Factual evidence for Causal Classification: Treatment Effect versus Outcome Prediction

By

Huanci Wang

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2021

Professor Marti G. Subrahmanyam Professor Foster Provost

Professor Christina Wang

Professor Wendy Jin

Faculty Advisers Thesis Adviser

# **Abstract**

Incentive marketing programs are often adopted by companies aiming for stimulating prospective or existing customers. Since such campaigns come with a cost, companies will gain most returns if they aim to target customers who will be led to purchase after being targeted instead of conducting random targeting. To predict customers’ behavior after being targeted with an advertisement, two kinds of models are usually considered: outcome prediction models and treatment effect models. Treatment effect models take into consideration the causal effect of treatment and outcome prediction models predict the behavioral outcomes. In theory, treatment effect models should give better payoff after implementation, since they estimate the real target: effect of treatment. Interestingly, outcome prediction is more often adopted in practice and the existing literature has not yet established which is the most suitable model with real-life dataset.

Choosing the model best at targeting influenceable customers will improve companies’ return in practice because this will reduce the cost of targeting customers who would purchase anyways. We define such customers as “sure-things”. If simple outcome prediction models are more effective than more complex treatment effect models, companies will not only target more suitable customers but also cut costs in terms of data collection as well as complex model building. Based on a real-world dataset, my research analyzes: (1) whether treatment effect models truly behave better on real-life data, and (2) when and why the outcome prediction model might lead to a better payoff when adopted. My findings so far show that the outcome prediction models are indeed better at predicting the most influenceable customers given varied data amounts, algorithms, evaluation methods and targets. The error-prone process of distinguishing between “sure-things” and “persuadables” together with the high variance of two-model uplift predictors lead to the differences between outcome prediction models and treatment effect models. This implies that treatment effect models could be replaced by outcome prediction models, which are less demanding in terms of costly customer data and are more effective and economical.

**Keywords:** treatment effect, outcome prediction

# **1.** **Introduction**

Predictive models are often applied to improve tasks such as targeted marketing. For instance, if company A wants to present a webpage advertisement for its app to its users and wants to target the people that the advertisement will have the most impact on, they will need a predictive model with users’ attributes as their input and users’ behavioral change as their output. The users’ behavioral change is to download the app after viewing the ad in this case when they otherwise would not have. This prediction model will improve company A’s decision making about which customer to target with the advertisement. This treatment decision is more important when the treatment comes with a cost. However, in practice, companies often use outcome prediction models which take people’s purchasing behavior as the target variable, regardless of the fact that many people’s purchasing behavior may not be a result of the treatment, and that in that case there is no causal relationship between purchasing and advertising. We are going to refer to these people who would take the action anyway as “Sure Things” in this paper. People who will be positively affected by the campaign and conduct purchasing are “Persuadables”.

表格

中度可信度描述已自动生成

Figure 1 Causal Classification Labels (Fernandez, Provost, 2019)

If the treatment effect model instead surpasses the outcome prediction model using real data, companies will experience an improvement in the payoff from making its treatment decisions due to less data collection cost and ineffective treatment cost. The payoffs of the two models are illustrated below. I will also analyze when an outcome prediction model possibly performs better and give it a theoretical explanation. With our findings, entities can adjust their data augmentation process and model building process for better targeted marketing outcomes.

图片包含 图示

描述已自动生成

Figure 2 Pro-Con Analysis of Two Models

Uplift models/causal inference models have been widely studied and compared. Popular methods include causal tree, single model learner, two model learner, x-learner etc. Outcome prediction is even more widely used in prediction practices, but no research has established which is better in terms of choosing the persuadable customers. In their 2019 study, Fernández-Loría and Provost argued that outcome prediction models are theoretically better than treatment effect models in terms of choosing the persuadable customers under certain circumstances. They also proposed the causal bias-variance trade-off theory that explains causal inference models’ potential worse performance--since the treatment effect models require the combination of multiple outcome prediction models and this leads to a larger variance of the prediction results. They showed this theoretically and with simulations, but they did not verify their simulations on empirical data. That is what I will do in this paper. Based on their theoretical papers, I further investigated the two models’ performance differences and the reason behind this difference with real-life data. My empirical result contributes to verifying their theoretical arguments.

