ECON 690(002) MACHINE LEARNING

CHOICES OF THE OPTIMAL PROMOTIONS STRATEGIES

Jinyu ZHOU *jzhou*337@*wisc.edu*
Xueqing YANG (Ivy) *xyang*398@*wisc.edu*
Zheming WANG *zwang*2236@*wisc.edu*
Yan WANG *wang*2264@*wisc.edu*
Department of Economics
University of Wisconsin-Madison , Madison

December 13, 2018

ABSTRACT

With the information of beer sold in Dominick's, franchised grocery stores scattered in Chicago area, we implemented the causal forest to find to what category of beer and in what time during the year would be the best strategy of promotion implementation. We concluded that for the beer price at around 10 dollar per bottle or can has been impacted the most by the promotion, and the promotion is most effective during the summer.

INTRODUCTION

We approached the problem by two methods: the causal forest and the random forest. The first one is evolved from the second method, it could help us to pick up the key variables which is important to the regression. It is a method based on decision tree and the random forest, and it also belongs to the family of non-parametric regression. "The causal forest model is pointwise consistent for the true treatment effect, and has an asymptotically Gaussian and centered sampling distribution."(Stefan Wager, Susan Arthey,2017).

Our data provided us with 80 variables, many of those have the problem of multi-col-linearity, so we chose causal forest to examine the treatment effect, since it would not be bothered by col-linearity issue when doing the regression.

From the regression result, we found it is better to put promotions on beers with higher prices (over the average price), and to put promotions during the summer time. Despite of this, the profit of beer brand "Miller" is sensitive to the promotion, so applying promotions on it might be a good choice.

Descriptive Analysis

Our data are collected from Domonick's Finer Foods franchises in Chicago area since the September of 1989, we picked up the data on beer sector, reshaped the beer sold by brands, types and the month they have been sold. The data also gave us information about the promotions which were applied on by weekly frequency.

The Description of Promotion applied

The promotions are usually applied to some fixed brands of beers. If we label the UPC (the universal product code), we could find the promotion of all three kinds: coupon, bonus and

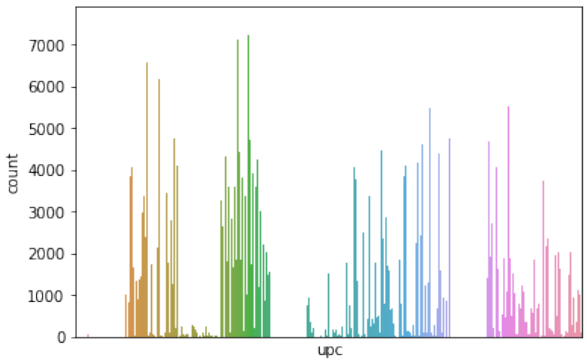

Figure 1: Bonus Allocation over UPC

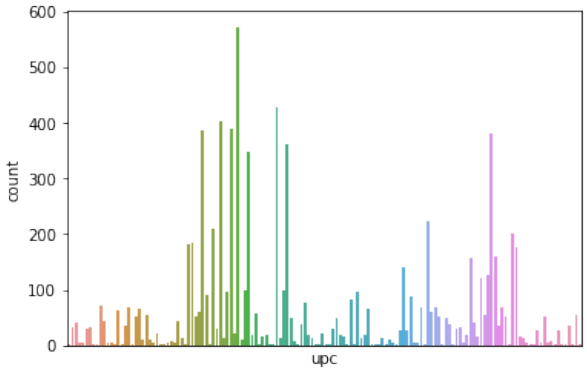

Figure 2: Sales Allocation over UPC

sales. Bonus and Sales are basically added on the same products (see figure 1 and figure 2). So, we could have some co-treatment effects when doing the analysis. Coupons were used only on four products, and the frequency is showed in figure 3.

Since the Coupon treatment effect could be nuance since it is rarely used, and Sales could have the co-treatment with Bonus, we decided to treat all three promotions as a whole.

