston, A. B. (2008), 'The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry', *Journal of retailing* **84**(2), 233–242.

Economist (2010), 'The data deluge'. `https://www.economist.com/leaders/2010/02/25/the-data-deluge`.

Gains, B. (2017), 'The 51 best referral program examples of 2016', *Saasquatch (February 17), https://www. referralsaasquatch. com/51-referral-program-examples* .

Gershon, R., Cryder, C. & John, L. K. (2020), 'Why prosocial referral incentives work: the interplay of reputational benefits and action costs', *Journal of Marketing Research* **57**(1), 156–172.

Ghose, A., Singh, P. V. & Todri, V. (2017), Got annoyed? examining the advertising effectiveness and annoyance dynamics, *in* 'Proceedings of the International Conference on Information Systems - Transforming Society with Digital Innovation, ICIS 2017, Seoul, South Korea, December 10-13, 2017'.

Godes, D. & Mayzlin, D. (2004), 'Using online conversations to study word-of-mouth communication', *Marketing science* **23**(4), 545–560.

Goh, K.-Y., Heng, C.-S. & Lin, Z. (2013), 'Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content', *Information systems research* **24**(1), 88–107.

Grevet, C., Choi, D., Kumar, D. & Gilbert, E. (2014), Overload is overloaded: Email in the age of gmail, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', CHI '14, Association for Computing Machinery, New York, NY, USA, p. 793–802.
URL: *https://doi.org/10.1145/2556288.2557013*

Grimmer, J., Messing, S. & Westwood, S. J. (2017), 'Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods', *Political Analysis* **Conditional Acceptance**.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R. et al. (2021), Data quality for machine learning tasks, *in* 'Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining', pp. 4040–4041.

Hitsch, G. J. & Misra, S. (2018), 'Heterogeneous treatment effects and optimal targeting policy evaluation', *Available at SSRN 3111957* .

Hosanagar, K. & Saxena, A. (2017), 'The democratization of machine learning: What it means for tech innovation', *Knowledge@Wharton* .

Iacus, S. M., King, G. & Porro, G. (2012), 'Causal inference without balance checking: Coarsened exact matching', *Political analysis* **20**(1), 1–24.

Imai, K. & Ratkovic, M. (2013), 'Estimating treatment effect heterogeneity in randomized program evaluation', *The Annals of Applied Statistics* **7**(1), 443–470.

Imai, K. & Strauss, A. (2011), 'Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign', *Political Analysis* **19**(1), 1–19.

Joulani, P., Gyorgy, A. & Szepesvári, C. (2013), Online learning under delayed feedback, *in* 'International Conference on Machine Learning', pp. 1453–1461.

Jung, J., Bapna, R., Golden, J. M. & Sun, T. (2020), 'Words matter! toward a prosocial call-to-action for online referral: Evidence from two field experiments', *Information Systems Research* **31**(1), 16–36.

Jung, J., Bapna, R., Golden, J. & Sun, T. (2019), 'Words matter! towards pro-social call-to-action for online referral: Evidence from two field experiments', *Information Systems Research, forthcoming* .

Kohavi, R. & Thomke, S. (2017), 'The surprising power of online experiments', *Harvard Business Review* **95**(5), 74–+.

Koning, R., Hasan, S. & Chatterji, A. (2022), 'Experimentation and start-up performance: Evidence from a/b testing', *Management Science* .

Kumar, V., Petersen, J. A. & Leone, R. P. (2010), 'Driving profitability by encouraging customer referrals: Who, when, and how', *Journal of Marketing* **74**(5), 1–17.

Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the National Academy of Sciences* **116**(10), 4156–4165.

Langford, J. & Zhang, T. (2007), The epoch-greedy algorithm for contextual multi-armed bandits, *in* 'Proceedings of the 20th International Conference on Neural Information Processing Systems', Citeseer, pp. 817–824.