# **2.** **Thesis Problem**

## **2.1** **Definition of treatment effect models/ uplift model**

According to Gerardy, “companies (typically in telecommunication or e-business sectors) are interested in uplift modeling to estimate the effect of an action on some customer outcome” (Gerardy, 3). In other words, treatment effect models measure the uplift (causal improvement) of a certain trigger. For the simplicity and coherency of the paper, I will not distinguish between uplift models and treatment effect models and use the terms uplift model and treatment effect model interchangeably. Treatment effect models are also referred to as causal effect models in other circumstances and we do not aim to distinguish between these two kinds of models, either.

## **2.2** **Problem Definition**

In this paper, there are three main questions to investigate:

1. Which model performs better in terms of maximizing the company’s reward of the targeting campaign, or, targeting the most persuadable people?

2. When does either model ever surpass another?

3. Why does this difference in performance happen?

**2.3 Thesis statement**

I argue that there are situations where the outcome prediction models are better at predicting the most influenceable customers given varied data amounts, algorithms and training data treatments. I also argue that outcome prediction models are a better choice for implementation. This implies that treatment effect models could be replaced by outcome prediction models which are less demanding in terms of customer data and are more effective and economical.

# **3.** **Methodology**

## **3.1** **Data**

Criteo dataset a large and widely used dataset for causal inference. Although we are only testing on this one dataset, it is suitable since it is almost the only dataset as such available and it is very large in size. The dataset contains eleven anonymous features for each customer. The company randomly treated 85% of the customers and documented their conversion and visiting behaviors. Conversion here means purchasing behaviors such as downloading an application. Visit means clicking into a website.

表格

描述已自动生成

Figure 3 Criteo Dataset

## **3.2** **Models**

Outcome prediction models: I adopted decision tree, random forest and logistic regression to predict conversion and visit. I choose them because they are some of the most widely used outcome prediction models and they all show consistent results. More reasons for choosing them will be provided in the result section.

Uplift models: I adopted the two-model approach due to its representativeness. According to Gerardy, “The Two-Model approach has been applied in several uplift papers (Radcliffe (2007), Nassif et al. (2013)) and is often used as a baseline model” (Gerardy, 3). The approach goes like this: build two outcome prediction models, one with treated data and one with untreated data, i.e., E[Yi(1)|Xi] and E[Yi(0)|Xi] . In this case, we build the first model with treated customer data (treatment = 1) to predict conversion or visit. We build the second model with untreated customer data (treatment = 0) to predict conversion or visit. Given a large enough dataset and random assignment of treatment, we can use the difference between model 1 and model 2’s prediction result to represent the estimated uplift of a treatment on a customer. Scholars have not yet reached a consensus on the best casual model and some support the two model approach (Olaya et al.) while others do not (Nie et al.) I will discuss more about the two-model approach in the discussion and limitations section.

## **3.3 Evaluation Methods**

We have adopted average reward (based on Li’s paper), and average treatment effect to evaluate the models. Average treatment effect is often adopted by scholars to measure uplift. It is calculated by the difference between average treatment effect on treated and average treatment effect on untreated. The treatment effect is 1 if one is both treated and decides to do purchase. Average reward is a relatively new measuring method by Li and his group. Their algorithm is presented below.

文本, 信件

描述已自动生成

Figure 4 Li et. al. Evaluation Method Algorithm (Li et. al., 665)

Both measurements are effective at measuring uplift. The new method introduced by Li compares the reward to existing data while the old methods calculate the uplift based on average uplift. I will adopt both methods in the following section.

# **4. Results**

## **4.1 Can Outcome Prediction Models Do Just As Well**

My analysis shows that outcome prediction models are better than treatment effect models with different algorithms, evaluation methods and increased amount of training data.