Another aspect we focused is at what time is the best to implement promotions, so we looked at when did those grocery stores generally apply promotions. The first week in data is from Sep.14 1989 to Sep.20 1989, and labeled consecutively. As we can see, the profit was dramatically drop in the 300^{th} week during observation (by figure 4). Also, there are some outliers from 150^{th} week to 250^{th} week with extraordinary profit.

The Description of Products

We have over 200 kinds of beer, some of which are produced by the same company. We gave each kind of beer with two labels: the brand and the type. The brand stands for the company

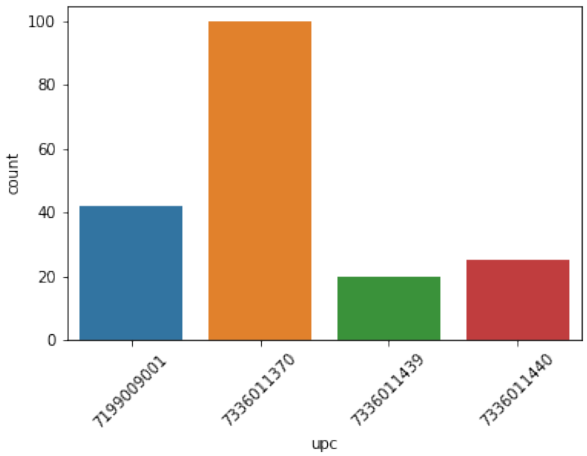

Figure 3: Coupon Allocation over UPC

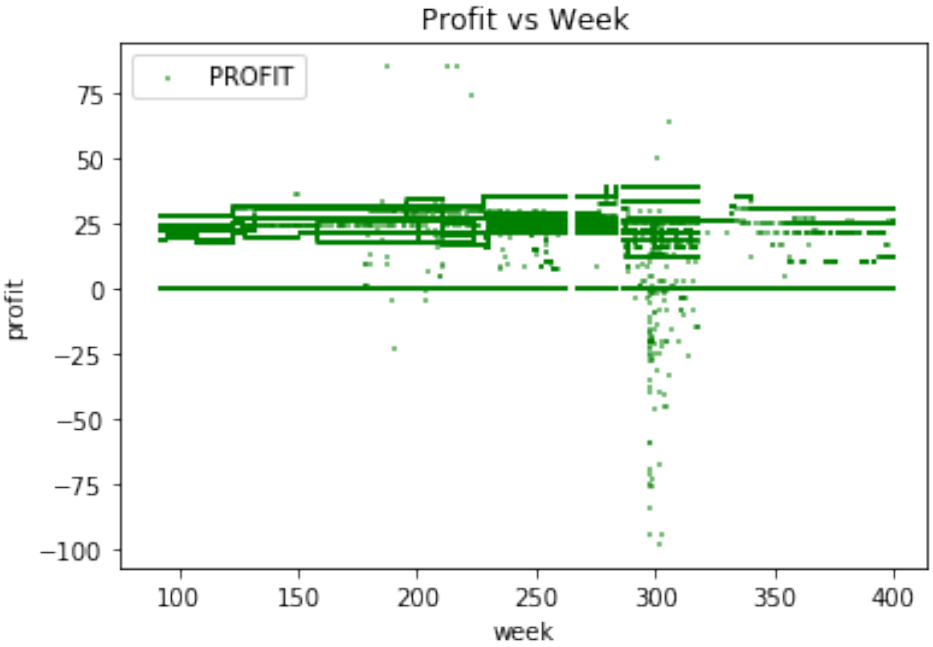

Figure 4: Profit over weeks

where the beer belongs to. The type, including Ale, Lager, Draft, Black and so on, is also labeled since it may contain some information about customer's preference.

Methodology

We firstly used Random Forest, and then applied Causal Forest to optimize our estimation. Also we ran the cross validation to assess the outcome, ensuring there is no over-fitting happened.

