Estimating the Causal Effect of Marketing Campaigns Using Doubly Robust and Tree-Based Methods

Susan Cherry
Final Project
STA 571: Advanced Machine Learning
Duke University

Abstract

This project uses the doubly robust and tree-based methods to estimate the causal effect of a bank marketing campaign. Previous approaches for estimating the causal effect of a policy, such as a marketing campaign, include the naive method and inverse propensity weighting which suffer for large bias and variance. This project compares estimates from all four methods and finds that the naive and inverse propensity weighting methods offer the best results. These methods are robust to misspecification and imply that it is optimal to contact a customer is 3 to 4 times. My results provide further evidence that the doubly robust and tree-based methods perform well on advertising data. Future work should be done to incorporate bagging into the tree-based method and to test these methods on larger, more complex datasets.

1 Introduction

1.1 Motivation and Introduction to Casual Inference

In an ideal world, marketers would run randomized experiments to understand how advertising campaigns effect consumer behavior. Unfortunately, randomized experiments are often expensive and difficult to implement, so causal methods are used to measure the effectiveness of advertising. This project determines the unbiased causal effect of a Portuguese bank's direct marketing campaign using doubly robust and tree-based methods.

The casual effect is measured by comparing the treatment and control groups. A naive comparison of outcomes between these two groups is biased, since treatment is highly correlated with user features such as demographic information. Instead, causal methods model treatment as a random variable that depends on a set of pre-treatment covariates [7]. These methods attempt to balance the features of the two different groups and rely on the following assumptions: 1) The outcome of one consumer is unaffected by the assigned treatment of other consumers. 2) Given the covariates X, the distribution of treatment assignments is independent of potential outcomes [6]. This project will investigate the effectiveness of the doubly robust method and tree-based method on a bank marketing dataset by comparing these methods to a naive estimator and inverse propensity weighting.

1.2 Related Work

This work builds upon a large literature of causal inference methods. The most closely related works are Dudik et al. and and Chang et al. [2 1] Dudik et al. provide a detailed explanation of doubly robust policy evaluation and compare this method to inverse propensity weighting. They show that the doubly robust method is robust to misspecification of either the outcome model or the propensity model. The authors test both methods on simulated datasets and find that the doubly robust method has low variance and bias [1]. Chang et al. propose a robust tree-based method which is computationally efficient, unbiased, nonparametric, and flexible. They apply the tree-based method to two datasets and compare its results to the naive method [2].

Other authors also test casual inference methods on marketing datasets. Lambert and Pregibon apply several methods (propensity score matching, direct outcomes model, and the doubly robust method) to advertiser trial data and and find that the doubly robust method gives results with the least variance [3]. Chan et al, compare inverse propensity weighting and the doubly robust method on simulated datasets and find that the doubly robust estimates are more robust to selection bias than other methods [1]. Want et al. begin by test experimenting with inverse propensity weighting and and the doubly robust method, but find that these methods do not work well for high dimensionality, sparsity, and imbalance of online advertisement data [6].

2 Model and Methods

2.1 Problem Definition and Data Sources

The goal of this project is to evaluate the effectiveness of the doubly robust and tree-based estimators on a bank marketing dataset. These estimators are compared to two more traditional methods: the naive method and inverse propensity weighting.

These estimators will be tested on the bank marketing dataset available from the UCI Machine Learning Repository. The dataset contains information that was gathered by a Portuguese bank during a direct marketing campaign. The goal of the project is to estimate the effect that the number of calls made to a consumer has on whether a customer decides to open a term deposit. The dataset contains 41,188 observations that include background information, such as age, marital status, education, employment, loan status, and whether the client had been contacted by a previous campaign. All of these features are used in the estimators unless otherwise specified. The treatment variable is the number of calls that the bank made to the client. The number of calls range from 0 to 43. Due to limited observations, individuals with 7 or more calls are grouped together. The outcome variable is whether the client opened a term deposit.

30% of the data is reserved for training data. The remaining 70% of the data is split equally into seven test datasets. The methods are estimated on each of the test datasets.

2.2 Model

Formally, for each treatment an individual has two potential outcomes: the outcome if he or she received the treatment and the outcome if he or she was in the control group. These outcomes are represented by $y_{1,i}$ and $y_{0,i}$, respectively. The average treatment effect (ATE) of a treatment is given by $\mathbf{E}[y_{1,i} - y_{0,i}]$, where $\mathbf{E}$ is the expectation operator. If both $y_{1,i}$ and $y_{0,i}$ were available in the data, we could simply estimate the ATE by $\sum_{i=1}^{n}(y_{1,i} - y_{0,i})$ where n is the number of observations in the dataset. Since we only observe either $y_{1,i}$ or $y_{0,i}$ in the real data depending on whether individual i received the treatment or not, we use causal methods to estimate the counterfactual. The counterfactual is the potential outcome if all subjects in the population were assigned to the treatment group.