I started with a decision tree algorithm to build the uplift model. I adopted nested cross-validation to test different parameters and in this case min\_samples\_leaf. The result is shown in the following graph. The average reward and effect for targets of uplift models are lower than either treated model or untreated model. This is a very promising first step that shows there are indeed circumstances where the treatment effect model performs worse.

图形用户界面

中度可信度描述已自动生成

Figure 5 Tree Algorithm Result. The outcome prediction model using the decision tree algorithm is presented in this graph as “uplift”. Model\_all, model\_ treated and model\_untreated use all data, treated customers’ data and untreated customers’ data as training data respectively. For\_target\_treated and for target\_untreated are average treatment effect on treated and average treatment effect on untreated respectively. Effect\_for\_targeted is average treatment effect.

This result might be particular to the decision tree algorithm. We incorporated a random forest model which is in a way the collective result of multiple decision trees and should give a more convincing and consistent result. The average reward and effect for targeted uplift models are again lower than either treated model or untreated model.

表格

中度可信度描述已自动生成

Figure 6 Random Forest Algorithm Result. The outcome prediction model using the random forest algorithm is presented in this graph as “uplift”. Model\_ treated and model\_untreated use treated customers’ data and untreated customers’ data as training data respectively. For\_target\_treated and for target\_untreated are average treatment effect on treated and average treatment effect on untreated respectively. Effect\_for\_targeted is average treatment effect. Avg\_reward is the average reward.

Finally, we tested logistic regression so that this research not only includes tree-based models. The average reward and effect for targeted uplift models are again lower than either treated model or untreated model.

图形用户界面, 应用程序

描述已自动生成

Figure 7 Logistic Regression Algorithm Result. The outcome prediction model using the logistic regression algorithm is presented in this graph as “uplift”. Model\_ treated and model\_untreated use treated customers’ data and untreated customers’ data as training data respectively. For\_target\_treated and for target\_untreated are average treatment effect on treated and average treatment effect on untreated respectively. Effect\_for\_targeted is average treatment effect. Avg\_reward is the average reward.

So far, we have shown that the average reward and causal effect for targeted uplift models are lower than either treated model or untreated model alone, using different algorithms. Remember that these models are built with the same amount of training data.

It is reasonable to question the effect of changing the amount of training data used for both theoretical and practical reasons. In theory, the good performance of outcome prediction might be a result of the large dataset and we should test this out whether data amount is the limitation. In many applications, such a large amount of data is extremely difficult to acquire. It is also possible that we cannot get the same amount of treat and untreated data. These issues prompt us to continue to our next analysis: vary the amount of training data.

We increased training data gradually and plotted the corresponding treatment effects and average rewards. These plots are called learning curves and the following plot shows treatment effect level as the amount of training data increases. Although the treatment effect for the uplift model increases gradually as we use more training data, it never reaches the outcome prediction models’ level, which also increases slightly as more training data is applied. This further confirms that outcome prediction models perform better than treatment effect models on this dataset. Also, we can see that increasing training data has a greater effect on uplift model’s performance than outcome prediction model’s. This is in line with the notion that learning treatment effect models will require more data, due to their higher variance.

图表, 折线图

描述已自动生成

Figure 8 Treatment Effect Level as Number of Training Data Increase

By far, all else being equal, algorithms and number of training data does not invalidate our claim. Next, we want to show that different evaluation methods do not invalidate our argument either. To do this, we plotted the learning curve that measures different evaluation figures such as average reward. The uplift model never performs as well as outcome prediction models no matter which evaluation method is adopted.

图表, 折线图

描述已自动生成

Figure 9 Average Reward Level as Number of Training Data Increase

The graphs above show the result of targeting at the top 10% targets from the validation group that have the highest prediction scoring. We choose to target a limited group to imitate practical cases when there are limited resources.