Decision Trees and CART

Decision trees consisted the basics of both the random forest and the causal forest. It uses the top-down strategy in which the root node could diverge into binary classification and the same divergence would generated again and again until reaching the terminal nodes.

By bootstrap aggregating (bagging), multiple decision trees would be set up to form a simple forest, where multiple training sets are generated with replacement. In our case, the "move" (the units sold) could be repeated. After the training, we would get training sets and we could apply CART (Classification and Regression Tree) model on each subgroup of our sample sets. However, in this step, the subgroups of the sample might be correlated.

The Basics of Random Forest

The random Forest is a popular ensemble to build the predictive models of both classification and regression problems. This model could reduce the correlation problems induced by CART model.

Given a set of simple trees, the random forest would select n variables at random from all N variables (N=80 in our case), and these variables are independent from the nodes we set in the simple forest. By these n variables, we repeat the decision tree process and a new series of nodes generated. We can do the above steps repeatedly, and finally come up with the Random Forest Prediction: $s = \frac{1}{K}\sum_{K=1}^{K} K^{th}$

The Rationale of Causal Forest

Although the Random Forest has already solved the problem of correlation issue, there are still some shortcomings of the Random Forest Method. It is hard to be interpreted and its

leaves (the terminal nodes) cannot represent a natural clustering of data into homogeneous groups.

Since we want to interpret our result better, we further applied the Causal Forest method. Those disadvantages of the Random Forest could be solved by clustering in the Causal Forest.

Here are the algorithms for the Causal Forest Method:

Suppose we have access to n independent and identically distributed training examples labeled i = 1, ..., n, each of which consists of a feature vector $X_i \in [0,1]d$,a response $Y_i \in R$,and a treatment indicator $W_i \in \{0,1\}$. The existence of potential outcomes Y (1) i and Y (0) i corresponding respectively to the response the i-th subject would have experienced with and without the treatment, and define the treatment effect at x as:

$$\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x] \tag{1}$$

To estimate this function $\tau(x)$, Athey and Wager made two important assumptions:

1. Assumption of Unconfoundedness: the treatment assignment W_i is independent of the potential outcomes for $Y_i|X_i$:

   $\{Y_i^{(0)}, Y_i^{(1)}\} \perp W_i|X_i$

2. Assumption of overlap:for some $\varepsilon > 0$ and all x$\in$ [0,1],

   ε<P[W=1|X=x]<1-ε

This condition effectively guarantees the enough treatment and control units near each test point x for local, when the sample size is enormous.

With assumptions above, the first step of causal forest is to use decision tree to group our observations. Athey and Wager used CART (Classification and Regression Tree) algorithm for decision tree since CART uses Gini coefficient as its standard, and CART could be used for both grouping and regression. The result of CART is a binary tree.

Based on decision tree, random forest could be the second step. In the random forest part, we firstly use bootstrap to sampling data with replacement, meanwhile, we use the decision tree method to make prediction for subsets in different features (independent variables X). The above process could be repeated for many times to produce huge numbers of trees. The result in final strategy is determined by the voting of all produced decision trees.

In the causal forest, after grouping by decision trees, it takes the median of treatment group minus the median of non-treatment group for every leaves L in trees. And the treatment effect for every leaves L comes out: for x in L:

$$\tau(x) = \frac{1}{|i : W_i = 1, X_i \in L|} \sum_{i:w_i=1, x_i \in L} Y_i - \frac{1}{|i : W_i = 0, X_i \in L|} \sum_{i:w_i=0, x_i \in L} Y_i \tag{2}$$

And under above two assumptions and continuity in math, such trees can be used to grow causal forests that are consistent for $\tau(x)$. In order to obtain the result of inference, a new concept called "honest tree" is created in causal forest model: for every observation i, its dependent variable Y can be used to calculate the treatment effect or used to generate the decision tree, but it cannot be used for both these two ways. Therefore, Athey and Wager created two new algorithms in causal forest: one is called Double-Sample Tree, another one is called propensity tree.

Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits. Input: n training examples of the form (X_i, Y_i) for regression trees or (X_i, Y_i,W_i) for causal trees, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k. Double-sample regression trees make predictions $\tilde{\mu}$(x) using $\mu(x) = \frac{1}{|i:W_i=1, X_i \in L|} \sum_{i:w_i=1, x_i \in L} Y_i$ on the leaf containing x, only using the I-sample observations. The splitting criteria is the standard for CART regression trees (minimizing mean-squared error of predictions). Splits are restricted so that each leaf of the tree must contain k or more I-sample observations. Double-sample causal trees are defined similarly, except that for prediction we estimate $\tilde{\mu}$(x) using (2) on the I sample. In addition, each leaf of the tree must contain k or more I-sample observations of each treatment class. Here is am important reminder: we used decision tree to sampling every time, however samples are randomly separated after random forest. Thus, we could be guaranteed to make use of all information in all samples.

Propensity trees use only the treatment assignment indicator W_i to place splits, and save the responses Y_i for estimating τ . Input: n training examples (X_i, Y_i, W_i), where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k. Three steps in this algorithm: The first step is drawing a random subsample $I \in \{1, ..., n\}$ of size $|I| = s$ (no replacement). Secondly, train a classification tree using sample I where the outcome is the treatment assignment, i.e., on the (X_i, W_i) pairs with i∈ I. Each leaf of the tree must have k or more observations of each treatment class. At last,estimate $\tau(x)$ using (2) on the leaf containing x. After these two algorithms, Athey and Wager proved their results are followed as asymptotically Gaussian and centered sampling distribution and use Jackknife's method to construct the confidence interval. And their experiments' results of causal forest and k-nearest neighbor algorithm shows causal forest has better performance in higher dimension than k_{nn} algorithm since its mean square error is relatively smaller.

Reason for Causal Forest

Here are advantages of causal random forest attract us to use it.

1. The causal random forest is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference. As we know, machine learning focus more on results of prediction rather than inference. This model does not only combine decision tree, random forest and causal inference, but also provides a powerful method to figure out the estimatesâĂŹ asymptotic normality distribution and confidence interval.

2. Compared with other traditional non-parametric method in machine learning, such as nearest-neighbor matching and kernel matching, causal forest performs much better on presence of irrelevant covariates, which means the causal forest could get rid of the problem of "dimension reduction curse".

Cross Validation Rationale

We also plan to use 10-folds cross validation on Lasso regression (Least Absolute Shrinkage and Selection Operator Regression) to evaluate our result. The reason we do not choose ridge regression is that it cannot perform co-variate selection and therefore does not help to interpret the model.

Lasso regression is better since it could evaluate and interpret the regression result by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients.

$$\hat{\beta}^{lasso} = argmin_{\beta\in\mathbb{R}}||y - X\beta||_2^2 + \lambda\sum_{j=1}^{p}|\beta_j| = argmin_{\beta\in\mathbb{R}}\underbrace{||y - X\beta||_2^2}_{loss} + \lambda\ \underbrace{||\beta||_1}_{penalty} \tag{3}$$

The predictor revenue is as same as Causal Forest, and we choose to use Gaussian distribution to deal with our variables. Since cross validation gives us bunch of λ, we choose the lambda with least standard error to regress revenue on all the quantitative variables:

$Y_{revenue} = \beta_1 X_{price} + \beta_2 X_{week} + \beta_3 X_{volume} + \beta_4 X_{sale} + \varepsilon$

The result is showed in table 1: From the regression we could verify the promotion contribute to the revenue significantly. We also need to do the cross validation which make sure there are no over-fitting issues for our regression.