In terms of this project, the potential treatments are 1 call, 2 calls, 3 calls, 4 calls, 5 calls , 6 calls, and 7 or more calls. I use four casual methods to estimate the counterfactual success rate for each treatment, or the estimated percent of customers who would have opened a term deposit under each treatment. For example, the counterfactual for 1 call is the estimated percent of customers who would open a deposit if every customer received exactly one call from the bank's marketing campaign.

2.3 Methods

This project uses the naive, inverse propensity weighting, doubly robust, and tree-based methods to estimate the counterfactual success rates. The methods are described below.

Naive Method: The naive method (NM) simply computes the average success rate in each treatment group. The naive estimator formula is given by $\hat{\mu}^{N}_{j,n_j} = n_j^{-1}\sum_{i=1}^{n_j} y_{j,i}$, where j is the treatment and n_j is the number of customers in the who received treatment j [6].

Inverse Propensity Weighting Method: The inverse propensity weighting (IPW) method is calculated by first estimating the propensity moddel $\hat{p_i}(j|X)$, or the probability that each individual is assigned a certain treatment given his or her characteristics. This is carried out using logistic regression and the resulting $\hat{p_i}(j|X)$ is used to calculate the estimator, $\hat{\mu}^{IPW}_{n,j} = n^{-1}\sum_{i=1}^{n}\frac{y_{j,i}\mathbb{1}_j}{\hat{p_i}}$ [2 3 7].

Doubly Robust Method: The doubly robust (DR) method combines the propensity model and an outcome model. Theory indicates that the method will be unbiased as long as one of these models is accurate. The outcome model is computed with a logistic regression $\hat{Q}_i(j, x)$ that predicts the outcome of each individual given characteristics X and treatment j. $\hat{p_i}(j|X)$ is calculated as it was for the IPW method. Finally, the DR estimator is computed as $\hat{\mu}^{DR}_{n,j} = n^{-1}\sum_{i=1}^{n}(\frac{y_{j,i}\mathbb{1}_j}{\hat{p_i}} - \frac{\mathbb{1}_j - \hat{p_i}}{\hat{p_i}}\hat{Q}_i(j, x))$ [2 3 7]

Tree Based Method: The tree based method is uses decision trees to compute an estimator that is robust to misspecification and is highly flexible [1]. The algorithm is as follows:
Input: Y_i, X_i, treatment T_i for $i = 1, 2, ...N$. **Output:** Estimated treatment effect.
Step 1: Fit a tree-based model with dependent variable T_i and independent variables X_i. **Step 2:** Within each leaf node, calculate the number of subjects and estimate the treatment effect for each treatment t. **Step 3:** Calculate the weighted final treatment effect.

3 Results

Next I present the results estimated using the naive, inverse propensity weighting, doubly robust, and tree-based methods.

3.1 Comparison of Methods

First, I calculate the mean and variance of the naive, inverse propensity weighting, and doubly robust estimates. Figure 1 describes the quality of the propensity model that was fitted in order to estimate $\hat{Q}_i(j, x)$. The ROC Curve shows the tradeoff between true positives and false positives. Random guessing would correspond to an ROC curve along the 45-degree line, so this model performs fairly well The AUC (are under the curve) for this model is 0.734. While far from perfect, the fitted propensity model performs reasonably well according to these metrics.

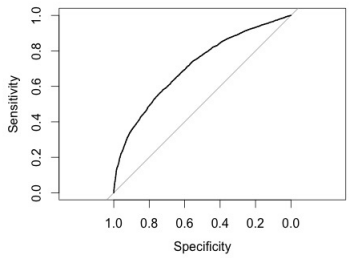

Figure 1: ROC Curve for Fitted Propensity Model

Table 1 presents the means and variance for each treatment and the naive, inverse propensity weighting, and doubly robust methods. The estimates vary substantially among the different estimators.

	Naive Method		Inverse Propensity		Doubly Robust	
# of Contacts	Mean	Variance	Mean	Variance	Mean	Variance
1	0.13046	0.00003	0.12326	0.00003	0.10039	0.00004
2	0.10368	0.00005	0.11380	0.00007	0.11444	0.00003
3	0.12632	0.00048	0.11849	0.00075	0.14228	0.00029
4	0.11111	0.00018	0.09649	0.00029	0.14671	0.00039
5	0.09524	0.00061	0.09165	0.00085	0.13728	0.00032
6	0.07368	0.00063	0.07524	0.00062	0.12754	0.00022
7+	0.02372	0.00019	0.04709	0.00038	0.11610	0.00081

Table 1: Mean and Variance for the Naive, IPW, and DR Methods

3.2 Tree-Based Method

Next, I applied the tree-based method to the dataset. For each treatment, I fit a tree using the CART algorithm. Figure 2 shows the tree for treatment 1. The trees for the other treatments are similar.

