Targeting Prospective Customers: Robustness of Machine Learning Methods to Typical Data Challenges

Duncan Simester*, Artem Timoshenko*, and Spyros I. Zoumpoulis†

*Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology

†Decision Sciences, INSEAD

March 2018

Abstract

We investigate how firms can use the results of field experiments to optimize the targeting of promotions when prospecting for new customers. We evaluate seven widely used machine learning methods using a series of two large-scale field experiments. The first field experiment generates a common pool of training data for each of the seven methods. We then validate the seven optimized policies provided by each method together with uniform benchmark policies in a second field experiment. The findings not only compare the performance of the targeting methods, but also demonstrate how well the methods address common data challenges.

Our results reveal that when the training data is ideal, model-driven methods perform better than distance-driven methods and classification methods. However, they also deteriorate faster in the presence of challenges that affect the quality of the training data, including the extent to which the training data captures details of the implementation setting. The challenges we study are covariate shift, concept shift, information loss through aggregation, and imbalanced data. Intuitively, the model-driven methods make better use of the information available in the training data, but the performance of these methods is more sensitive to deterioration in the quality of this information.

The classification methods we tested performed relatively poorly. We explain the poor performance of the classification methods in our setting and describe how the performance of these methods could be improved.

1. Introduction

When prospecting for new customers, firms must decide both which customers to target and what promotions to offer them. A standard approach is to run a pilot experiment to measure the response to different promotions. The firm can then use this training data to design a targeting policy that identifies which promotion to send to each prospective customer, and then implement the resulting policy on a larger scale. The performance of targeting policies depends upon the choice of the optimization method and the quality of the training data, including the extent to which the training data captures details of the implementation setting. Several practical challenges can arise. We test the robustness of different targeting methods by evaluating their performance when confronted with data challenges that are typical of a prospecting setting.

Our performance comparison uses a sequence of two field experiments. The optimized policies are trained using data from the first experiment and validated in the second experiment. This requires that we finalize which methods we will compare before implementing the second field experiment. The criterion we used to select methods was whether a moderately sophisticated retailer would be able to implement the method.

While many firms have built extensive data warehousing capabilities in recent years, the sophistication of the targeting methods they use varies greatly. Firms such as Amazon and Google have very sophisticated machine learning capabilities. Other firms outsource their targeting decisions to third-party consultants or data analysis services. However, many firms rely upon relatively simple targeting models. For example, the retailer that participated in this study has USD revenue in the tens of billions. It has a well-organized and extensive data warehouse, together with a team of data scientists. Before this study, the firm's data science team built its targeting models using OLS. The team did not have experience implementing the machine learning methods that we evaluate in this paper.

We compare seven machine learning methods, including two model-driven methods (Lasso and finite mixture models), three distance-driven methods (k-nearest neighbors, kernel regression and hierarchical clustering), and two classification methods (SVM and CHAID). Overall, the model-driven methods perform best. This is particularly true in regions of the parameter space in which the training data provides an accurate representation of the validation data. However, when the quality of the information in the training data deteriorates, the deterioration affects the model-driven methods more than the other methods.

We evaluate four types of data challenges that affect the quality of the information and the performance of the methods. The machine learning literature labels the first challenge "covariate shift". Most methods assume that the distribution of the targeting variables in the training data will be representative of the implementation data, but this is not always the case. For example, if the targeting method is implemented in different geographic regions than the regions in which the training data was gathered, the characteristics of the targeting pool may differ from the implementation pool.

The second data challenge is commonly labeled "concept shift". If the underlying response function is different in the training data than in the implementation setting in ways that are not captured by the predictor variables, this will generally result in sub-optimal targeting policies. The risk of this is high if the targeting policy is implemented several months after the training data is collected. In the intervening period, changes in the environment through shifts in macroeconomic conditions, competitors' actions, the firm’s other marketing activities, or just seasonality can all contribute to changes in how customers respond to the firm’s actions.

The remaining two data challenges are particularly relevant when prospecting for new customers. Firms generally have much less information about prospects than about existing customers. For existing customers they can use past purchasing decisions, but purchasing histories generally do not exist for prospective customers.[1] Instead, firms are generally forced to rely upon demographic measures aggregated at the zip code, census block or carrier route level. We investigate how this aggregation affects the performance of the different methods.

When prospecting for new customers, firms typically trade-off a low average response rate with a large long-run expected profit from the new customers that do respond. This leads to what the machine learning literature refers to as “imbalanced” data with an asymmetric cost of errors. The imbalance in the data results from the very low response rate, while the asymmetric cost reflects a much higher cost of false negatives (not mailing to customers who would respond) than false positives (mailing to customers who will not respond). Additionally, the low response rate introduces an imprecise labels problem, as the policy that performed best at the finite sample experiment is not necessarily truly optimal. The problem of imprecise labels reinforces the imbalanced data and asymmetric cost problems. As we will discuss, these problems are of particular relevance to the two classification methods (CHAID and SVM).

Our results reveal an important general finding. Model-driven methods perform better than distance-driven and classification methods when the data is ideal. However, they also deteriorate faster in the presence of covariate shift, concept shift and information loss through aggregation. Intuitively, the model-driven methods make the best use of the available information. However, when the quality of the information deteriorates, the performance of these methods is more sensitive to this deterioration. These findings contribute to the debate in the machine learning literature about whether complex methods are less robust than simpler methods. This debate has focused on the impact of covariate shift and concept shift. Our findings indicate that complex methods are more sensitive to these data challenges, as well as to loss of precision through aggregation.

These data challenges are extensively studied in the machine learning literature. Common applications include natural language processing, image analysis, computer vision, robot control, software engineering, bioinformatics and brain-computer interfacing. For marketing applications, the susceptibility of machine learning methods to these data challenges is not well understood. Moreover, the empirical investigations of these challenges in the machine learning literature are almost universally based on simulations. This is one of the few papers to explore the practical importance of these issues using experimental field data.

The paper proceeds in Section 2 where we review the literature on the four data challenges. In Section 3 we describe the Stage 1 field experiment used to generate the training data, and the Stage 2 field experiment used to validate the trained methods. In the Stage 1 experiment two promotions are sent to randomly selected households. The US Postal Service groups households by carrier route, and because different households in each carrier route received each treatment, the training data provides a measure of

[1] One possible source of past purchasing data for prospective customers is purchasing from competitors. While this information is almost never publically available, there are some markets in which third parties (such as EPSILON Abacus) will score prospective customers using purchases from competitors. However, these opportunities are relatively limited.

the response to each treatment in each carrier route. The targeting variables all vary at the carrier route level, and so we train the targeting methods at the carrier route level, and in the Stage 2 experiment we randomly select which carrier routes will receive each targeting policy. All of the households in a selected carrier route receive the policy recommended for that carrier route. In Section 3 we also describe the twelve-month expected profit measure used to train the methods and evaluate their performance. Under the direction of the retailer that participated in the study, this measure uses the membership type that households signed up for (if any), together with their initial store spending, to project total profits over a twelve-month period.

We present preliminary results in Section 4, including the average profit earned from each of the seven methods, and a discussion of the estimated parameter values for the different targeting variables. In Section 5 we evaluate the robustness of the methods to the different data challenges. The paper concludes in Section 6.

Section 2. Literature Review

As the breadth of machine learning applications has grown, attention has increasingly turned to how robust methods are to different types of data challenges. We investigate the robustness of the seven targeting methods to four data challenges that are typical in the customer acquisition setting. In this section we review the existing literature on each of these challenges, beginning with covariate shift.

2.1 Covariate Shift

Most machine learning methods assume that the underlying distribution of the data used to train the method matches the distribution of the data on which the trained model will be implemented. However, in many situations this assumption does not hold. For example, Bickel and Scheffer (2007) demonstrate that current approaches to collecting training data for spam email filtering algorithms generate datasets with different distributional properties from both the global distribution of all emails and the distribution of emails received by a given user. Similar examples can be found in other fields, including natural language processing (Sugiyama and Kawanabe, 2012), computer vision (Ueki, Sugiyama and Ihara, 2010), robot control (Sutton and Barto, 1998), software engineering (Turhan, 2012), bioinformatics (Baldi and Brunak, 2001; Borgwardt et al., 2006) and brain-computer interfacing (Wolpaw et al., 2002; Sugiyama et al., 2007).[2]

The formal definition of covariate shift is that the distribution of the predictive variables in the training data, $P_{train}(\mathbf{x})$, is different than the distribution in the implementation data: $P_{train}(\mathbf{x}) \neq P_{impl}(\mathbf{x})$ (Shimodaira, 2000; Yamazaki et al., 2007; Moreno-Torres et al., 2012). This can occur for a variety of reasons. In our setting, the training data is obtained from a randomized field experiment implemented in two geographic regions. However, the implementation data extends beyond these two regions, and includes locations that did not participate in the initial experiment. This introduces the possibility (indeed high likelihood) that the characteristics of the implementation data will not match the training data.

Covariate shift is closely related to sample selection bias, which has been well studied in both economics and marketing. If the data used to generate estimates has sample selection bias, the parameter estimates may not be representative of a more general population (Heckman, 1979). For example, Kim and Rossi (1994) demonstrate that consumers with high purchase frequency or high purchase volume are more

[2] Sugiyama and Kawanabe (2012) provide a comprehensive review of this literature.

price-sensitive than other customers. This can lead to inefficient decisions when the models are applied to customers not represented in the training sample.

Various tests have been proposed for detecting covariate shift. These statistical tests are all designed to detect differences in multivariate distributions. They include the multivariate t-test, the generalized Kolmogorov-Smirnov test (Smirnov, 1939; Friedman and Rafsky, 1979), the nonparametric distance-based tests (Biau and Gyorfi, 2005; Hall and Tajvidi, 2002) and the maximum-mean-discrepancy-based test (MMD; Gretton et al., 2012). We use the MMD test in this paper. As we will discuss, this test is distribution-free, computationally efficient and performs strongly across a variety of applications (Gretton et al., 2012; Sejdinovic et al., 2013; Li et al., 2015).

The machine learning literature recognizes that not all methods are equally susceptible to covariate shift. Moreover, covariate shift can impact performance in different parts of the parameter space differently. Regions of the parameter space in which performance is most likely to deteriorate are regions that are under-represented in the training data, but over-represented in the implementation data. While accuracy in these regions is important for implementation, the methods have relatively little information in the training data to learn from.

We group methods into three categories: classification methods, distance-driven methods, and model-driven methods. Distance-driven methods make predictions by weighting *local* observations more heavily than distant observations, while the model-driven methods and classification methods considered in our paper learn *globally*. These differences can affect how the methods perform in regions of the parameter space that are sparsely populated in the training data (Zadrozny 2004). Because they weight the local observations more heavily, the standard errors in the distance-driven methods may increase quickly when there are relatively few local observations. In contrast, the classification and model-driven methods extrapolate from areas with more observations to make predictions about the areas that are sparsely populated. If the models are perfectly specified, this can work well. However, specification errors mean that these projections can lead to large errors. For example, training a linear model to predict age from income can lead to unrealistically high age predictions at the right tail of the income distribution. For this reason, covariate shift can lead to predicted value explosions in methods that estimate globally (classification and model-driven methods), while distance-driven methods may be more robust to this problem.

Chernozhukov et al. (2017) cite predicted value explosions as an explanation for relatively poor out-of-sample predictions using *Lasso* (a model-driven method). They argue that the out-of-sample prediction errors tend to be larger than other methods because of the linear extrapolation outside the range of the training data. Similarly, Shimodaira (2000) demonstrates that covariate shift coupled with parametric model misspecification lead to inefficient MLE estimates.

Hand (2006) provides evidence that more flexible models outperform simpler methods when the covariate distribution is stable, but demonstrate similar performance when covariates shift. Hand's argument is that more complex models are more susceptible to over-fitting features of the training data that may not be stable in the validation data. Hand (2006b) further argues that identifying aspects of the problem that deviate from the standard supervised classification paradigm may have more substantial impact on performance than using a more complex method. Alaiz-Rodriguez and Japkowicz (2008) revisit Hand's conclusion. They use simulations set in a medical domain, predicting the prognosis of patients infected with the flu, and compare the robustness of four methods in the face of covariate shift (and also concept shift). The four methods include two simple methods (a simple 1R classifier and a simple neural network with one node in the hidden layer), and two more complex methods (a C4.5 decision tree and a neural network with ten nodes). They conclude that the performance of more complex prediction methods does not deteriorate faster than the performance of simpler methods in the presence of covariate shift.

Our findings contribute to the debate about whether complex methods are as robust to covariate shift as simple methods. Among the seven methods that we compare, the model-driven methods generally have more parameters than the distance-driven and classification methods. The model-driven methods are arguably more flexible and more complex. They also perform the best in the absence of covariate shift. Consistent with Hand (2006), and unlike Alaiz-Rodriguez and Japkowicz (2008), we find a disadvantage for more complex methods in the face of covariate shift. The model-driven methods deteriorate more quickly than the distance-driven and classification methods.

The second data challenge we study is concept shift, which also focuses on differences between the training data and the implementation data.

2.2 Concept Shift

Our discussion of covariate shift began by recognizing that machine learning methods assume that the distribution of the predictive variables in the training data matches the implementation data. The methods also assume that the relationship between the outcome variable, y, and the predictor variables is stable. Changes in the response function between the two datasets is commonly referred to as concept shift (it also sometimes called concept drift or functional relation change). More formally, concept shift refers to situations in which the conditional distribution $P(y \mid \boldsymbol{x})$ is different in the training and implementation data, while the distribution of covariates $P(\boldsymbol{x})$ remains unchanged (Schlimmer and Granger, 1986; Widmer and Kubat, 1998; Hand, 2006b; Moreno-Torres et al., 2012).

Similar to most targeting settings, our training data is collected in a different time period than our implementation data. The training data was sourced from an initial experiment in spring, and we implemented the targeted policies six months later in fall. This temporal separation between the training data and the implementation data is almost inevitable, although it may be longer or shorter than the six-month interval in our setting. This interval provides opportunities for changes in the environment that may affect the underlying response function. Changes could include seasonality, wearin or wearout of promotions, other actions by the firm, competitors' actions, or even macroeconomic shifts. In our case, the retailer reported that customers were exposed to a lot more mass media advertising during the fall implementation period than during the spring training period.

The general problem of concept shift has been recognized as important and addressed in various domains, including financial prediction (Harries and Horn, 1995), business cycles prediction (Klinkerberg, 2003), clinical studies in medicine (Kukar, 2003; Tsymbal et al., 2006), credit card fraud detection (Wang et al., 2003), computer security and detection of anomalous users (Lane and Brodley, 1998), monitoring and control in industrial operations (Pechenizkiy et al., 2010), and student learning modeling in education (Castillo et al., 2003). More applications are illustrated by Gama et al. (2014), who also provide a broad review of concept shift handling techniques.