Since the result of targeting the top 10% group is consistent, our last variation is on the people targeted. We divided the validation group into ten sub-groups from people with the highest prediction score to the lowest and plotted their result. This is also done with different evaluation methods.

图表, 折线图

描述已自动生成

Figure 10 Treatment Effect by Percentage Targeted

图表, 折线图

描述已自动生成

Figure 11 Average Reward by Percentage Targeted

We target each percentage range, and see the reward for each bin. This rise in the last bin means that the uplift model considers this group of people unlikely to be affected by the treatment, but they have a higher reward/treatment effect (because some of them do indeed convert). More precisely, this lift means that some people who are actually persuadables were considered as sure-things by the uplift model. We will further investigate this in the following sections.

## **4.2 Why is there a lift at the bottom 10%**

The rise in the last bin seems to be an important indicator of the difference between outcome prediction model and treatment effect model. To carefully investigate who are these people in the last ten percent and why do they lead to such a difference, we graphed in Figure 12 the percentage overlap of people being targeted by each model in each 10% bucket. In the top left corner of the second matrix (0.73 cell) for instance, we can see there is 73% overlap of customers between the uplift and outcome prediction model in each of their top 10% bins. If we have a look at the lower left corner of the matrix in the middle, we can see a group (0.26 population of column 1) left out by the uplift but considered likely to convert according to outcome prediction models. These people have higher untreated scores and relatively lower treated scores.

The difference between the first matrix (comparing treated and untreated) and the following two matrices points to the different targets of the two models. In the first matrix, most dark cells align in the diagonal which means that treated and untreated models categorize people in relatively the same way. In other words, people who are considered influenceable targets by one outcome prediction model are also considered influenceable by the other outcome prediction model. However high overlap does not always happen along the diagonal in the second and third matrix. When a group of people is categorized in the top 10% by the treated uplift model, 73% of these people are also in the top 10% of the uplift model but 26% of them are surprisingly in the bottom 10%. This coincides with the surprising rise in the last bin in the graph above.