price	7.1037
week	-0.0269
volume	-0.0049
sale	40.1993

Table 1: Coefficients for variables

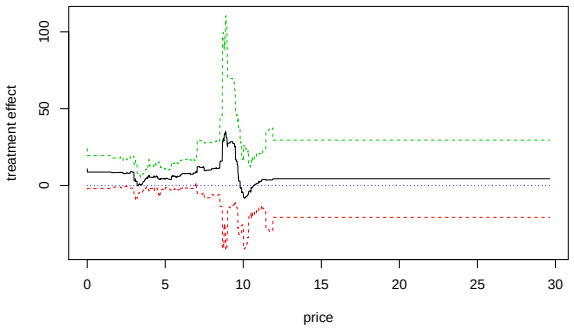

Figure 5: Treatment Effect over Unit Price
[1]the green and red lines described the confidence interval

Results of Causal Forest Method

We labeled the brand name and beer types. For the prices and volume in a bottle, we chopped them by quantile, so we could examine at which price level and at which size, are those beers the most sensitive to the promotions. The Causal Forest picked "price", "week", "brand", "type" to be the main variables, so we examined over these variables separately.

Treatment Effects on different prices

From the Causal Forest regression, we find our profit is improved significantly when the promotions were applied to the beers whose price (denoted as P) $P \in [6,8]$ (see figure 5), but the treatment effects suddenly dropped to be negative when P=10. Hence, the promotion should only be applied to those beers priced at [6,8], and we cannot implement any kind of promotion to those beers exceed this range. Applying promotions to certain brands of beer also brings a significant profit improvement. As illustrated in figure 6, those brands labeled at around 130 are influenced by promotions the most, those beer are all from the Miller Brewing company. It was not surprising since the Miller Brewing located in Milwaukee, and the stores are scattered in Chicago area, so customers may prefer the local brand and change their behavior elastically when receiving the promotion.

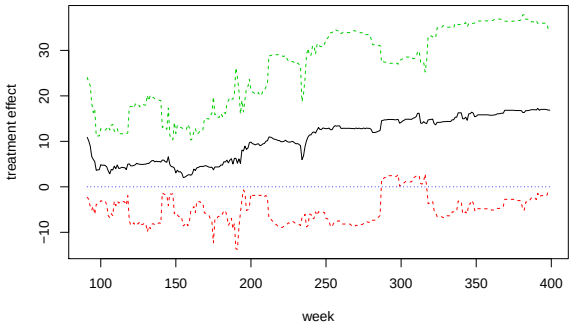

Figure 6: Treatment Effect over time

Treatment Effects over Time

The treatment effects increased gradually as time pass by. It seems there is nothing we can manipulate to boost the treatment effect in a short period of time. Despite of this, when we examine the treatment effect, we could find the treatment effects goes in slight waves, higher in summer times, also we could find that promotions would lower the profit (negative treatment effects) during those weeks have festivals, hence, we should not implement promotions during festivals.

The figure 7 illustrates the treatment effects over prices and weeks separately, it accords to the regression plot above.

Treatment Effects on Beer Brands

Applying promotions to certain brands of beer also brings a significant profit improvement. As illustrated in figure 6, those brands labeled at around 130 are influenced by promotions the most, those beer are all from the Miller Brewing company. It was not surprising since the Miller Brewing company is located in Milwaukee, and the stores are scattered in Chicago area, so customers may prefer the local brand and change their behavior elastically when receiving the promotion.

Treatment Effect over Types

From figure 9, we can find for some of the types, the treatment effects are strong, and they are labeled at around 15, which are brown ale, dark ale, dark dry. Also, we could find types labeled as 60 to 70 have negative treatment effects, those are lager and light beers.