The machine learning community has worked on concept shift for over 30 years, starting with the first issue of the journal *Machine Learning* (Schlimmer and Granger, 1986). In later years the machine learning community produced theoretical results with guarantees for learning under concept shift:

Helmbold and Long (1991, 1994), Kuh et al. (1991) and Kuh et al. (1992) all characterize the maximum severity or frequency of concept changes that is tolerable by a learner. Widmer and Kubat (1996) propose a family of algorithms that flexibly react to concept shift in online learning problems. In a special issue of *Machine Learning* on concept drift, Dietterich, Widmer and Kubat (1998) compiled a collection of works that encompass a wide spectrum of theoretical and practical concept shift-related issues (Movellan and Mineiro, 1998; Harries et al., 1998; Auer and Warmuth, 1998; Herbster and Warmuth, 1998; Vicente et al., 1998).

Research on the robustness of machine learning methods to concept shift is scarce. Alaiz-Rodriguez and Japkowicz's (2008) simulations studying the prognosis of patients infected with the flu are perhaps the closest research to our paper. Recall that we cited this paper as an example of research comparing the performance of methods when covariates shift. Alaiz-Rodriguez and Japkowicz (2008) also study the performance of methods in the presence of concept shift. They reach a similar conclusion; more complex prediction methods do not deteriorate faster than simpler methods in the presence of either covariate shift or concept shift.

Because we focus on a different domain and different machine learning methods, our findings help to reveal to what extent Alaiz-Rodriguez and Japkowicz's conclusion that complex methods are as robust to concept shift as simple methods can be generalized. Recall that model-driven methods are arguably more complex than distance-driven and classification methods. They also perform the best in the absence of concept shift. Our results again deviate from Alaiz-Rodriguez and Japkowicz's findings. We find that while model-driven methods perform better than other methods when conditions are ideal, if the response function changes, the performance of these methods deteriorates more quickly than the performance of distance-driven and classification methods.

The direct mail promotions used in this study represent a form of advertising. There has been previous work in the marketing literature recognizing that the response to advertising may be dynamic. This research has focused on developing models of these dynamics, and incorporating them into estimates of the response function. For example, Naik et al. (1998) recognize that the effectiveness of an advertisement varies over time. They incorporate repetition wearout, copy wearout and ad quality restoration in their dynamic brand awareness model. The model is then used to optimize advertising sequencing (see also Erdem and Keane, 1996). In order to incorporate dynamics into the response function, the dynamics must be well understood. In our setting, the change in prospective customer responses to promotions may reflect many different factors. Some of these factors may not be regularly occurring (such as competitors' actions or changes in macroeconomic conditions), so that the change may not be consistent over time. This makes the concept shift that we study difficult to either predict or model.

2.3 Aggregation of the Targeting Variables

Data aggregation is an important topic in marketing and forecasting research. Because individual-level data is often unavailable or expensive, previous studies have sought to understand the value of individual versus aggregate data, and how aggregation affects the performance of a model. In the prospecting problem we study, targeting models are often trained using aggregate data. Individual purchasing histories are not available because the customers have not previously purchased. In the absence of past purchasing histories, firms are often forced to rely upon demographic data. However, demographic data is expensive to purchase at the household level, and is often not accurate. Instead, many firms target prospective customers by purchasing demographic data aggregated to the zip code or carrier route level.

The targeting variables used in this study are all aggregated at the carrier route level. However, some carrier routes have more households in them than other carrier routes. The degree of aggregation thus varies across carrier routes, and it is this variation that we exploit. In carrier routes with relatively few households, the information is less aggregated and more precise than in larger carrier routes that contain more households.[3]

The marketing literature contains many examples of papers studying the performance of models at different levels of aggregation. Notably, aggregate level models often perform as well as models built using disaggregate data. For example, Andrews, Currim and Leeflang (2011) compare sales promotion response predictions with individual-level or store-level data. They find that in general the individual-level models, which capture customer heterogeneity, do not produce more accurate sales response predictions. Gupta, Chintagunta, and Kaul (1996) also compare models built using household versus store-level data. They reach a similar conclusion that disaggregate models may not improve performance.

However, not all results favor aggregate models. Foekens, Leeflang, and Wittink (1994) estimate alternative models at the store-level, chain-level, and market-level model. They compare the models both on whether the parameters are reasonable, and on how accurate the sales forecasts are. In general, the less aggregate models performed better than the market-level model on both benchmarks. More recently, Abhishek, Hosanagar, and Fader (2015) compare models built using daily sponsored search data versus disaggregate data that includes intraday variation in ad position. They show that the aggregate (day-level) data yields biased estimates. In an example from economics, Pesaran, Pierse, and Kumar (1989) compare aggregate (groups of industries) and disaggregate (industry-level) models for predicting the demand for labor. The disaggregate models predict more accurately.

There is also an older literature that characterizes when summing forecasts from less aggregate models is expected to perform better than using a single aggregate model. Grunfeld and Griliches (1960) and Edwards and Orcutt (1969) conclude that the superiority of the less aggregate models depends upon the quality of the model specification. Aigner and Goldfeld (1973 and 1974) show analytically why this is true, and provide more general conditions under which less aggregate models will perform better. The conditions favoring less aggregate models include uncorrelated errors (across sub-samples), and negative correlations in the predictor variables (across sub-samples). On the other hand, if there is substantial measurement error in the predictor variables, this favors the performance of the aggregate models. Aggregate models will also tend to have lower errors when the errors across the sub-samples are negatively correlated or the predictors are positively correlated. After reviewing these characteristics, Foekens, Leeflang, and Wittink (1994) conclude that in practice the relative performance of aggregate and disaggregate models is difficult to anticipate and may vary across settings. This explains why the comparison of aggregate and disaggregate models continues to be an important research topic. In our setting, where we target prospective customers using aggregate demographic data, it is unclear whether the loss of information due to greater aggregation will outweigh the possible cancellation of disaggregate measurement error.

The machine learning literature addresses the issue of aggregation through the lens of information loss. The performance of predictive models deteriorates with lower quality predictors, but the rate of deterioration varies between methods. Model-driven methods can adjust to the weak predictors by

[3] When interpreting the results, we recognize that the size of the carrier route may be related to other carrier route features.

assigning smaller weights to these predictors. For example, *Lasso* assigns a zero weight on the subset of variables with low prediction power (Tibshirani, 1996; Friedman et al., 2009). On the other hand, distance-driven methods may be more sensitive as weak predictors introduce noise to the standard definitions of distance, such as Euclidean distance (Yang and Jin, 2006; Xiang et al., 2008; Bellet et al., 2013). Distance-driven methods also suffer from the curse of dimensionality, which can be exacerbated by the inclusion of weak predictors (Indyk and Motwani, 1998; Jain and Zongker, 1997).

We compare how much the performance of the different methods deteriorates when data is more aggregated. We find that the performance of model-driven methods deteriorates more quickly than distance-driven methods. However, this is not because the model-driven methods perform worse than other methods with more aggregate data; all of the methods perform similarly. Instead, the model-driven methods perform a lot better than distance-driven methods when the data is less aggregated, and they deteriorate faster as the level of aggregation increases. Intuitively, the model-driven methods appear to make the best use of the more precise information in the less aggregate data, and this performance is most susceptible to erosion in the quality of that information.

2.4 Imbalanced Data and Asymmetric Costs for Classifiers

The fourth data challenge that we study is the role of imbalanced data and asymmetric costs. This data challenge is particularly relevant to the performance of the classification methods (CHAID and SVM). In Section 6.4 we will present an example that illustrates why classification methods can perform poorly when data is imbalanced and the cost of errors is asymmetric. There is an extensive machine learning literature on this topic, but we defer a discussion of this literature to Section 6.4. The discussion will be easier to interpret when presented alongside the illustrative example.

We complete our review of the literature by discussing other comparisons of machine learning methods.

2.5 Other Comparisons of Methods

Comparisons of methods have been conducted in various domains independently of the four data challenges we identify. We start with marketing and, in particular, customer segmentation. Similar to our setting, McCarty and Hastak (2007) compare segmentation methods for direct marketing. They report that CHAID outperforms basic recency, frequency and monetary value (RFM) analysis in settings where the response rate is low, and the marketer can target only a small portion of the entire database. We note that their RFM analysis uses transaction variables about the past behavior of customers that are not available in prospecting campaigns. Olson et al. (2009) find that data mining methods (logistic regression, decision trees, neural networks) performed better than basic RFM analysis for customer segmentation; but found little variation in the performance of the data mining methods.

In economics, Stock and Watson (2005) empirically compare methods for macroeconomic forecasting and find that models that reduce dimensionality using factor analysis are more accurate than models that do not, such as OLS regression and bagging. In microfinance, Wu et al. (2010) find that Naive Bayes and Bayesian networks perform better than clustering methods when making decisions about giving loans to sub-prime borrowers.

Finally, there is a rich literature in comparing methods for medical predictions, yet no overarching statement can be made about the relative performance of families of methods. We provide some characteristic examples. Tong et al. (2016) report that Lasso logistic regression predicted all-cause non-elective readmission risk of patients better than stepwise logistic regression, AdaBoost, and a readmission risk index (especially when the sample size is small). A decision tree (C5) outperformed a neural network and logistic regression in predicting breast cancer survivability in a study by Delen et al. (2005). SVM outperformed logistic regression, discriminant analysis, neural networks, fuzzy clustering and random forests in predicting diabetes in a study by Tapak et al. (2013). Penny and Chesney (2006) find that neural nets outperformed CART, logistic regression, and another tree-based classification algorithm (C5) for predicting death following injury.

In the next section we introduce the research context, describe the design of the experiments, and discuss the targeting variables and the output measure used to train and evaluate the models.

3. Data Overview

3.1 Research Context

Data for this study was provided by a large retailer that operates a large number of stores in the US. The retailer sells a broad range of products including perishables, sundries, and durables. Customers can only purchase if they have signed up for a membership. The retailer regularly mails promotions to prospective members in order to increase its membership base. We study two types of promotional offers. The first offer is a $25 paid 12-month membership, which represents a 50% discount off the full price. The second offer is a 120-day free trial. Customers who want to maintain their memberships after the trial period must purchase a regular membership at the full price. When new customers register for a membership, they provide their name and mailing address. This is used to identify which households responded to which direct mail promotions.

3.2 Design of the Stage 1 Experiment

The first-stage (training) experiment included 1,185,141 households in a mailing list purchased from a third party commercial data supplier. The households were all located in two geographic regions. The households were assigned to three experimental conditions, randomized at the household level. The three conditions included a control group that received no promotion offer, a group that received the $25 paid offer, and a group that received the 120-day free trial. The promotions were highlighted on the front cover, inside front cover and back cover of a 48-page book of product-specific coupons. The treatments were repeated twice approximately six weeks apart, so customers received the same offer twice. The first treatment was implemented in early February 2015, and it was then repeated in late March.

The United States Postal Service organizes households into carrier routes. These routes identify the households that each letter carrier visits. There are typically 200 to 400 households per carrier route and they all fall within the same zip code. Both when training the models using the Stage 1 field experiment, and when implementing the targeting policies in Stage 2, we used data aggregated to the carrier route level.

The Stage 1 experiment included 5,976 carrier routes. Because we randomized at the household level in this stage, each carrier route includes households randomly assigned to each of the three experimental treatments. Therefore, for each carrier route we can calculate the profit earned under each of the experimental treatments: $25 paid offer, 120-day free trial, and control.

3.3 Targeting Variables

The Stage 1 experiment was used to provide data for training the seven targeting models. Before training the models we needed to decide which targeting variables to use. Recall that our goal is to compare typical implementations of each method (by a moderately sophisticated retailer), and to evaluate how robust their performance is to different data challenges. This goal of implementing a version of the model that a moderately sophisticated retailer would implement is a different goal than identifying the variant of each method that yields the optimal performance. For this reason we chose to use the same variables that the firm used in its OLS-based training models. Moreover, like the retailer, we restricted attention to linear relationships, and did not include any interactions between the variables or non-linear transformations of the variables. We recognize that including additional targeting variables, or allowing for non-linear relationships, may have improved performance.

The retailer provided thirteen targeting variables for the carrier routes in Stage 1, and then subsequently provided the same targeting variables for the carrier routes in Stage 2. All variables vary at the carrier route level. Five of the variables, *Age*, *Home Value*, *Income*, *Single Family*, and *Multi-Family* variables were purchased by the retailer from a third-party commercial data supplier. The remaining variables were constructed by the retailer using its own data. A complete list of variables together with their definitions and summary statistics are presented in the Appendix.

In preliminary analysis we also investigated estimating the models at the household level instead of the carrier route level. To facilitate this comparison, the retailer purchased household-level data for the *Age*, *Home Value*, and *Income* variables from the third-party commercial data supplier. We used a holdout sample and the outcomes from the Stage 1 experiment to compare whether training the models at the household level improved performance over models trained at the carrier route level. There was no improvement in the ability to predict outcomes in the holdout sample when using household-level data. One possible explanation is that the household-level data provided by the data provider is not sufficiently accurate to improve performance. The household-level data is more expensive to purchase, and so throughout the study we use data aggregated to the carrier route level.

3.4 The Targeting Methods

The Stage 1 experiment provided an outcome measure for each of the three treatments for all 5,976 carrier routes that participated in that experiment (we describe the outcome measure in detail later in this section). This outcome measure, together with the thirteen targeting variables, comprise the data that we use to train the seven targeting models.

The targeting methods that we compare use two different approaches to design optimal targeting policies. Most of the methods use a regression-based approach, under which a separate model is trained for each of the three treatments (No Mail, $25 paid promotion, and the 120-day trial promotion). These models are then used to predict the outcomes for a new observation (i.e., a new carrier route). The methods assign to the new observation the treatment with the highest predicted profit. Regression-based methods can be further categorized as non-parametric distance-driven methods, or parametric model-driven methods. We implement three distance-driven methods, together with two model-driven methods (see below).

An alternative approach to the problem is to classify the observations in the training data according to which of the treatments yielded the highest profit. The thirteen targeting variables are then used to predict this classification of the training data outcomes. We evaluate two classification methods.

Distance-Driven Methods	**Model-Driven Methods**	**Classification Methods**
kernel regression	Lasso regression	chi-square automatic interaction detection (CHAID)
k-nearest neighbors (k-NN)	finite mixture models (FMM)	support vector machines (SVM)
hierarchical clustering (HC)		

We also evaluate three naïve benchmarks. These benchmarks implement "uniform" policies, which assign each new observation the same treatment: the policy assigning the $25 paid membership uniformly, the policy assigning the 120-day free trial uniformly, and the policy assigning the no-mail treatment uniformly.