Langford, J. & Zhang, T. (2008), The epoch-greedy algorithm for multi-armed bandits with side information, *in* 'Advances in neural information processing systems', pp. 817–824.

Li, L., Chu, W., Langford, J. & Schapire, R. E. (2010), A contextual-bandit approach to personalized news article recommendation, *in* 'Proceedings of the 19th international conference on World wide web', ACM, pp. 661–670.

Li, X. & Hitt, L. M. (2008), 'Self-selection and information role of online product reviews', *Information Systems Research* **19**(4), 456–474.

Martelo, S. & Toth, P. (1990), 'Knapsack problems', *Wiley* **1995**, 306.

McFowland III, E., Gangarapu, S., Bapna, R. & Sun, T. (2021), 'A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects.', *MIS Quarterly* **45**(4).

McFowland III, E., Somanchi, S. & Neill, D. B. (2018), 'Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection', *Working Paper* .

McFowland III, E., Speakman, S. D. & Neill, D. B. (2013), 'Fast generalized subset scan for anomalous pattern detection', *The Journal of Machine Learning Research* **14**(1), 1533–1561.

Misra, K., Schwartz, E. M. & Abernethy, J. (2018), 'Dynamic online pricing with incomplete information using multi-armed bandit experiments', *SSRN Working Paper* .

Neill, D. B. (2012), 'Fast subset scan for spatial pattern detection', *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* **74**(2), 337–360.

Neill, D. B., McFowland III, E. & Zheng, H. (2013), 'Fast subset scan for multivariate event detection', *Statistics in medicine* **32**(13), 2185–2208.

Raschka, S. (2018), 'Model evaluation, model selection, and algorithm selection in machine learning', *arXiv preprint arXiv:1811.12808* .

Rogers, D. (2016), *The digital transformation playbook*, Columbia University Press.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika* **70**(1), 41.

Rzepakowski, P. & Jaroszewicz, S. (2010), Decision trees for uplift modeling, *in* 'Data Mining (ICDM), 2010 IEEE 10th International Conference on', IEEE, pp. 441–450.

Schmitt, P., Skiera, B. & Van den Bulte, C. (2011), 'Referral programs and customer value', *Journal of marketing* **75**(1), 46–59.

Schwartz, E. M., Bradlow, E. T. & Fader, P. S. (2017), 'Customer acquisition via display advertising using multi-armed bandit experiments', *Marketing Science* **36**(4), 500–522.

Simester, D., Timoshenko, A. & Zoumpoulis, S. I. (2020), 'Targeting prospective customers: Robustness of machine-learning methods to typical data challenges', *Management Science* **66**(6), 2495–2522.

Somanchi, S., McFowland III, E. & Neill, D. B. (2018), 'Discovering heterogeneous patterns of care using observational data: Evidence from studies of healthcare', *Working Paper* .

Speakman, S. D., McFowland III, E. & Neill, D. B. (2015), 'Scalable Detection of Anomalous Patterns With Connectivity Constraints', *Journal of Computational and Graphical Statistics* **24**(4), 1014–1033.

Sun, T., Gao, G. & Jin, G. Z. (2019), 'Mobile messaging for offline group formation in prosocial activities: A large field experiment', *Management Science* **65**(6), 2717–2736.

UMAA (2019), 'About university of minnesota alumni association'. `https://www.umnalumni.org/about`.

Van den Bulte, C., Bayer, E., Skiera, B. & Schmitt, P. (2018), 'How customer referral programs turn social capital into economic capital', *Journal of Marketing Research* **55**(1), 132–146.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B. & Brueckner, M. (2018), 'Linear bandits with stochastic delayed feedback', *arXiv preprint arXiv:1807.02089* .

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**(523), 1228–1242.

Zhou, Z., Xu, R. & Blanchet, J. (2019), Learning in generalized linear contextual bandits with stochastic delays, *in* 'Advances in Neural Information Processing Systems', pp. 5197–5208.