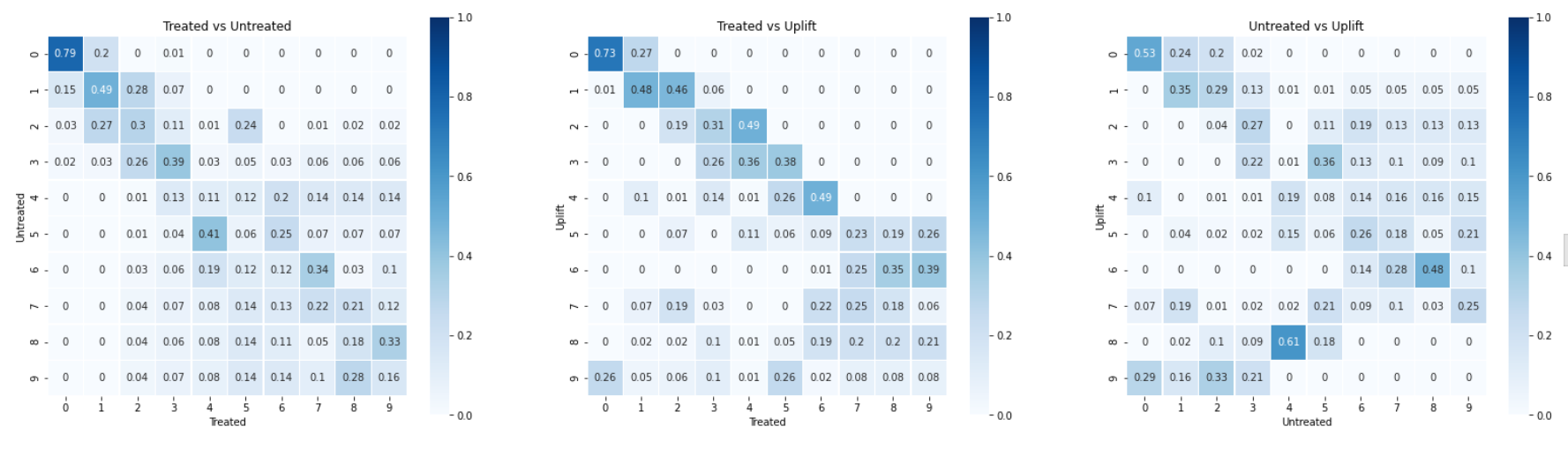


Figure 12 Matrix for Percentage of Overlap of Customers (Tree Algorithm). Figure 12 Customers in the validation set are ranked by treated outcome prediction and treatment effect model’s prediction scores respectively. We divide the sorted population into ten deciles and in each decile, there are some overlaps between the two models’ rankings. For instance, people in the top left cell are considered top 10% by both models’ scoring and they make up 73% of the overall population that is top 10% by either one of the models.

This graph shows us two results: first, the uplift model and treatment effect models are highly correlated. They are interchangeable in a way not only due to their similar performance according to the different evaluation methods but also due to the fact that they relatively rank the targets in the same way. A better illustration of this correlation is shown in the following graph plotted using logistic regression algorithm models. There are large overlaps happening between uplift model and treatment effect models when using logistic regression algorithms.

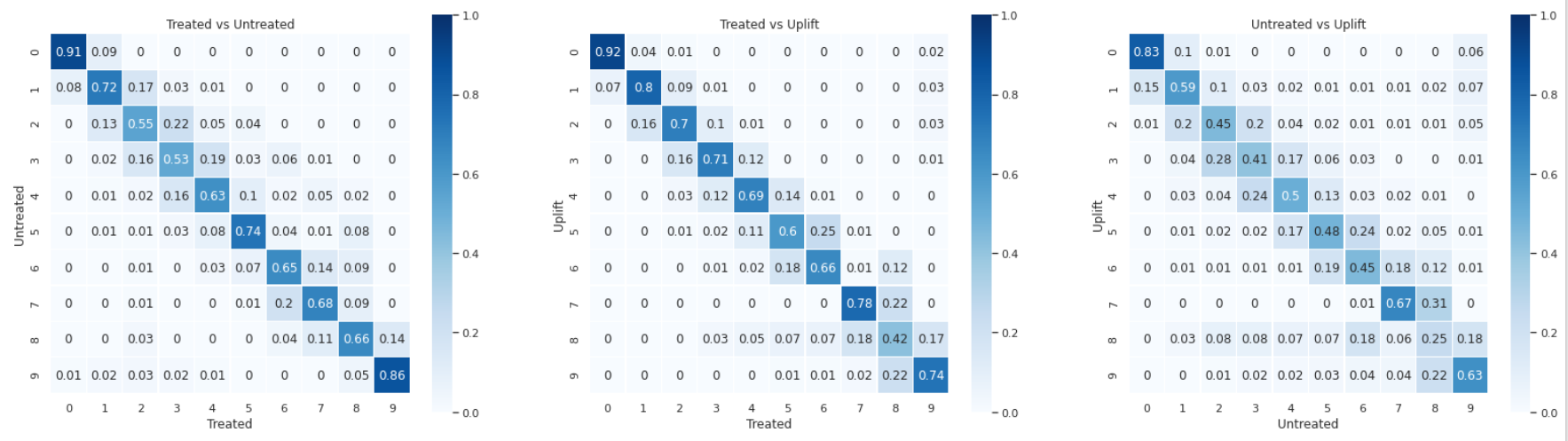


Figure 13 Matrix for Percentage of Overlap of Customers (Logistic Regression Algorithm). Figure 13 Customers in the validation set are ranked by treated outcome prediction and treatment effect model’s prediction scores respectively. We divide the sorted population into ten deciles and in each decile, there are some overlaps between the two models’ rankings.

Second, the matrix points at some possible limitations of the uplift model that leads to its worse performance. In the next section, we are going to investigate the reason for a worse performance of this treatment effect model from observing the 10% lift, the matrix and score distributions.