Recall that our criterion for selecting these methods was whether a moderately sophisticated retailer would be able to implement the method. We used the same criterion when formulating each of the methods. Because the implementations are not novel (by design), we relegate a discussion of each method and its implementation to the Appendix.

3.5 Design of the Stage 2 Experiment

The second-stage (validation) experiment involved 4,119,244 households organized into 10,419 carrier routes. The mailing list for this experiment was purchased from the same third-party data supplier that provided the mailing list for the stage 1 experiment. The 10,419 carrier routes were randomized into ten experimental conditions. These included the seven optimized conditions (one for each of the seven segmentation methods), and three uniform conditions ($25 paid offer, 120-day free trial, no mailing).

In Stage 2 we randomized by carrier route instead of households, so that every household in a carrier route was in the same experimental condition. This decision was motivated by lower mailing costs. The United States Postal Service offers cheaper mailing rates if every household in a carrier route receives the same mailing. Notice that this is a difference from the Stage 1 experiment, where randomization occurred at the household level, thus ensuring a measure of the response to each of the three treatments ($25 paid offer, 120-day free trial, and control) in each carrier route in Stage 1.

The Stage 2 experiment shared many of the same features as the Stage 1 experiment. The study involved the same retailer sending direct mail solicitations to prospective customers identified through rented mailing lists. As in the Stage 1 experiment, the mailing treatments in the Stage 2 experiment were repeated six weeks apart, with customers receiving the same treatment in each wave. The promotions were also the same, including a $25 discounted membership, a 120-day free trial, and a no-mail (control) condition. We used the same thirteen targeting variables that were available in the first-stage experiment (summary statistics for the Stage 2 mailing list are also provided in the Appendix).

However, there are several differences in the design of Stages 1 and 2. First, the households were in a much broader geographic area in Stage 2, with just 2% of the households located in the same geographic area as the first-stage experiment. Second, the second-stage experiment was conducted in the fall (starting on August 23, 2015), while the first-stage experiment was conducted in the spring (of the same year). Third, the Stage 1 treatments were randomly assigned at the household level, while in Stage 2 the ten mailing policies were randomly assigned at the carrier route level. Fourth, in Stage 2 the promotions were mailed to prospective customers using a postcard, which was printed on both sides and highlighted the offer on each side. Recall that in Stage 1 the offers were printed on the outside and inside covers of a 48-

page book of coupons. Finally, the Stage 2 experiment coincided with a mass media advertising campaign by the retailer. There was no such mass media advertising during the Stage 1 experiment.

These types of differences between the experiments are typical in a field setting. They are likely to lead to differences in the response functions between the training and validation experiments, which will tend to diminish the accuracy of the optimized policies. However, the randomization ensures that this affects all of the optimization methods. Moreover, this variation provides an opportunity to compare the methods in realistic conditions. In practice, these types of differences are common, and a comparison of the methods across identical experiments is perhaps less informative than a comparison of the methods in realistic conditions. We will exploit these differences in Section 5, where we evaluate how robust the different methods are to differences in the training and validation data.

3.6 Profit Measure

We train the methods and compare their performance using the "profit" earned from each household. This profit measure estimates profits in the twelve months after each experiment's first mailing date. Recall that the two experiments are only six months apart, and so it was not possible to measure the full twelve-month outcome from the first experiment before the start of the second experiment. Instead, the retailer provided a formula for projecting twelve-month profits using an initial "tracking window".

Mailing decisions for the Stage 2 experiment needed to be made several weeks before the mail date of that experiment. In order to make these mailing decisions, many steps were required to construct the Stage 1 training data, train the models, and finalize the experimental design for Stage 2. These steps all had to be performed after the end of the tracking window for the Stage 1 experiment. As a result, we had to extract the Stage 1 data and start preparing for the second experiment well before the mailing date of the second experiment. Consequently, the training data is constructed using a tracking window from the Stage 1 experiment that includes 63 days after the first mailing date to identify new members, and 77 days after the first mailing date to track transactions from these new members.[4] To ensure that we validated the seven methods using the same measure that we trained them on, we used an identical tracking window and estimation formula to compare their performance in the Stage 2 experiment.

The retailer's profit estimation incorporates key factors that contribute to customer retention and profitability. First, it recognizes that membership revenue has no cost of goods sold and contributes directly to profit. Second, it applies a constant profit margin to in-store purchases (which is consistent with the retailer's pricing practices).[5] Third, it distinguishes between households that did not respond, households that signed up for a trial membership, and households that signed up for a regular paid membership. The distinction between trial and regular membership is important. The retailer only receives membership revenue from the paid memberships. Moreover, because not all trial members convert to regular memberships at the end of their trial, the membership type plays an important role in customer retention over the twelve months. This conversion probability has been carefully studied by the retailer, and is incorporated into the profit calculation. Finally, the projection of in-store purchasing over the twelve months was specific to each household and is based upon the household's in-store spending during the first 77 days.[6]

[4] Different data extraction processes are used to identify new members and track purchases.

[5] Charging a constant markup is common for this type of retailer.

[6] A small number of customers purchased products (primarily cigarettes) from the retailer to (presumably) sell at their own retail locations. This introduced outliers to measures of store revenue. At the suggestion of the retailer,

Although the profit calculation is confidential, we document the sources of variation in the Appendix. The mailing costs only varied according to whether a household received a promotional mailing. Because this variation is small, it explains little variation in the twelve-month expected profit measure. Instead, the variation in expected profit is almost equally attributable to a household's membership decisions (including the choice of membership type), and the amount that new members spent in the store during the first 77 days.

In the Appendix we also discuss the possibility that a retailer may impose a ceiling on the total number of mailings. We discuss how this type of constraint could be accommodated by the model-driven and distance-driven methods. For ease of exposition, throughout the paper we use the term *Profit* to refer to the expected twelve-month profit measure described in this sub-section.

4. Preliminary Results

4.1 Average Profit in Each Experimental Condition

We compare the *Profit* earned from the households in each of the experimental conditions in Figure 4.1. The findings reveal that the policy produced by Lasso yielded the highest average profit. The average profit was significantly higher than CHAID and SVM, and it also significantly outperformed all uniform policies (p < 0.05). The k-NN method also significantly outperformed CHAID (p < 0.01). However, it did not significantly outperform the uniform $25 condition.

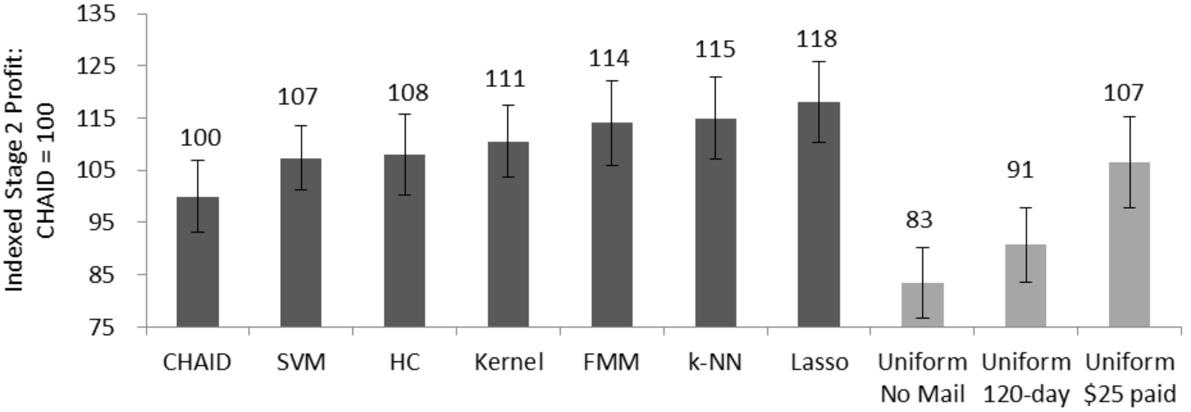

Figure 4.1: The figure reports the average Stage 2 *Profit* averaged across each of the households in each experimental condition. To preserve confidentiality, the profits are indexed to 100 for the CHAID data point. The error bars represent 95% confidence intervals. Complete findings including sample sizes and standard errors are reported in the Appendix.

We can also compare the performance of the methods when grouping the methods using our taxonomy of methods. In Figure 4.2 we report the analysis when pooling the households according to this taxonomy. Notice that this grouping preserves the benefits of randomization, although the sample sizes vary because there are more distance-driven methods than model-driven or classification methods. The results clearly favor the distance- and model-driven methods over the classification methods. The two classification methods (CHAID and SVM) are the two worst performing methods. Indeed, when we pool the outcomes

store revenue in the first 77 days that exceeded $15,000 was truncated to $15,000. This affected just eight of the 4,119,244 observations in the Stage 2 validation data. It did not affect any of the 1,185,141 observations in Stage 1.

by method type, the distance- and model-driven methods yield policies that are both significantly better than the classification outcomes (p < 0.01). The difference between the distance- and model-driven methods is not significant.

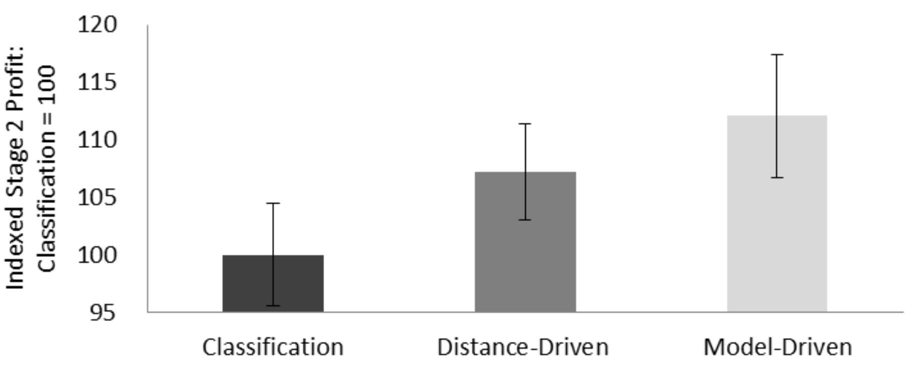

Figure 4.2: The figure reports the average Stage 2 *Profit* when pooling the households using the taxonomy of methods. To preserve confidentiality, the profits are indexed to 100 for the Classification methods data point. The error bars represent 95% confidence intervals. Complete findings including standard errors and sample sizes are reported in the Appendix.

4.2 Comparison with the Uniform $25 Policy

Although the Lasso method significantly outperforms the uniform policy of sending every household the $25 paid offer, it is the only optimization method to do so. While most of the other optimized policies earn higher average profits than the uniform policies, the differences compared to the uniform $25 paid policy are generally not statistically significant. Although this may seem disappointing, it is not surprising. Most of the methods choose to send the $25 paid offer to the majority of households. For households for which a policy would recommend sending the $25 paid offer, there is obviously no difference in the expected profit compared to the uniform $25 policy. If we want to focus on the differences between an optimized policy and this uniform policy, we need to focus on the households for which the policies are different. In particular, we need to compare the households for which the optimized policy would not send the $25 paid offer. Because of the experimental variation, we can make this comparison using a randomly selected group of customers who received the $25 paid offer (those assigned to the uniform policy) and a randomly selected group of customers that received another treatment (those assigned to the optimized policy).

We can illustrate this logic more clearly using an example. Consider the households in the Lasso condition and the households in the uniform $25 paid condition. We can ask what action Lasso would have recommended for all approximately 800,000 customers (pooled across both conditions). It would have recommended sending the $25 paid offer to 73.63% of them, sending the 120-day free trial to 8.00% of them, and not mailing to the remaining 18.37%. These three sub-groups of households are systematically different (otherwise Lasso would not have chosen to treat them differently). However, within each sub-group there is a sample of customers randomly assigned to the Lasso policy, and an equivalent sample randomly assigned to the uniform $25 paid policy. We can safely compare these equivalent samples. The results are summarized in Figure 4.3. Where Lasso chose the 120-day or no-mail treatments over the $25 paid treatment, it outperformed the uniform $25 policy. The difference is statistically significant (p < 0.01) for the no-mail treatment, and when pooling the 120-day and no-mail groups (the columns at the far right in the figure).

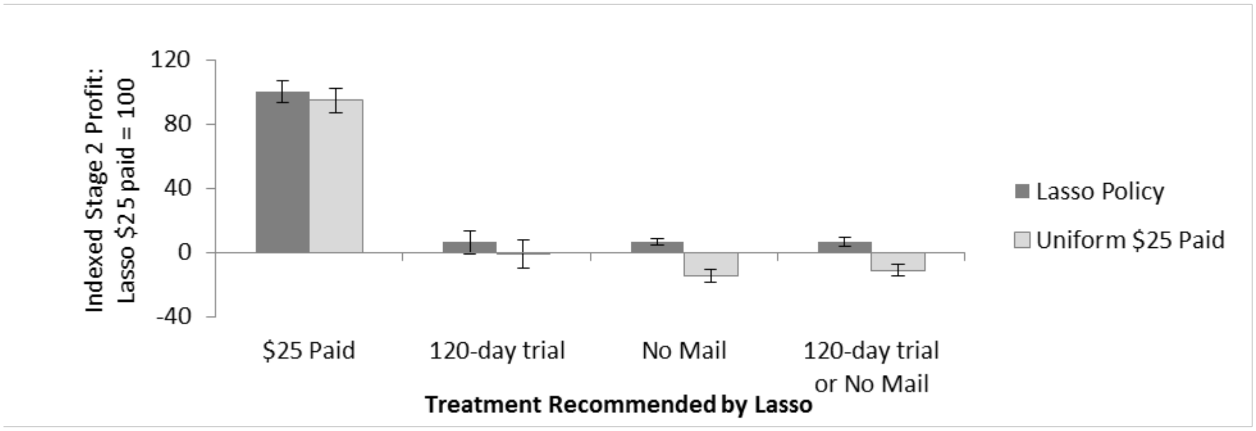

Figure 4.3: The figure focuses on households in the Lasso and uniform $25 paid conditions. It reports the average Stage 2 *Profit* when grouping households by the treatment recommended by Lasso. To preserve confidentiality, the profits are indexed to 100 for the Lasso Policy $25 Paid offer data point. The error bars represent 95% confidence intervals. Complete findings, including standard errors, are reported in the Appendix.

Reassuringly, we do not observe a significant difference between conditions for the households for which Lasso recommended sending the $25 paid offer. This is true even though this is the largest sub-group of households (73.63% of them). For these households, the comparison between conditions represents a randomization check. They received the same treatments and so any differences in the outcome could only be attributed to differences in the households themselves. The findings also reveal another distinctive pattern. Lasso recommends mailing the $25 paid offer to the most valuable households. It is only the less valuable households to which it chooses to send the 120-day free trial (or not to mail). In the Appendix we repeat this analysis for all of the optimized models.