After fitting a tree for each treatment, I calculate the treated success rate in each leaf node. Using the node success rate, I calculate the overall success rate weighted by the number of observations that fall into each node. The table in the left in Table 2 shows the calculations for treatment 1. Similar calculations were carried out for the remaining treatments. The table to the right in Table 2 shows the weighted success rates for each of the 7 treatments.

3.3 Discussion

The estimates from the different methods vary substantially. The naive method implies that making one call to consumers is optimal and that increasing the treatment leads to a steady decrease in the success rate, with only 2% of customers receiving seven or more calls opening a deposit. The IPW results follow roughly the same pattern, though the decrease in the success rate is not as sharp. The DR method offers quite different estimates. Here, it appears that calling customers 4 times is optimal

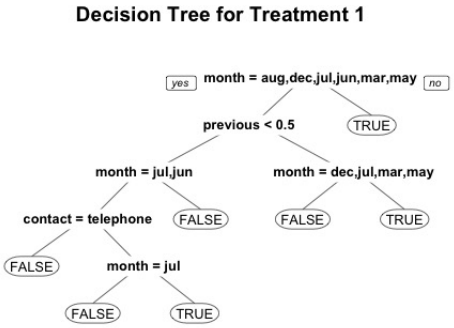

Figure 2: Classification Tree for Treatment 1

Treatment 1		
Node	**Treated Success Rate**	**Weighted Success Rate**
Five	0.03973168	0.1193662
Seven	0.07923497	
Eight	0.3712375	
Nine	0.0788084	
Eleven	0.2251712	
Twelve	0.4219114	
Thirteen	0.2157559	

# of Contacts	Weighted Success Rate
1	0.1193662
2	0.110916
3	0.1443517
4	0.1411046
5	0.1392108
6	0.1123825
7+	0.1074705

Table 2: Estimates from the Tree-Based Method

with a success rate of 14.7%. Even more striking, the 7+ treatment results in a success rate of 11.6%, much higher than the estimates from the naive and IPW methods. Interestingly, the tree-based method provides estimates that are very similar to the DR method. Here, 3 calls results in a success rate of 14.4% and 4 calls results in 14.1%. The 7+ treatment also has a similar success rate of 10.7%.

The variation in estimates is striking, but not altogether surprising. The naive method is almost certainly biased, so I would not expect these results to be accurate. The IPW method, while shown to have nice mathematical properties, often does not work well in practice especially when some observations have a low probability of being observed [6]. Furthermore, the goal of the propensity model is not to fit the data well, but to balance the covariates X across the treatment and control groups. To test this assumption, I group the observations into seven groups (as suggested by [1]) according to their estimated $\hat{p}_i(j|X)$ [1]. While a formal test is needed to draw definite conclusions, it does not appear that the covariates are balanced. Specifically, the individuals that are more likely to be treated with 7+ calls appear to be older and less likely to be married than individuals that are more likely to be in the control group. The doubly robust and tree-based methods are robust to this propensity model misspecification, so they likely lead to better estimates. In fact, the DR and tree-based estimates follow a pattern that is typical of marketing data: a linear increase in the success rate for first few contact frequencies, but then a leveling off and decreasing success rate after a certain number of contacts. Overall, my results support previous work that suggest that the DR and tree-based methods provide simple and unbiased estimates of the effect of campaigns and policies.

4 Conclusion and Future Work

This project provides further evidence that the doubly robust and tree-based methods are good options for estimating the causal effect of marketing campaigns, especially when the datasets are unbalanced. Future work should compare these two methods on larger and more complex datasets,

such as data from online advertising campaigns. Future projects should also modify the tree-based method to include bagging.

5 References

[1] Chan, D., Ge, R., Gershony, O., Hesterberg, T., & Lambert, D. (2010, July). Evaluating online ad campaigns in a pipeline: causal models at scale. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 7-16). ACM.

[2] Dudk, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. arXiv preprint arXiv:1103.4601.

[3] Lambert, D., & Pregibon, D. (2007, August). More bang for their bucks: Assessing new features for online advertisers. In Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising (pp. 7-15). ACM.

[4] Stitelman, O., Dalessandro, B., Perlich, C., & Provost, F. (2011). Estimating the effect of online display advertising on browser conversion. Data Mining and Audience Intelligence for Advertising (ADKDD 2011), 8.

[5] Varian, H. R. (2016). Causal inference in economics and marketing. Proceedings of the National Academy of Sciences, 113(27), 7310-7315.

[6] Wang, P., Liu, Y., Meytlis, M., Tsao, H. Y., Yang, J., & Huang, P. (2014, February). An efficient framework for online advertising effectiveness measurement and comparison. In Proceedings of the 7th ACM international conference on Web search and data mining (pp. 163-172). ACM.

[7] Wang, P., Sun, W., Yin, D., Yang, J., & Chang, Y. (2015, February). Robust tree-based causal inference for complex ad effectiveness analysis. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. 67-76). ACM.