## **4.3 Why the treatment effect model is error-prone**

The graph below plots the data points in the first column of the second matrix. X-axis means the scoring by treated outcome model and y-axis means the scoring by untreated outcome model. By the nature of the two-way uplift model, the difference of the two is the scoring of the uplift model. The orange dots in this graph belong to the 0.26 cell and they are the differences between the two models (people that are considered very persuadable by outcome prediction models but unpersuadable by treatment effect models). In other words, they are categorized in the top 10% of the uplift model but according to the uplift model, they are in the last 10%.

图表, 散点图

描述已自动生成

Figure 14 Treated Score and Untreated score for 0.73 and 0.26 Cell Customers. Figure 14 The figure extracts customers from the first column of figure 12. The x-axis and y-axis are treated score and untreated score respectively. Orange dots and blue dots are customers within 0.26 and 0.73 cells respectively.

The orange point on the top right corner has both a high treated score and untreated score, but the difference between the two is small. Thus, these points have low uplift scores and are categorized in the last bin by the uplift model. They are targetable people according to outcome prediction models because they have high treated/untreated outcome prediction model scores. However, while the uplift model is trying to eliminate sure-things by subtracting the untreated score from treated score, some persuadables might be eliminated at the same time. Although we cannot prove that the orange dots on the top right corners are “mistakes” made by the uplift model, we can definitely see that this is where the difference happens and that this process is error-prone.

To further illustrate this difficulty for the uplift model--of differentiating between sure-things and persuadables--I will show the score distribution of the treated outcome prediction model, untreated outcome prediction model and the first few branches of the tree-based outcome prediction model below.

图表, 直方图

描述已自动生成

Figure 15 Treated Score Distribution

图表, 直方图

描述已自动生成

Figure 16 Untreated Score Distribution

According to this score distribution, people ranked top 10% by the outcome prediction models have outstanding high scores (0.22) which is an indication of an especially predictive feature. By observing the decision tree, we identified this feature F11 which leads to a very high score by the outcome prediction model. As long as F11 is small enough, a group of people will be given a high outcome prediction score. By using an uplift model, the effect of such a feature is effectively eliminated but does not essentially lead to a better performance. In other words, in the process of eliminating “sure-things” and validating “persuadables”, uplift is not only eliminating some “sure-things” but also eliminating the causal effect predicting power of features that also possess strong outcome prediction power. Some persuadables are incorrectly left out.

一些文字和图片的手机截图

中度可信度描述已自动生成一些文字和图片的手机截图

低可信度描述已自动生成

Figure 17 Tree Treated and Untreated Model Visualization

Finally, according to Fernández-Loría and Provost’s study, the variance is another important issue. We believe that the orange dot in the bottom left corner is very likely the result of variance. In other words, these people might not consistently be categorized in the bottom 10% by uplift model since the two-way uplift model is a combination of two outcome prediction models and this process inevitably increases the variance. To verify this, we extracted 10 groups of different training data each time to build different outcome prediction models, and we built 10\*10 uplift models using the combination of every two outcome prediction models. We documented the ranking of the same validation set each time to track where each customer ranked by each uplift model. The result is as follows.

图表, 散点图

描述已自动生成

Figure 18 Probability of Each Customer Ending up in Either 0.73 or 0.26 Cell

The axis of this graph means the percentage of time each person ends up in the position of the 0.73 (top left) or 0.26 (bottom left) cell according to the matrix between uplift and treated model. Many dots are scattered throughout the graph which means that many people are seldom present in either of those two positions. For the ones that are present, some are often present in the 0.73 position and they appear in the left of the graph, but few often present in the 0.26 position. This means that few customers are consistently categorized in the 0.26 cell. Also, the people in the bottom left cell or orange dots (low ranking by uplift but high ranking by outcome prediction) are not stable. This also shows that the uplift model is error-prone and it cannot give consistent results due to its high variance.