As a final set of preliminary results, in the next sub-section we report how the profits earned from the three experimental treatments in Stage 1 varied with respect to each of the targeting variables.

4.4 Parameter Values

To demonstrate the relationship between the *Profit* and the targeting variables, in the Appendix we report the parameter estimates when using OLS to regress the *Profit* outcome measure on the thirteen targeting variables. There are no significant differences in the direction of the relationship between profits and the targeting variables across the three treatment conditions, but coefficients vary in their magnitudes. This variation provides an opportunity for the targeting methods to vary the optimal action across carrier routes. Across the three models, the strongest indicator that a carrier route will yield large profits is a high previous response rate (*3yr Response*). The coefficient on this variable is approximately three times larger than any other coefficient (in absolute value). Other significant coefficients indicating larger expected profits include: a short distance to the nearest own store (*Distance*), a long distance to the competitors' store (*Comp. Distance*), a concentration of single family housing (*Single Family*), a low average age (*Age*), and a high proportion of households that were previously paid members (*Past Paids*).

Our initial findings indicate that the model-driven and distance-driven methods had a better overall performance than the classification methods. In the next section we investigate the robustness of this finding by re-examining the performance of the methods in different regions of the parameter space. This allows us to evaluate the relative performance of the methods when confronted with data challenges that are typical in a prospecting setting.

5. How Well Do the Methods Address the Challenges of Targeting Prospects?

In the Introduction we identified typical challenges of using the response from an initial experiment to target prospective customers. The first challenge is the risk of covariate shift. If the distribution of the targeting variables in the Stage 1 training data is different than the Stage 2 validation data, then the trained policies may not extrapolate well to the validation data.

The second challenge is that the response function itself may not be stationary, which the machine learning literature refers to as "concept shift". This could reflect a wide variety of factors, including seasonality, macroeconomic conditions, competitive actions, or exposure to intervening promotions. If the response function changes from the training data to the validation data, this may contribute to deterioration in the performance of the optimized policies.

A third obstacle is that the targeting variables are often measured at an aggregate level. This is particularly common when targeting prospective customers. Because these customers have not previously purchased, there is no past purchasing data to use as a targeting variable. Instead, firms often rely upon demographic variables, which are typically aggregated across census blocks, zip codes or carrier routes.

We start this section by evaluating the seven targeting methods on different segments of the data to assess how well the methods address these first three challenges. We conclude the section by focusing on the fourth data challenge to explain why the classification methods performed so poorly. This discussion focuses on data imbalance and asymmetric costs of error. Prospecting for new customers typically yields a very low response rate, but a large profit from the prospective customers that do respond. The machine learning literature recognizes that these characteristics can cause difficulties for classification methods. We discuss how the methods can be modified to overcome these difficulties and improve performance. We start by investigating the challenge posed by covariate shift.

5.1 Covariate Shift

One factor that could influence the performance of targeting models is the extent to which the distribution of the customers' characteristics in the Stage 2 validation data matches the distribution in the Stage 1 training sample. We would expect targeting models to perform better (compared to naïve benchmarks) in regions of the parameter space that are well represented in the training data.

This issue is particularly relevant in settings such as ours, where the geographic regions in the training data and validation data vary. Recall that in our study, the training data is drawn from just two geographic regions, while the validation data is drawn from a much broader geographic area. As a result, it should not be surprising if the validation data contains regions of the parameter space that are not well represented in the training data.

We first employ a statistical test to establish the covariate shift between our spring training data and fall validation data. We formally identify the covariate shift between training and validation data using the Maximum Mean Discrepancy test (MMD). The MMD test is a statistical test of the null hypothesis that two distributions, f and g, are equal, $H_0: f = g$, against the alternative hypothesis $H_A: f \neq g$. Suppose there are n samples drawn from the multivariate distribution f, and m samples drawn from the distribution g. The test first finds a smooth function that is large on the points drawn from f, and small (negative) on the points drawn from g. The test statistic, MMD, is then the difference between the mean function values on the two samples. When the MMD is large, the samples are likely drawn from different distributions.

The choice of the class of smooth functions is important for the performance of the test statistic. There is a tradeoff between the richness of the class of functions and the convergence properties of the estimator. We want MMD to vanish only when $f = g$, but the empirical estimate of MMD converges to its expectation slowly for richer classes. Gretton et al. (2012) derive the test thresholds for the MMD statistic with functions from a reproducing kernel Hilbert space. In our test, we use the Gaussian kernel with a bandwidth equal to the median Euclidean distance between the data points (Caputo et al., 2002). For computational efficiency, we reduce the sample size by applying the MMD test to 5,000 carrier routes randomly drawn from each of the training and validation datasets. The test rejects the null hypothesis of equal distributions at the 1% significance level.

As an initial investigation of the effect of covariate shift on the performance of the methods, we first calculated the mean and standard deviation of each of the thirteen targeting variables using the Stage 1 training data. We then identified carrier routes in the Stage 2 validation data for which one or more of the variables was at least two standard deviations away from the (training data) mean. This procedure revealed that 60.5% of Stage 2 carrier routes fell outside the two standard deviation range on at least one of the thirteen targeting variables. We classified these validation sample carrier routes as “Outside the Range” of the training data, and the remaining carrier routes as “Inside the Range”. We then recalculated the average outcome for each method using the two groups of carrier routes. The findings are summarized in Figure 5.1.1.

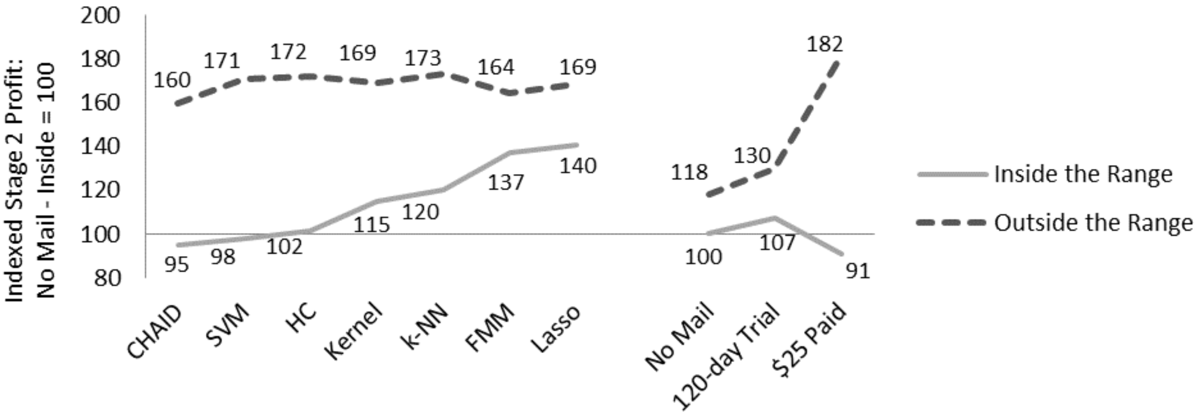

Figure 5.1.1 The figure summarizes the average Stage 2 *Profit* when grouping the households according to whether they are inside or outside the range of the training data. The profits are indexed to 100 in the No Mail control for the Inside the Range data point. We categorize the households in the Stage 2 validation data as falling inside the range of the Stage 1 training data if they have values on all thirteen targeting variables within two standard deviations of the average value in the training data. Complete findings are reported in the Appendix.

The findings reveal several interesting patterns. First, we see that the households in the Stage 2 validation data that were outside the range of the Stage 1 training data were generally more valuable households.

Second, the $25 paid uniform policy outperforms the other uniform policies on the households that are outside the range of the training data, but the 120-day trial is the best uniform policy inside the range of the training data. This result is consistent with our observation in Section 4.2 that Lasso’s optimized policy recommends mailing the $25 paid offer to the most valuable households, and the 120-day free trial to the less valuable households.

Third, when we focus on the households that were inside the range of the training data, the power of the model-driven methods is revealed. Lasso and FMM sharply outperform the other methods, and also all three of the uniform policies. The implication is that these methods make the best use of the information in the training data.

Finally, the seven optimized methods all perform similarly on the households outside the range of the training data. Moreover, all seven methods yield optimized policies that underperform the $25 paid uniform policy. This highlights the cost of covariate shift. Training models on data that is not representative of the data that the models will be implemented on yields relatively poor policies compared to the naïve benchmark of mailing the $25 paid promotion to everyone.

We highlight how covariate shift contributes to the deterioration in the performance of the methods by comparing their performance to a naïve benchmark. For the naïve benchmark we use the most profitable uniform policy. Inside the range of the training data the most profitable uniform policy was the 120-day trial, while outside the range of the training data the $25 paid promotion was the most profitable. In Figure 5.1.2 we compare the methods grouped using our taxonomy against these two benchmarks.

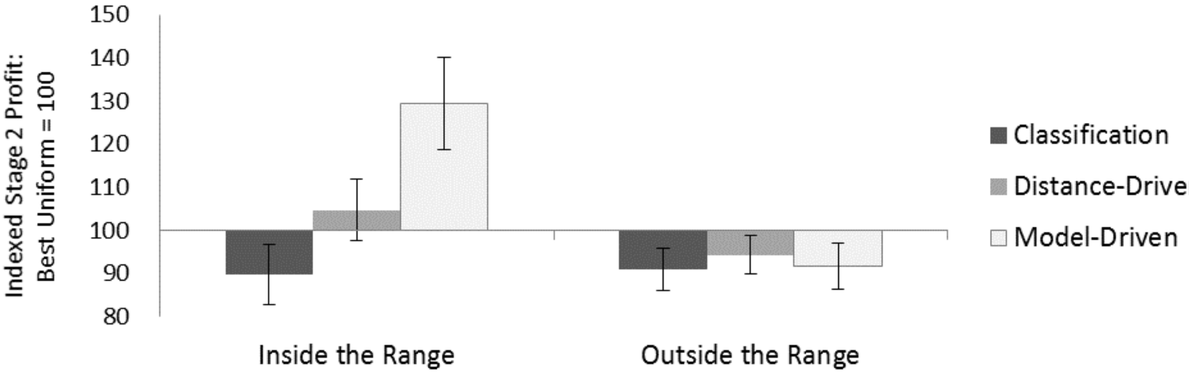

Figure 5.1.2 The figure summarizes the average Stage 2 *Profit* when grouping the households according to whether they are inside or outside the range of the training data. Profits are indexed to 100 for the most profitable uniform condition (in that group of carrier routes). We categorize the households in the Stage 2 validation data as falling inside the range of the Stage 1 training data if they have values on all thirteen targeting variables within two standard deviations of the average value in the training data. The error bars represent 95% confidence intervals.

Inside the range of the training data, the model-driven methods significantly improve upon their benchmark. However, outside the range all methods perform equally poorly compared to that benchmark. The deterioration in performance due to covariate shift is particularly striking for the model-driven methods. The advantage they enjoy within the range of the training data is eliminated.

One interpretation of these findings is that model-driven methods make better use of the information provided by the covariates in the training data. When the validation data matches the training data, this results in improved performance. However, when the training and validation data are poorly matched, exploiting the information in the training data does not help to improve performance in the validation data. Under this interpretation, the deterioration in performance is larger on the methods that make the best use of the training data.

Recall from our discussion in Section 2 that there are conflicting arguments in the machine learning literature about how robust different types of methods will be in the presence of covariate shift. These arguments focus on the way that the methods predict outcomes in regions of the training data that are sparsely populated. Distance-driven methods make predictions by weighting *local* observations more heavily than distant observations, while the model-driven methods considered in our paper learn *globally* (Zadrozny 2004). When we examine the predicted profits in the sparsely populated regions of the training data, we see more extreme predictions and larger variance in the predictions from the distance-driven methods than the model-driven methods. This is not consistent with an argument that model-driven methods yield more extreme predictions when they extrapolate from areas with more observations to make predictions about the areas that are sparsely populated. Instead, the large variance in the predictions from the distance-driven methods is consistent with local estimation in sparsely populated regions.

The covariate shift findings in Figures 5.1.1 and 5.1.2 indicate that when the distance between the training data means and the validation data is too *large*, then the targeting methods perform worse than a naïve benchmark. We next evaluate whether the targeting methods may also perform worse than a naïve benchmark if the distance is too *small*. It is possible that the validation data may be different from the training data because the validation data is too few standard deviations from the training data means.

To illustrate this we group the carrier routes according to the largest deviations from the mean values of the thirteen targeting variables in the training group data. In particular, we use the Stage 1 training data to calculate the mean and standard deviation of each of the thirteen targeting variables. For each variable and carrier route, we can then calculate the number of standard deviations between the value of this variable in the carrier route and the training group mean. We group the carrier routes using the largest absolute deviation across the thirteen variables (we label this the *Maximum Deviation*).[7] In Figure 5.1.3 we report a histogram of the *Maximum Deviation* for carrier routes in the training and validation samples.

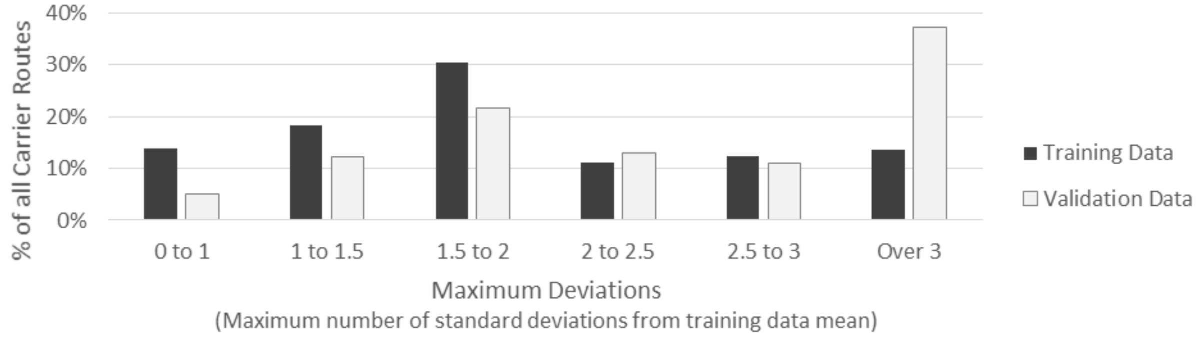

Figure 5.1.3 The figure is a histogram illustrating the distribution of carrier routes in the training and validation samples according to the maximum (across the thirteen targeting variables) of the number of standard deviations from the training data mean. The columns for the Training Data and the Validation Data each add to 100%.

We see that relatively few of the carrier routes in the validation data have *Maximum Deviations* less than two standard deviations from the training data means. In contrast, many of these carrier routes have *Maximum Deviations* at least three standard deviations from the training data means.

In Figure 5.1.4 we see how this affects the performance of the targeting methods in Stage 2. In particular, we report the average profits within each group of Stage 2 carrier routes, where the profits are indexed at

[7] In this analysis we focus on the maximum deviation across the thirteen targeting variables. However, we also obtain similar findings when we repeat the analysis using the average deviation across the thirteen variables.

100 using the most profitable uniform policy (within each group) as a benchmark.[8] The findings indicate where the targeting policies were able to improve upon the naïve benchmarks. The ability of the targeting models to improve upon the naïve benchmarks does not deteriorate monotonically with the largest deviations from the training data means. The distance- and model-driven targeting methods both perform best on carrier routes where the *Maximum Deviation* is 1.5 to 2 standard deviations from the training data mean. This is the group of carrier routes that are best represented in the training data (Figure 5.1.3).

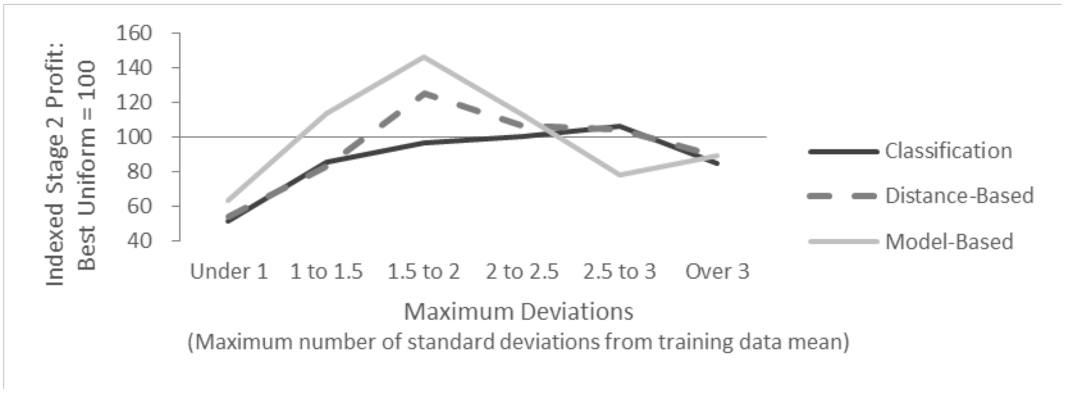

Figure 5.1.4 The figure summarizes the average Stage 2 *Profit* when grouping the carrier routes according to the largest deviation (across the targeting variables) from the average in the training data (measured in training data standard deviations), and when pooling using the taxonomy of methods. The profits are benchmarked at 100 using the most profitable uniform policy (in that group of carrier routes).

Among carrier routes with larger *Maximum Deviations* from the training data means, the targeting methods add essentially no value over the naïve benchmarks. Perhaps more surprisingly, we also see that the targeting methods add no value in carrier routes that have the smallest deviations from the training data means.

We note that this last finding cannot simply be attributed to these carrier routes being less valuable. In Figure 5.1.4 the profits are all indexed against the most profitable uniform policy for that group of carrier routes. This benchmark controls for the variation in the relative value of the groups of carrier routes. When the thirteen targeting variables all have values close to the training data mean, none of the targeting methods beat the naïve benchmark. Instead, it appears that a tight grouping around the training data mean is another form of covariate shift. We see from Figure 5.1.3 that fewer than 14% of the training data carrier routes have values on all thirteen targeting variables that are tightly grouped within one standard deviation of the training data mean. Because this region of the parameter space is not well represented in the training data, the targeting methods have relatively little information to make predictions in this region of the parameter space.

For completeness, we also investigated which targeting variables most frequently shifted between the two stages of the study. The most common deviations are *Income* and *Home Value* (higher in the validation data), distance to the competitors' and retailer's own stores (higher in the validation data), and the proportion of past trialists in the carrier route (higher in the validation data). These findings are summarized in the Appendix, where we also report which deviations had the largest impact on the performance of the targeting methods.

[8] The 120-day benchmark was used for the *Under 1* group, the *No Mail* benchmark was used for the *1 to 1.5* group, and the $25 benchmark was used for the remaining groups.

In the next sub-section we analyze how the methods performed as the underlying demand conditions varied between Stages 1 and 2.

5.2 Concept Shift

The interval between the first mailing date in Stage 1 and the first mailing date in Stage 2 was just over six months. In this intervening period it is possible (even likely) that underlying demand conditions changed in different ways in different carrier routes. This non-stationarity could lead to changes in how prospective customers responded to the promotions. The machine learning literature refers to this as "concept shift". In this section we investigate how concept shift affected the relative performance of the targeting methods.

Our discussion of concept shift proceeds in the following steps. First, we provide initial evidence of concept shift by comparing the outcome of the uniform policies in Stages 1 and 2. Second, we construct a measure of how underlying demand conditions changed between the two stages of the study. Changes in underlying demand conditions are one source of concept shift, and so this measure provides an indicator of which carrier routes were most likely to experience concept shift. Third, we support our claim that underlying demand changes are associated with concept shift by comparing our initial measure of concept shift (the differences in the Stage 1 and 2 uniform policy outcomes) with our measure of changes in underlying demand conditions. Finally, we compare the performance of the seven optimized policies in Stage 2 when grouping the carrier routes using our measure of demand changes. The methods all perform worse when there are positive or negative changes in underlying demand, compared to when demand changes are flat. This is true for all three types of models, but the deterioration in performance is particularly large for the model-driven methods. We start by providing initial evidence of concept shift by comparing the performance of the uniform policies in Stages 1 and 2.

5.2.1 Initial Evidence of Concept Shift

As initial evidence of concept shift, in Figure 5.2.1 we compare how the *Lift in Profits* attributable to the $25 paid and 120-day trial promotion conditions varied between Stages 1 and 2. The *Lift in Profits* is calculated as the average profit in the promotion condition minus the average profit in the no mail control condition. In this figure (and throughout this sub-section) we restrict attention to the 234 carrier routes that received uniform treatments in both stages of the study. Focusing on this common sample of carrier routes essentially eliminates covariate shift, as there was almost no variation in the covariates within a carrier route between the two stages. Any remaining difference in the performance of the promotions across the two stages can thus be attributed to concept shift. To protect confidentiality, in Figure 5.2.1 and the other findings reported in this sub-section (5.2), we multiply the profits by a (common) random number.

The findings reveal an important difference between the Stage 1 and Stage 2 results for these 234 carrier routes. The *Lift in Profits* attributable to the promotions were positive in Stage 1 but negative in Stage 2.[9] The differences in these outcomes between the two stages could reflect a wide variety of factors, including seasonality, wear-in or wear-out of the promotions, competitor's actions, or other activities by the retailer. Whatever the reason, these differences are examples of concept shift that could affect the

[9] The Stage 2 revenue was higher in the promotion conditions than in the no mail control. However, the profits in the promotion condition were lower than in the control because the profit contributed by the incremental revenue was less than the cost of mailing the promotions.

performance of the targeting models. This suggests that our data provide an opportunity to understand how robust the seven targeting methods are to concept shift in a real-world marketing application.

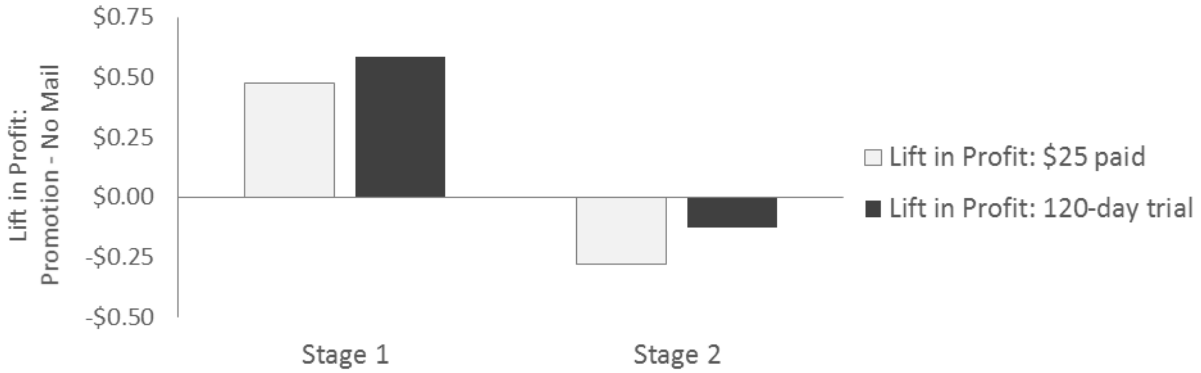

Figure 5.2.1 The figure summarizes the average difference in *Profit* between each promotion and the no mail control. The results are reported separately for the $25 paid and 120-day trial promotions and for the Stage 1 and Stage 2 experiments. The average profits are calculated using the 234 carrier routes that received uniform treatments in both Stages 1 and 2. To protect confidentiality, we multiply the profits by a (common) random number.

5.2.2 Measuring Changes in Underlying Demand Conditions

To study the impact of concept shift, we first quantify how underlying demand conditions changed between Stages 1 and 2. We construct a demand change measure using purchases by existing customers. In particular, we identify customers who had been members for at least five years at the start of Stage 1. Using these existing customers we calculate total revenue during Stage 1 and total revenue during Stage 2 for each of the retailer's stores. We then construct a measure of *Revenue Change* for each store:

$$Revenue\ Change = \frac{(Stage\ 2\ Revenue - Stage\ 1\ Revenue)}{0.5 *\ Stage\ 1\ Revenue + 0.5 *\ Stage\ 2\ Revenue}$$

Using the average of Stage 1 and 2 revenue as the denominator ensures that increases and decreases are treated symmetrically. Variation in this measure may result from almost any local factor, including macro-economic variation, seasonality, competitors' actions, or exposure to promotions. Our findings are robust to how this measure is constructed. For example, if we use all existing customers who were members prior to the start of the study we obtain a very similar pattern of results. We also obtain similar results if we restrict attention to revenue from a common set of existing customers who made purchases in both Stage 1 and Stage 2. This ensures that the findings are not affected by attrition of existing customers from the panel.

Each carrier route is associated with a single store (the correspondence between the stores and the carrier routes is defined by the retailer), and our 234 carrier routes map to fifteen different stores. The distribution of *Revenue Change* across these fifteen stores is summarized in Figure 5.2.2, where we use different colored shading to group the stores as follows:[10]

Negative Growth	*Revenue Change* less than -0.02
Flat Growth	*Revenue Change* between -0.02 and 0.02
Positive Growth	*Revenue Change* over 0.02

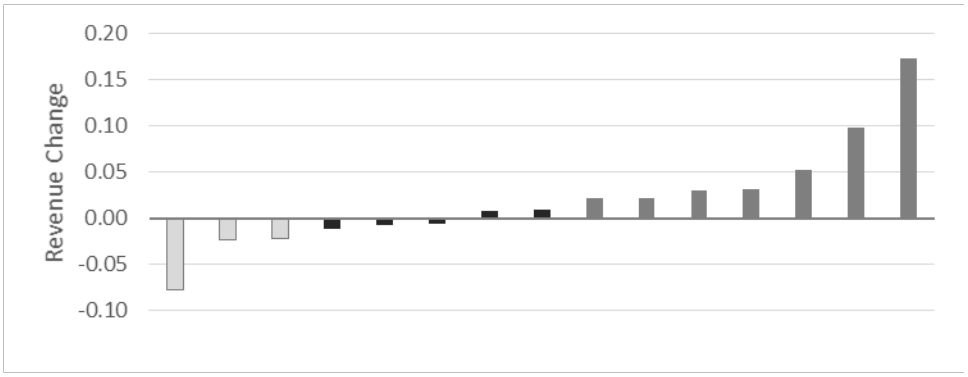

Figure 5.2.2. This figure illustrates the distribution of the *Revenue Change* measure across the fifteen stores associated with the carrier routes that participated in both stages of the study. Each column represents a single store. The colors of the bars distinguish between *Negative Growth*, *Flat Growth* and *Positive Growth* stores.

The *Revenue Change* measure is not a direct measure of concept shift. Instead it measures changes in underlying demand conditions, which is not the same as the response to the promotions. Our assumption is that changes in demand conditions contribute to concept shift, and so we use *Revenue Change* as an indicator of which carrier routes were most likely to experience concept shift. We investigate this assumption next.

5.2.3 Concept Shift and Changes in Underlying Demand Conditions

To support our claim that *Revenue Change* is associated with concept shift, we conducted a preliminary analysis using the three uniform mailing conditions. In particular, we repeated our earlier analysis of the *Lift in Profits* for the two promotions using the same 234 carrier routes. However, we now calculate the *Lift in Profits* separately for the 53 *Negative Growth* carrier routes, the 50 *Flat Growth* carrier routes, and the 131 *Positive Growth* carrier routes.

Changes in underlying demand conditions could affect the response to the three experimental conditions in unpredictable ways. Therefore, we do not make a specific prediction about how the *Lift in Profits* varies between Stages 1 and 2 across the different *Revenue Change* groups. Instead, we simply report the change in the *Lift in Profits* for each group. To do so we first calculate the difference in profit between the promotion condition and the no mail control (this is the *Lift in Profits*). We do this separately for the two promotion conditions ($25 paid and 120-day trial), and for the carrier routes in the different *Revenue Change* groups. We then calculate the change in the *Lift in Profits* between Stages 1 and 2 for each group. These differences are reported in Figure 5.2.3. For example, the figure reports that, compared to the no

[10] We also investigate using thresholds of -0.01 and 0.03 to balance the size of the groups. This yielded a similar pattern of results.

mail control, the $25 paid promotion generated a *Lift in Profits* that was $1.05 smaller in Stage 2 than in Stage 1 in carrier routes with negative *Revenue Growth.*[11]

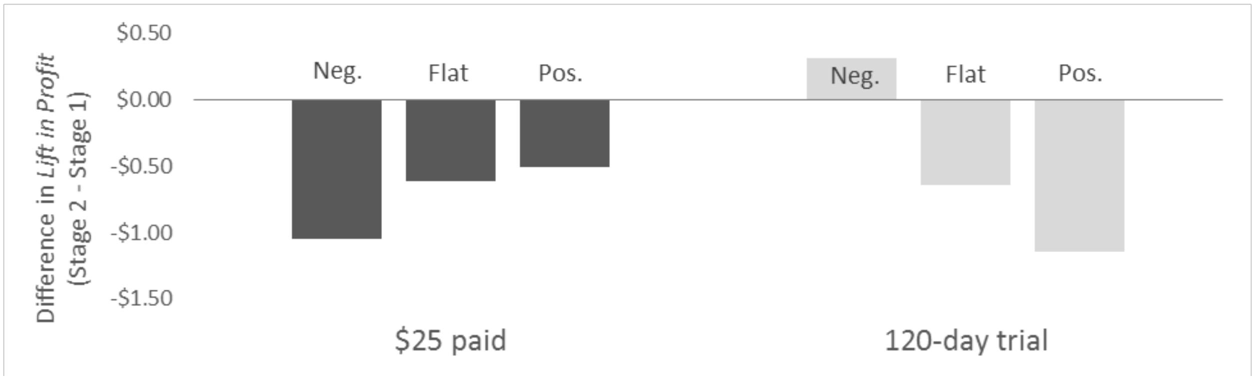

Figure 5.2.3 The figure summarizes the average change between Stages 1 and 2 in the *Lift in Profit* attributable to each promotion condition. The *Lift in Profit* is calculated as the average profit in the promotion condition minus the average profit in the no mail control condition. The change between Stages 1 and 2 is calculated as Stage 2 minus Stage 1. The results are reported separately for the $25 paid and 120-day trial promotions. The analysis uses the 234 carrier routes that received uniform treatments in both Stages 1 and 2. To protect confidentiality, we multiply the profits by a (common) random number.