# **5. Discussion and Limitations**

To present a more thorough result, we still need to investigate more causal inference models including meta-learner approach, class transformation method and so on. Although the two-model approach is considered the baseline for treatment effect models, it is not the best model and there is no universal best choice when it comes to treatment effect models. Two-models uplift is controversial according to Gerardy. “Some authors (Zaniewicz and Jaroszewicz (2013), Athey and Imbens (2015b)) show that it is often outperformed by other methods. One reason is that the two models focus on predicting the outcome separately and can therefore miss the ‘weaker’ uplift signal” (Gerardy, 4). That being said, the results of the two-model approach at least proves that treatment effect models are in no way completely better than outcome prediction models. This practical result should be considered a thought-provoking start for further studies.

For future work, we will also study more into the variance issues that lies behind the worse performance of treatment effect models. For instance, so far we have not shown the definite effect of high variance. If we could indeed show that high variance results in aggregately worse performance of the uplift model, then we could further evaluate how much the variance impairs the uplift models’ performance. The same applies to evaluating the aggregated impact of categorizing a persuadable person as sure-thing. Interesting work will be done later in this direction.

Moreover, our practical research is built upon one dataset only for now due to the limitation of resources. We cannot prove that this result is suitable for other dataset so far. We would love to continue our studies on other datasets when they become available.

# **6.** **Conclusion**

In this paper, I used imperial analysis with real-life dataset to show that the outcome prediction models are better at predicting the most persuadable customers given varied data amount, algorithms, evaluation methods and targeted customers.

In terms of the reason behind the performance difference, we first showed that the two types of models have large correlation between each other. However, the people left out by the uplift model might explain its poor performance compared with the outcome prediction model. They were left out because both treated and untreated models give high scores of conversion while the uplift model eliminates this effect. In addition to the previous causal bias-variance theory, we have a new reason for performance gap: overlap between outcome predictive feature and causal predictive feature. This indicates that to further improve ability to predict causal influence, we need to solve for the shortcomings of the existing two model approach uplift model that eliminate causal effect while trying to eliminate “sure-thing” targets.

In terms of practical application, companies can use simple outcome prediction models to replace or as a proxy of treatment effect models. Outcome prediction models are not only easier to build but also requires less data, and less labeled data. While the uplift model requires an adequate amount of both treated and untreated data, outcome prediction models can work well with either treated or untreated data. This is great news for companies in real life where most customers are untreated (never exposed to a new product). In terms of reducing the cost and increasing profit from marketing campaigns, we believe outcome prediction models have a promising effect.

# **References**

Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. Journal

of Marketing Research, 55(1):80-98, 2018.

Brian Dalessandro, Daizhuo Chen, Troy Raeder, Claudia Perlich, Melinda Han Williams,

and Foster Provost. Scalable hands-free transfer learning for online advertising. In Pro-ceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1573-1582, 2014.

Brian Dalessandro, Rod Hook, Claudia Perlich, and Foster Provost. Evaluating and

optimizing online advertising: Forget the click, but there are good proxies. Big data, 3(2):

90-102, 2015.

Diemert Eustache, Betlei Artem, Christophe Renaudin, and Amini Massih-Reza. “A Large

Scale Benchmark for Uplift Modeling”. In: Proceedings of the AdKDD and Target Ad

Workshop, KDD, London, United Kingdom, August, 20, 2018. ACM, 2018.

Fernández, Carlos and Provost, Foster, Causal Classification: Treatment Effect Estimation vs.

Outcome Prediction (June 22, 2019). Available at

SSRN: <https://ssrn.com/abstract=3408524> or [http://dx.doi.org/10.2139/ssrn.3408524](https://dx.doi.org/10.2139/ssrn.3408524)

Gerardy, Jean-Yves. “Causal Inference and Uplift Modeling: A Review of the Literature”.

JMLR: Workshop and Conference Proceedings 67:1–13, 2016.

Lihong Li et al. “A Contextual-Bandit Approach to Personalized News Article

Recommendation”. In: Proceedings of the 19th international conference on World wide web.

2010.