We see clear evidence that concept shift impacted the profitability of the two promotions. The impact was different for the two types of promotions, and also varied systematically across the *Revenue Change* groups. To evaluate how this concept shift could affect the accuracy of Stage 2 decisions made using the Stage 1 outcomes, it is helpful to distinguish between three types of decisions: (a) whether to mail the $25 paid promotion versus not mail; (b) whether to mail the 120-day trial versus not mail, and (c) whether to mail the $25 paid versus the 120-day promotion.

For the first decision, the concept shift in Figure 5.2.3 means that using the Stage 1 data to make Stage 2 decisions is likely to lead to errors in all three *Revenue Change* groups. The $25-paid promotion generates a smaller profit lift in Stage 2 than in Stage 1, and so training on the Stage 1 data will tend to result in over-mailing the $25 paid promotion in Stage 2. This is particularly true for the carrier routes with negative *Revenue Growth*, because the change in *Lift in Profit* is particularly large in this group.

For the second decision, using the Stage 1 data to make Stage 2 decisions will lead to different types of errors in the different *Revenue Growth* groups. Training on the Stage 1 data will tend to result in under-mailing the 120-day trial promotion to negative *Revenue Growth* groups (though this effect appears small), and over-mailing in the flat and positive *Revenue Growth* groups. The errors are likely to be particularly large in the carrier routes with positive *Revenue Growth*, as the change in *Lift in Profit* is noticeably larger in that group than in the other groups.

The third decision focuses on “what to mail”, which depends solely upon the relative profits in the $25 paid and 120-day trial conditions. To isolate concept shift that affects this decision, we calculate the

[11] Recall that we multiply the profits by a random number.

difference in profits between the two promotion conditions. We then compare this profit difference in Stages 1 and 2 and report the findings in Figure 5.2.4.[12]

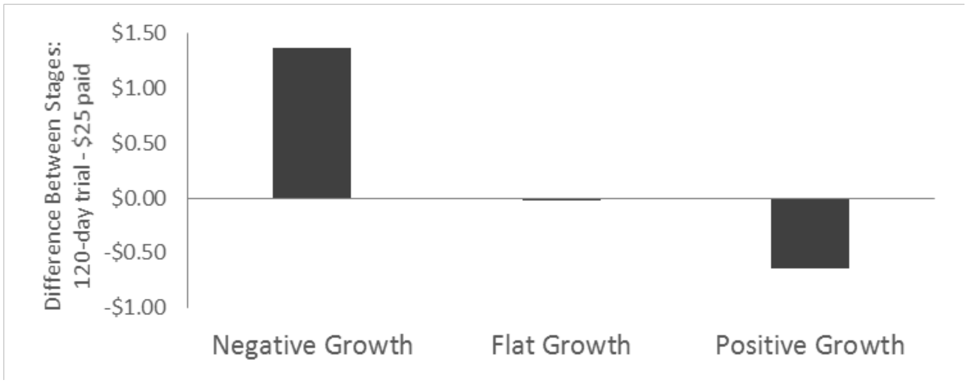

Figure 5.2.4 The figure reports how the difference between the *Profit* earned in the 120-day trial and $25 paid uniform policies changed between Stages 1 and 2. The difference in profits is calculated as the 120-day trial *Profit* minus the $25 paid *Profit*. The change between Stages 1 and 2 is calculated as Stage 2 minus Stage 1. The analysis uses the 234 carrier routes that received uniform treatments in both Stages 1 and 2. To protect confidentiality, we multiply the profits by a (common) random number.

The changes in the relatively profitability of the two promotions clearly vary across the three *Revenue Change* groups. In carrier routes with negative change in underlying demand, the 120-day trial promotions were relatively more profitable than the $25 paid promotions in Stage 2 compared to Stage 1. The reverse is true for carrier routes with positive demand change. Relying on the Stage 1 data to decide "what to mail" in Stage 2 could lead to errors in carrier routes with positive or negative *Revenue Change*.

In contrast, in carrier routes for which *Revenue Change* is flat, the relative profitability of the two promotions did not change between Stages 1 and 2. In these carrier routes, this type of concept shift did not occur, and so the Stage 1 data provides accurate information about which promotion to mail in Stage 2. This has an important implication for the performance of the seven optimized policies. The seven methods had more accurate information to select which promotion to mail in carrier routes for which *Revenue Change* was flat, compared to carrier routes in which *Revenue Change* was positive or negative.

We have reported evidence of concept shift, together with evidence that concept shift varies across the three *Revenue Change* groups. We next evaluate how the performance of the seven optimized targeting policies varied across the three *Revenue Change* groups.

5.2.4 Revenue Change and the Performance of the Seven Optimized Policies

In Figure 5.2.5 we compare how well the seven methods performed when *Revenue Change* was negative or positive, compared to when it was flat. We report the Stage 2 performance of the different targeting methods within each *Revenue Change* group. The profit is indexed at 100 using the most profitable uniform policy within each group. This indexing reveals how well the methods performed compared to the most profitable naïve benchmark.

[12] This is equivalent to calculating the difference between the *Profit Lift: $25 paid* and *Profit Lift: 120-day* columns in Figure 5.2.3.

Concept shift anticipates that the performance of optimized policies will deteriorate if the underlying response function changes. One reason that the response function may change is because of changes in the underlying demand conditions. Our *Revenue Change* measure provides a measure of demand changes, and so concept shift would predict that the methods will perform best when *Revenue Change* is flat, but performance will deteriorate when *Revenue Change* is negative or positive. This is precisely what we see in Figure 5.2.5. Both positive and negative demand changes are associated with lower performance than flat demand growth. This is true for all three types of targeting methods.

The poor performance of the targeted policies when *Revenue Change* is positive or negative cannot be attributed to lower value customers in these carrier routes. Recall that the findings in Figure 5.2.5 are indexed against the most profitable uniform policy within each group of carrier routes. This controls for differences in the absolute value of the customers in each group.

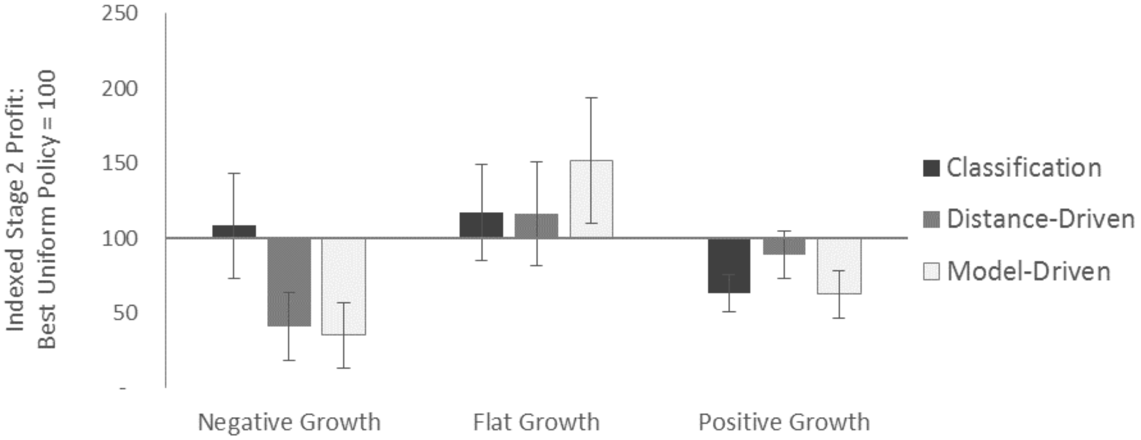

Figure 5.2.5. This figure illustrates the average Stage 2 *Profit* earned from carrier routes that participated in both stages of the study. The profits are grouped using our taxonomy of methods. *Profit* is indexed at 100 in the optimal uniform policy (for that *Revenue Change* group). The error bars indicate 95% confidence intervals.

The deterioration in performance appears to be larger for the model-driven methods than for the other methods. In the Appendix we formally compare the deterioration in the performance across methods. The findings confirm that profits are lower in carrier routes for which revenue among existing customers changed between the two stages. This is true for either positive or negative *Revenue Change*. Moreover, the deterioration in profit between the *Flat Growth* and *Positive Growth* carrier routes is larger for model-driven methods than distance-driven methods. We also see that model-driven methods deteriorate more than classification methods when *Revenue Change* is negative (compared to when it is flat).

The evidence that the performance of the model-driven methods is more sensitive to demand changes than the other methods is consistent with the covariate shift results reported in the previous sub-section (5.1). In the absence of covariate shift or concept shift, the model-driven methods yield the largest improvements over the naïve benchmarks. However, as the quality of the data deteriorates due to either covariate shift or concept shift, the advantages of using the model-driven methods quickly disappear. In the next set of results in this section, we analyze how the methods performed when the quality of the training data deteriorates due to aggregation.

5.3 Aggregation of the Targeting Variables

When targeting existing customers, firms can use past purchasing history to make targeting decisions. In the absence of individual purchasing histories, firms generally rely upon demographic variables, which are typically aggregated across census blocks, zip codes or carrier routes. This feature is an important difference between targeting prospective customers and targeting existing customers. In this section we investigate how this aggregation affects the performance of the targeting methods.

Because the number of households in a carrier route varies across carrier routes, the degree of aggregation also varies. For households in carrier routes with relatively few other households (i.e., small carrier routes), the targeting variables contain more information about the households than in carrier routes with many households. To investigate how the precision of the information affected the outcomes, we calculated the findings separately using a median split of the number of households in each carrier route. The findings are summarized in Figure 5.3.1 where we report findings for each of the ten treatments, and in Figure 5.3.2 where we group the methods using our taxonomy.

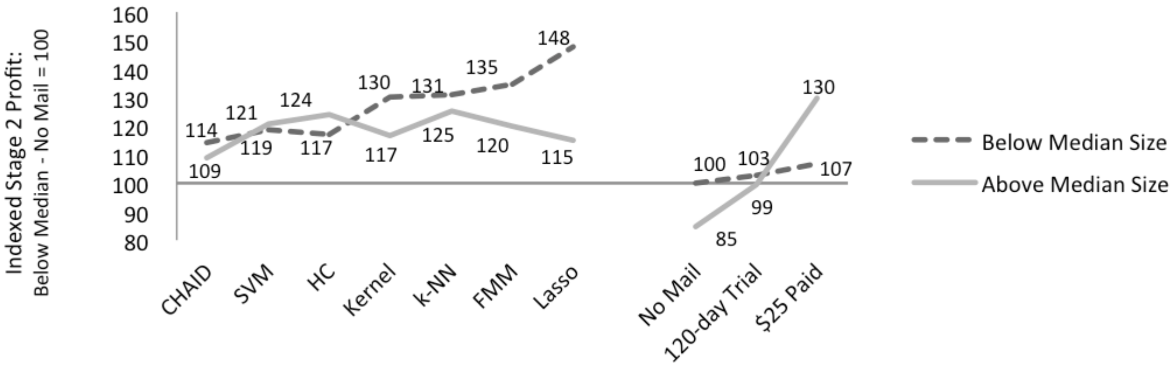

Figure 5.3.1. This figure illustrates the average profits earned in Stage 2 treatment condition when grouping the carrier routes according to whether they contain more or less than the median number of households. The profits are indexed to 100 in the No Mail - Below Median Size data point. Complete findings are reported in the Appendix.

In the smaller carrier routes, where the targeting variables are more informative, the two model-driven targeting methods (Lasso and FMM) perform better than the other methods. This improvement is particularly clear when grouping the methods using our taxonomy of methods. This suggests that the model-driven methods make the best use of the increased precision of the information in smaller carrier routes. Notice that this finding cannot be explained simply by smaller carrier routes being more profitable. We would expect this to benefit all methods.

However, in the larger carrier routes, where the demographic variables contain less precise information about each household, all targeting methods perform similarly. Notably, none of them out-perform the benchmark of uniformly mailing the $25 paid promotion to every household. The implication is that when the carrier routes are large, the targeting variables provide so little information that none of the targeting methods can improve upon a naïve benchmark. Moreover, this deterioration in performance when there is more aggregation is more pronounced for the model-driven methods than for the other methods.

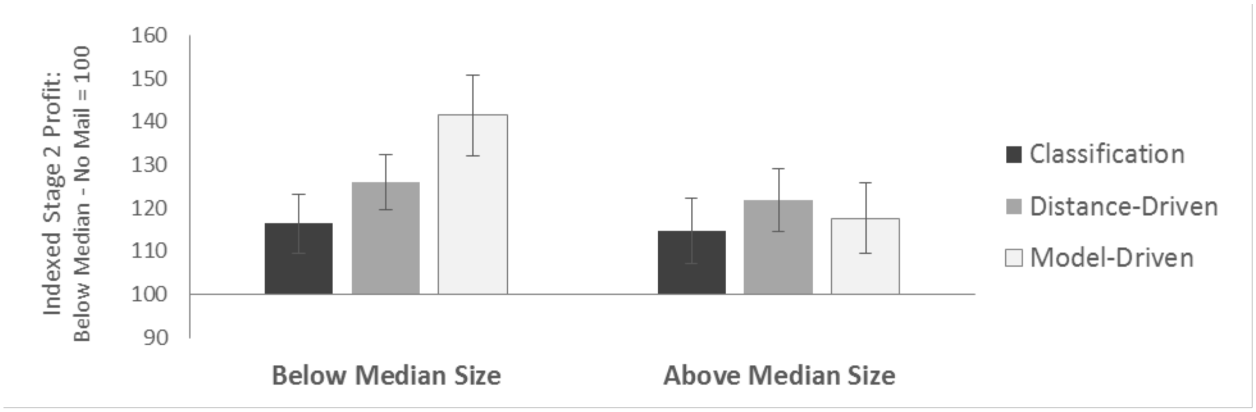

Figure 5.3.2. This figure illustrates the average Stage 2 *Profit* when grouping the carrier routes according to whether they contain more or less than the median number of households. The profits are indexed to 100 in the No Mail - Below Median Size data point. The error bars indicate 95% confidence intervals. Complete findings are reported in the Appendix.

While we have interpreted the evidence that the model-driven methods perform better in smaller carrier routes as evidence that they make better use of more precise information, this is not the only possible interpretation. There may be other features of smaller carrier routes that distinguish them from larger carrier routes, and which also contribute to the improvement in the performance of the model-driven methods. However, the findings throughout this section have revealed a consistent pattern; model-driven methods perform better, but only when the training data provides accurate information about the validation data. As the quality of the information deteriorates, we consistently see that the performance of the model-driven methods deteriorates faster than the performance of the other methods. We interpret this pattern as evidence that the model-driven methods make better use of the available information, but are also more likely to be misled when the quality of this information deteriorates. This pattern survives when we vary the "quality of the information" along a variety of different dimensions.

5.4 Why did the Classifiers Perform So Poorly?

A common thread in our findings is that the classification methods (CHAID and SVM) performed poorly compared to the distance- and model-driven (regression-based) methods. Understanding the reasons for this poor performance will help us evaluate the extent to which this result generalizes to other settings. In this section we show that the poor performance of the classification methods is at least partly attributable to loss of information. We also discuss modifications to the classification methods that have been proposed in the machine learning literature to address this limitation.

To help understand why the classification methods performed poorly in our study, we start by considering the following simplified setting. Suppose the training set consists of N carrier routes, each described by targeting variables $\mathbf{x}_i \in \mathbb{R}^p$. For each carrier routes i in the training set, we know the profit outcomes for the *no mail* and *mail* treatments, $\left(y_i^{no\ mail}, y_i^{mail}\right)$. We define a matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ and vectors $\boldsymbol{y}^{no\ mail} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{y}^{mail} \in \mathbb{R}^{N\times 1}$ to capture the targeting variables and the profit outcomes of all customers in the training data. We also have a new carrier route with targeting variables $\mathbf{x}_{new}$ for which we want to select an optimal treatment.

In the model-driven and distance-driven (regression-based) methods, we use $(\mathbf{X}, \boldsymbol{y}^{no\ mail}, \boldsymbol{y}^{mail})$ to train two models that predict $\left(\hat{y}_i^{no\ mail}, \hat{y}_i^{mail}\right)$ given $\mathbf{x}_i$. We then predict $(\hat{y}_{new}^{no\ mail}, \hat{y}_{new}^{mail})$ for $\mathbf{x}_{new}$ and assign the treatment that yields the largest predicted profit.

In the classification methods, we first transform $(\boldsymbol{y}^{no\ mail}, \boldsymbol{y}^{mail}) \rightarrow \boldsymbol{t}^*$, where $t_i^* = argmax_t\ (y_i^t)$ is the optimal treatment *in hindsight* for carrier route i in the training set. We then use $(\mathbf{X}, \boldsymbol{t}^*)$ to train a classifier that predicts $\widehat{t_i}$ given $\mathbf{x}_i$ and assign to a new observation the treatment that the classifier predicts is optimal: $t_{new}^* = \hat{t}(\mathbf{x}_{new})$.

The explanation for why the classification methods perform poorly focuses on the transformation $(\boldsymbol{y}^{no\ mail}, \boldsymbol{y}^{mail}) \rightarrow \boldsymbol{t}^*$ and on the training of the prediction function $\hat{t}(\cdot)$.

We illustrate the explanation using an example. Assume that the carrier routes are identical, so that they are not distinguishable by targeting variables $\mathbf{X}$. If a household responds to a promotional mailing, the firm earns $1,000 after deducting costs. If the household does not respond, the firm loses $1 representing mailing and printing costs. Not mailing yields a payoff of zero with certainty.

Consider first a model with a 100% response rate. If all households that were mailed respond to the mailing, the firm earns $1,000 from each of these customers and zero profit from the customers who were not mailed. Regression-based methods will predict that mailing yields an expected profit of $1,000 per customer, and that not mailing yields an expected profit of zero per customer. The regression-based methods will therefore correctly recognize that mailing is more profitable than not mailing in every carrier route.

Classification methods will also recognize that *in hindsight* mailing is on average more profitable than not mailing in every carrier route. The transformation $(\boldsymbol{y}^{no\ mail}, \boldsymbol{y}^{mail}) \rightarrow \boldsymbol{t}^*$ will result in every carrier route in the training data receiving a label indicating mailing was the most profitable policy ($\boldsymbol{t}_i^* = mail$). Consequently, the classification methods will also correctly recommend mailing to every carrier route. We conclude that with a 100% response rate, both types of methods will recommend the optimal policy.

Now assume that the probability a household (in any carrier route) responds to a mailing is just 1%. Assume further that every carrier route has 100 households, and in the training data a randomly selected sub-sample of 50 households receives promotional mailings. The remaining households are not mailed. Regression-based methods will predict that mailing yields an expected profit of $9.01 per household ($1,000*0.01 -$1*0.99) and correctly recognize that mailing is more profitable than not mailing.

In contrast, classification methods will compare the profits earned from mailing and not mailing in each carrier route. If at least one customer responds, the total profits exceed the mailing costs, in which case the carrier route receives a label indicating mailing is the most profitable policy ($t_i^* = mail$). In contrast, if no customer responds, the label indicates not mailing is more profitable in that carrier route ($t_i^* = no\ mail$). With 50 customers receiving mailings and a 1% probability of a response, the probability there is at least one response is 39.5% (i.e., $1 - 0.99^{50}$), assuming independence. In expectation, 39.5% of the carrier routes receive a label *mail* and the remaining 60.5% receive a label *no mail*. Because the carrier routes are not distinguishable by $\mathbf{X}$, and not mailing is optimal more frequently than mailing, the classification methods will incorrectly recommend not to mail.

This error reflects a loss of information due to the transformation $\left(\boldsymbol{y}^{\text{no mail}}, \boldsymbol{y}^{\text{mail}}\right) \rightarrow \boldsymbol{t}^*$. The transformation recognizes which treatment performed best for each carrier route in the training data, but discards information about the magnitude with which it outperformed the other treatments.

The machine learning literature has recognized this limitation and characterized when it is most likely to lead to sub-optimal policies. Errors are likely when the data is "imbalanced" (i.e., the two labeled classes have different sizes in the training set), and the cost of errors is asymmetric. Asymmetric costs arise when the cost of a false positive (mailing when there is no response) is very different than the cost of a false negative (not mailing when there would be a response).

Data imbalance and asymmetric costs are clearly illustrated in our example. Recall that if 100% of the customers responded to the promotion mailing, the classification methods recommend the correct policy. They only make an error when the response rate is so low that the majority of carrier routes do not receive a single response. Notice also that if the profit when a customer responds to a mailing were identical to the profit lost when a customer does not respond to a mailing, then the classification methods would also perform well. It is the difference in these outcomes that makes the costs "asymmetric".

Targeting prospective customers with direct mail yields a low average response rate, but a large profit when customers do respond. This makes the imbalanced data with asymmetric costs issue particularly relevant in our setting. This is not true for all marketing problems. For example, when targeting existing customers, average response rates tend to be a lot higher than for prospects,[13] while using a sales force to target customers (instead of direct mail) can also increase response rates. In both examples we would expect data that is a lot less imbalanced. Moreover, for existing customers, the difference in long-run profits from customers who respond to a promotion, versus those who do not, may be relatively small. If the response to a promotion results in temporal substitution from future demand, the response to the promotion may result in little variation in long-run profits.[14] In these examples the cost of errors is more symmetric.

Our illustrative example can also be used to highlight an additional characteristic that makes targeting prospective customers even more challenging for classification methods. The labeling of the optimal policy in each carrier route is made based on a relatively small sample of households in each carrier route. This introduces a finite sample problem, which is not just a source of variance, but also a source of bias. We can illustrate this bias by assuming that instead of 100 households per carrier route with 50 receiving the mailings in the training data, there are 200 households per carrier route and 100 receive the mailings. Mailing is still profitable in a carrier route if at least one household responds to the mailing. With 100 households receiving mailings, the probability that at least one household will respond in a carrier route is 63.4% (i.e., $1 – 0.99^{100}$), assuming independence. As a result, we would expect that over half of the carrier routes (63.4%) would receive a label indicating "mail" was the most profitable action, and the classification methods will now correctly recommend mailing to every carrier route. This example confirms that the size of the training data does not just contribute to variance in the recommended policy; it can also introduce bias by changing the expected policy. The machine learning literature refers to this as an "imprecise label" problem (Frénay and Verleysen, 2014).

These arguments are reflected in the outcome of our study. To illustrate the finite sample size bias, we counted the number of households that received the $25 paid promotion in each Stage 1 carrier route. We then divide the Stage 1 carrier routes into two sub-samples using a median split of this count (for ease of exposition we will label them the "large" and "small" carrier routes). Within each sub-sample we

[13] For example, the DMA reports that average direct marketing response rates are approximately three times higher when targeting existing customers compared to prospective customers (DMA 2005).

[14] Temporal demand substitution in response to promotions is well studied in marketing (see for example Krishna, 1992, 1994; Thompson and Noordewier, 1992). One dramatic example occurred in 2005 when the three US domestic automobile manufacturers all offered a promotion in which customers could buy at the "employee prices". This resulted in record increase in sales at the time of the promotion, but this simply pulled demand forward, so that there was almost no change in any of the firms' annual sales (Busse et al., 2010).

compare how often the promotion was more profitable than the no mail control. Among the large carrier routes, the profits earned in the \$25 paid treatment exceeded the no mail control in 32.9% of the carrier routes. However, among the small carrier routes, just 24.9% of carrier routes had higher profits in the \$25 paid treatment than in the control. The difference in these proportions is highly significant ($p < 0.0001$), and is replicated when we conduct the same analysis for the 120-day free trial.[15] For large carrier routes it appears to the classification methods that mailing is profitable more frequently than in small carrier routes. However, this is an artifact of the transformation $\left(\boldsymbol{y}^{\text{no mail}}, \boldsymbol{y}^{\$25}, \boldsymbol{y}^{120-\text{days}}\right) \rightarrow \boldsymbol{t}^{*}$, and will occur even if the optimal policy is the same in large and small carrier routes.

This bias leads to sub-optimal policies. As the no mail control condition is more profitable (in hindsight) in most of the carrier routes in the training data, the classification methods recommend not mailing to too many carrier routes. This is precisely what we see. SVM and CHAID recommend not mailing either promotion to 82.2% and 66.3% of the Stage 2 carrier routes respectively. In contrast, the model-driven and distance-driven (regression-based) methods choose not to mail to 25.3% and 22.3% of carrier routes respectively.

The machine learning literature has proposed two approaches to address these problems. One class of methods involves pre-processing the training data (these are sometimes call *external* solutions), while the second approach involves adjusting the decision rule (these solutions are sometimes called *internal* solutions).

The external solutions involve sampling from the universe of data in the training data. Many different sampling methods have been proposed (see He and Ma, 2013 for a collection of papers describing different re-sampling methods). For example, Batuwita and Palade (2010) study SVM and recommend first estimating the prediction function $\hat{t}(\cdot)$ using all of the data in the training sample. This original estimate is then used to select the data points (carrier routes) lying close to the hyperplane that separates the different classes of observations (in $\mathbf{X}$-space). By oversampling in this region, and re-estimating the prediction function, the data imbalance issue is addressed and performance may be improved.

Internal methods incorporate the asymmetric costs of errors into the policy design (see for example Eitrich and Lang, 2006). In particular, if the cost of a false negative is higher than the cost of a false positive, then the decision rule can be adjusted to give more weight to avoiding false negatives (or vice versa). In our example, the cost of a false negative (not mailing to a carrier route that would yield one or more responses) is much higher than the cost of a false positive (mailing to a carrier route that does not yield any responses). The decision rule could be adjusted to account for this by recommending mailing to all carrier routes whenever at least a third of the carrier routes in the training data receive the label *mail* (instead of at least a half of the carrier routes). In practice, this adjustment is made by adjusting the prediction function $\hat{t}(\cdot)$, which shifts the separating hyperplane distinguishing the different classes of observations. Our illustration highlights that the cost adjustment could also potentially account for the bias introduced by the imprecise labels problem.

As we discussed, asymmetric costs, imbalanced data and label imprecision are very relevant when prospecting for new customers, and are also likely to arise in other customer targeting problems. Although there is now an extensive machine learning literature that proposes solutions to these problems,

[15] For the 120-day free trial, among the larger carrier routes (for which the number receiving the 120-day promotion was above the median) the promotion was more profitable than the no mail control in 29.6% of the carrier routes, compared to 24.0% in the smaller carrier routes. This difference is again highly significant.

the problems and the proposed solutions have received little attention in marketing.[16] The proposed modifications are likely to improve the performance of the classification methods, but it is unlikely that a moderately sophisticated retailer would implement these solutions. We believe that the findings we report are representative of the performance that many retailers will obtain when using these classification methods. Hopefully, our findings will help to highlight the practical importance of these issues to both marketing academics and practitioners.

6. Conclusions

As more firms embrace field experiments to optimize their marketing activities, the focus on how to derive more value from field experiments is likely to intensify. One way to derive additional value is to segment customers and evaluate how to target different customers with different marketing treatments. Fortunately the literature on customer segmentation and targeting methods is vast, and so firms have a broad range of methods to choose from. However, the literature offers little guidance as to which of these methods are most effective in practice.

In this paper, we have evaluated seven widely used segmentation methods using a series of two large-scale field experiments. The first field experiment is used to generate a common pool of training data for each of the seven methods. We then validate the optimized policies provided by each method in a second field experiment. Our detailed comparison of the methods reveals an important general finding. Model-driven methods perform better than distance-driven and classification methods when the data is ideal. However, they also deteriorate faster in the presence of covariate shift, concept shift and information loss through aggregation. Intuitively, the model-driven regression methods make the best use of the available information, but the performance of these methods is more sensitive to deterioration in the quality of the information.

We also investigated why the model-driven and distance-driven methods outperformed the classification methods in our setting. Classification methods perform poorly when there is a low probability of a response, but each response is very profitable. We describe modifications to the classification methods that have been proposed in the machine learning literature to address these challenges.

As with most comparisons of methods, there are two important limitations to these findings. First, each of the methods that we implemented is representative of a general class of related methods. It is obviously not possible to test every version of every class of methods. While we chose what we believe to be a commonly implemented version of each class of methods (by moderately sophisticated retailers), we recognize that other versions may perform differently.

A second limitation is that the findings may not generalize to every market setting. In particular, the experiments in this study involved prospecting for new customers. When prospecting for new customers, firms generally have a lot less information than if they are targeting existing customers. With existing customers, firms can often observe each customer’s past purchasing from the firm, which provides a rich source of information for predicting future responses. The performance of the segmentation methods may be very different if they have access to this type of information. In future research we hope to extend our results to existing customers and to other marketing decisions.

[16] Although not in a marketing journal, Kim, Chae and Olson (2012) recognize the data imbalance problem, and use random under-sampling methods to reduce the degree of imbalance.

References

Abhishek, Vibhanshu, Kartik Hosanagar, and Peter S. Fader. "Aggregation bias in sponsored search data: The curse and the cure." *Marketing Science* 34, no. 1 (2015): 59-77.

Aigner, D.J. and S.M. Goldfeld, 1974, "Estimation and prediction from aggregate data when aggregates are measured more accurately than their components". *Econometrica*, 42 (January), 113-134.

Aigner. D.J. and S.M. Goldfeld, 1973, "Simulation and aggregation: a reconsideration". *The Review of Economics and Statistics*, 55 (February), 114-118.

Alaiz-Rodríguez, Rocío, and Nathalie Japkowicz. "Assessing the impact of changing environments on classifier performance." In *Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 13-24. Springer, Berlin, Heidelberg, 2008.

Andrews, Rick L., Imran S. Currim, and Peter SH Leeflang. "A comparison of sales response predictions from demand models applied to store-level versus panel data." *Journal of Business & Economic Statistics* 29, no. 2 (2011): 319-326.

Auer, P., and M. K. Warmuth. "Tracking the best disjunction." *Machine Learning* 32(2), 1998: 127-150.

Baldi, Pierre, and Søren Brunak. *Bioinformatics: the machine learning approach*. MIT press, 2001.

Batuwita, Rukshan, and Vasile Palade. "Efficient resampling methods for training support vector machines with imbalanced datasets." In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1-8. IEEE, 2010.

Bellet, Aurélien, Amaury Habrard, and Marc Sebban. "A survey on metric learning for feature vectors and structured data." *arXiv preprint arXiv:1306.6709* (2013).

Biau, Gérard, and Laszlo Gyorfi. "On the asymptotic properties of a nonparametric l/sub 1/-test statistic of homogeneity." *IEEE Transactions on Information Theory* 51, no. 11 (2005): 3965-3973.

Bickel, Steffen, and Tobias Scheffer. "Dirichlet-enhanced spam filtering based on biased samples." In *Advances in neural information processing systems*, pp. 161-168. 2007.

Borgwardt, Karsten M., Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. "Integrating structured biological data by kernel maximum mean discrepancy." *Bioinformatics* 22, no. 14 (2006): e49-e57.

Busse, Meghan R., Duncan I. Simester, Florian Zettelmeyer. 2010. "The Best Price You'll Ever Get": The 2005 Employee Discount Pricing Promotions in the U.S. Automobile Industry. *Marketing Sci.* 29(2):268-290.

Caputo, Barbara, K. Sim, F. Furesjo, and Alex Smola. 2002. "Appearance-based object recognition using SVMs: which kernel should I use?." In *Proc of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision*, Whistler, vol. 2002.

Castillo, Gladys, João Gama, and Ana M. Breda. "Adaptive Bayes for a student modeling prediction task based on learning styles." In *International Conference on User Modeling*, pp. 328-332. Springer, Berlin, Heidelberg, 2003.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. "Double/Debiased/Neyman machine learning of treatment effects." *American Economic Review* 107, no. 5 (2017): 261-65.

Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial intelligence in medicine* 34, no. 2 (2005): 113-127.

Dietterich, Thomas G., Gerhard Widmer, and Miroslav Kubat. "Special issue on context sensitivity and concept drift." *Machine Learning 32*, no. 2 (1998).

DMA. (2005). *Statistical Fact Book*. Direct Marketing Association, New York NY.

Edwards, J.B. and G.H. Orcutt, 1969, "Should estimation prior to aggregation be the rule?" *The Review of Economics and Statistics*, 51 (November), 409-420.

Eitrich, Tatjana, and Bruno Lang. "Efficient optimization of support vector machine learning parameters for unbalanced datasets." *Journal of computational and applied mathematics*196, no. 2 (2006): 425-436.

Erdem, T; Keane, MP 1996. “Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets,” *Marketing Science*, 15(1), 1-20.

Foekens, Eijte W., Peter SH Leeflang, and Dick R. Wittink. "A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation." *International Journal of Forecasting* 10, no. 2 (1994): 245-261.

Frénay, Benoît, and Michel Verleysen. "Classification in the presence of label noise: a survey." *IEEE transactions on neural networks and learning systems* 25, no. 5 (2014): 845-869.

Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." *The Annals of Statistics* (1979): 697-717.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.

Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. "A survey on concept drift adaptation." *ACM Computing Surveys (CSUR)* 46, no. 4 (2014): 44.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test." *Journal of Machine Learning Research* 13, no. Mar (2012): 723-773.

Grunfeld, Y. and Z. Griliches, 1960, “Is aggregation necessarily bad?” *The Review of Economics and Statistics*, 42 (February), 1-13.

Gupta, Sachin, Pradeep Chintagunta, Anil Kaul, and Dick R. Wittink. "Do household scanner data provide representative inferences from brand choices: A comparison with store data." *Journal of Marketing Research* (1996): 383-398.

Hall, Peter, and Nader Tajvidi. "Permutation tests for equality of distributions in high-dimensional settings." *Biometrika* 89, no. 2 (2002): 359-374.

Hand, David J. "Classifier technology and the illusion of progress." *Statistical science* (2006): 1-14.

Hand, David J. "Rejoinder: Classifier Technology and the Illusion of Progress. " *Statistical science* 21 (2006b): 30-34.

Harries, Michael Bonnell, Claude Sammut, and Kim Horn. "Extracting hidden context." *Machine learning* 32, no. 2 (1998): 101-126.

Harries, Michael, and Kim Horn. "Detecting concept drift in financial time series prediction using symbolic machine learning." In *AI-Conference*, pp. 91-98. World Scientific Publishing, 1995.

He, Haibo, and Yunqian Ma, eds. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

Heckman, James J. "Sample Selection Bias as a Specification Error." *Econometrica* 47, no. 1 (1979): 153-61.

Helmbold, David P., and Philip M. Long. "Tracking drifting concepts using random examples." In *Proceedings of the fourth annual workshop on Computational learning theory*, pp. 13-23. Morgan Kaufmann Publishers Inc., 1991.

Helmbold, David P., and Philip M. Long. "Tracking drifting concepts by minimizing disagreements." *Machine learning* 14, no. 1 (1994): 27-45.

Herbster, Mark, and Manfred K. Warmuth. "Tracking the best expert." *Machine learning* 32, no. 2 (1998): 151-178.

Indyk, Piotr, and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality." In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604-613. ACM, 1998.

Jain, Anil, and Douglas Zongker. "Feature selection: Evaluation, application, and small sample performance." *IEEE transactions on pattern analysis and machine intelligence* 19, no. 2 (1997): 153-158.

Kim, Gitae, Bongsug Kevin Chae and David L. Olson. 2013. "A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models." *Service Business*, 7(1): 167–182.

Kim, Byung-Do, and Peter E. Rossi. "Purchase frequency, sample selection, and price sensitivity: The heavy-user bias." *Marketing Letters* 5, no. 1 (1994): 57-67.

Klinkenberg, Ralf. "Predicting phases in business cycles under concept drift." In *Proc. of LLWA*, pp. 3-10. 2003.

Krishna, Aradhna. 1992. The normative impact of consumer price expectations for multiple brands on consumer purchase behavior. *Marketing Sci.* 11(3) 266–286.

Krishna, Aradhna. 1994. The impact of dealing patterns on purchase behavior. *Marketing Sci.* 13(4) 351–373.

Kuh, Anthony, Thomas Petsche, and Ronald L. Rivest. "Incrementally learning time-varying half-planes." In *Advances in Neural Information Processing Systems*, pp. 920-927. 1992.

Kuh, Anthony, Thomas Petsche, and Ronald L. Rivest. "Learning time-varying concepts." In *Advances in Neural Information Processing Systems*, pp. 183-189. 1991.

Kukar, Matjaž. "Drifting concepts as hidden factors in clinical studies." In *Conference on Artificial Intelligence in Medicine in Europe*, pp. 355-364. Springer, Berlin, Heidelberg, 2003.

Lane, Terran, and Carla E. Brodley. "Approaches to Online Learning and Concept Drift for User Identification in Computer Security." In *KDD*, pp. 259-263. 1998.

Li, Yujia, Kevin Swersky, and Rich Zemel. "Generative moment matching networks." In *International Conference on Machine Learning*, pp. 1718-1727. 2015.

McCarty, John A., and Manoj Hastak. "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression." *Journal of business research* 60, no. 6 (2007): 656-662.

Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. "A unifying view on dataset shift in classification." *Pattern Recognition* 45, no. 1 (2012): 521-530.

Movellan, Javier R., and Paul Mineiro. "Robust sensor fusion: Analysis and application to audio visual speech recognition." *Machine Learning* 32, no. 2 (1998): 85-100.

Naik, Prasad A., Murali K. Mantrala, and Alan G. Sawyer. "Planning media schedules in the presence of dynamic advertising quality." *Marketing science* 17, no. 3 (1998): 214-235.

Olson, David L., Qing Cao, Ching Gu, and Donhee Lee. "Comparison of customer response models." *Service Business*3, no. 2 (2009): 117-130.

Pechenizkiy, Mykola, Jorn Bakker, I. Žliobaitė, Andriy Ivannikov, and Tommi Kärkkäinen. "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift." *ACM SIGKDD Explorations Newsletter* 11, no. 2 (2010): 109-116.

Penny, Kay I., and Thomas Chesney. "Imputation methods to deal with missing values when data mining trauma injury data." In *Information Technology Interfaces, 2006. 28th International Conference on*, pp. 213-218. IEEE, 2006.

Pesaran, M. Hashem, Richard G. Pierse, and Mohan S. Kumar. "Econometric analysis of aggregation in the context of linear prediction models." *Econometrica: Journal of the Econometric Society* (1989): 861-888.

Schlimmer, Jeffrey C., and Richard H. Granger. "Incremental learning from noisy data." *Machine learning* 1, no. 3 (1986): 317-354.

Sejdinovic, Dino, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing." *The Annals of Statistics*(2013): 2263-2291.

Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference* 90, no. 2 (2000): 227-244.

Smirnov, Nikolai V. "On the estimation of the discrepancy between empirical curves of distribution for two independent samples." *Bull. Math. Univ. Moscou* 2, no. 2 (1939): 3-14.

Stock, James H., and Mark W. Watson. "An empirical comparison of methods for forecasting using many predictors." *Manuscript, Princeton University* (2005).

Sugiyama, Masashi, and Motoaki Kawanabe. 2012. *Machine Learning in Non-Stationary Environments*, Cambridge: MIT Press.

Sugiyama, Masashi, Matthias Krauledat, and Klaus-Robert Müller. "Covariate shift adaptation by importance weighted cross validation." *Journal of Machine Learning Research* 8, no. May (2007): 985-1005.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1, no. 1. Cambridge: MIT press, 1998.

Tapak, Lily, Hossein Mahjub, Omid Hamidi, and Jalal Poorolajal. "Real-data comparison of data mining methods in prediction of diabetes in Iran." *Healthcare informatics research*19, no. 3 (2013): 177-185.

Thompson, Patrick A., Thomas Noordewier. 1992. Estimating the effects of consumer incentive programs on domestic automobile sales. *J. Bus. Econom. Statist.* 10(4) 409–417.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

Tong, Liping, Cole Erdmann, Marina Daldalian, Jing Li, and Tina Esposito. "Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk." *BMC medical research methodology* 16, no. 1 (2016): 26.

Tsymbal, Alexey, Mykola Pechenizkiy, Padraig Cunningham, and Seppo Puuronen. "Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections." In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pp. 679-684. IEEE, 2006.

Turhan, Burak. "On the dataset shift problem in software engineering prediction models." *Empirical Software Engineering* 17, no. 1-2 (2012): 62-74.

Ueki, K., M. Sugiyama and Y. Ihara. “Perceived age estimation under lighting condition change by covariate shift adaption. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 897-904, 2008.

Vicente, Renato, Osame Kinouchi, and Nestor Caticha. "Statistical mechanics of online learning of drifting concepts: A variational approach." *Machine Learning* 32, no. 2 (1998): 179-201.

Wang, Haixun, Wei Fan, Philip S. Yu, and Jiawei Han. "Mining concept-drifting data streams using ensemble classifiers." In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 226-235. AcM, 2003.

Widmer, Gerhard and Miroslav Kubat. “Guest editors’ introduction.” *Machine Learning* 32, 83–84. 1998.

Widmer, Gerhard, and Miroslav Kubat. "Learning in the presence of concept drift and hidden contexts." *Machine learning* 23, no. 1 (1996): 69-101.

Wolpaw, Jonathan R., Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. "Brain–computer interfaces for communication and control." *Clinical neurophysiology* 113, 6 (2002): 767-791.

Wu, Jia, Sunil Vadera, Karl Dayson, Diane Burridge, and Ian Clough. "A comparison of data mining methods in microfinance." In *2010 2nd IEEE International Conference on Information and Financial Engineering (ICIFE)*, pp. 499-502. IEEE, 2010.

Xiang, Shiming, Feiping Nie, and Changshui Zhang. "Learning a Mahalanobis distance metric for data clustering and classification." *Pattern Recognition* 41, no. 12 (2008): 3600-3612.

Yamazaki, Keisuke, Motoaki Kawanabe, Sumio Watanabe, Masashi Sugiyama, and Klaus-Robert Müller. "Asymptotic bayesian generalization error when training and test distributions are different." In *Proceedings of the 24th international conference on Machine learning*, pp. 1079-1086. ACM, 2007.

Yang, L., and R. Jin. "Distance metric learning: A comprehensive survey." *Michigan State Univ.* 2(2) (2006).

Zadrozny, Bianca. "Learning and evaluating classifiers under sample selection bias." In *Proceedings of the twenty-first international conference on Machine learning*, p. 114. ACM, 2004.