The figure 10 depicted the treatment effect over brands and types:

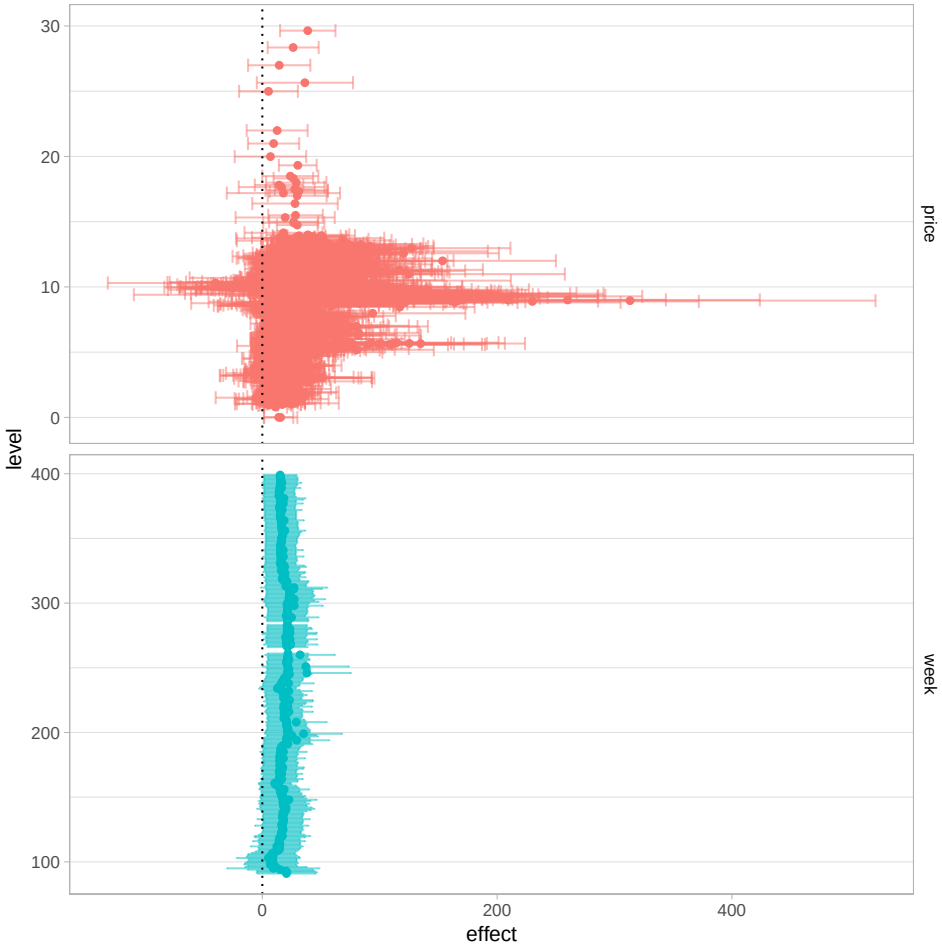

Figure 7: Treatment Effects by week and price

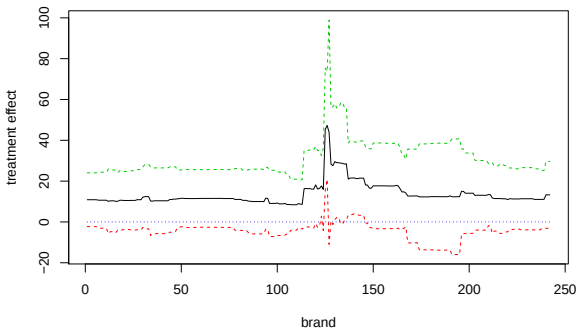

Figure 8: Treatment Effect over Unit Brands

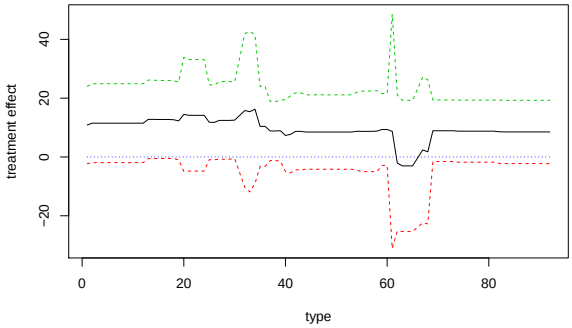

Figure 9: Treatment Effect over Unit Types

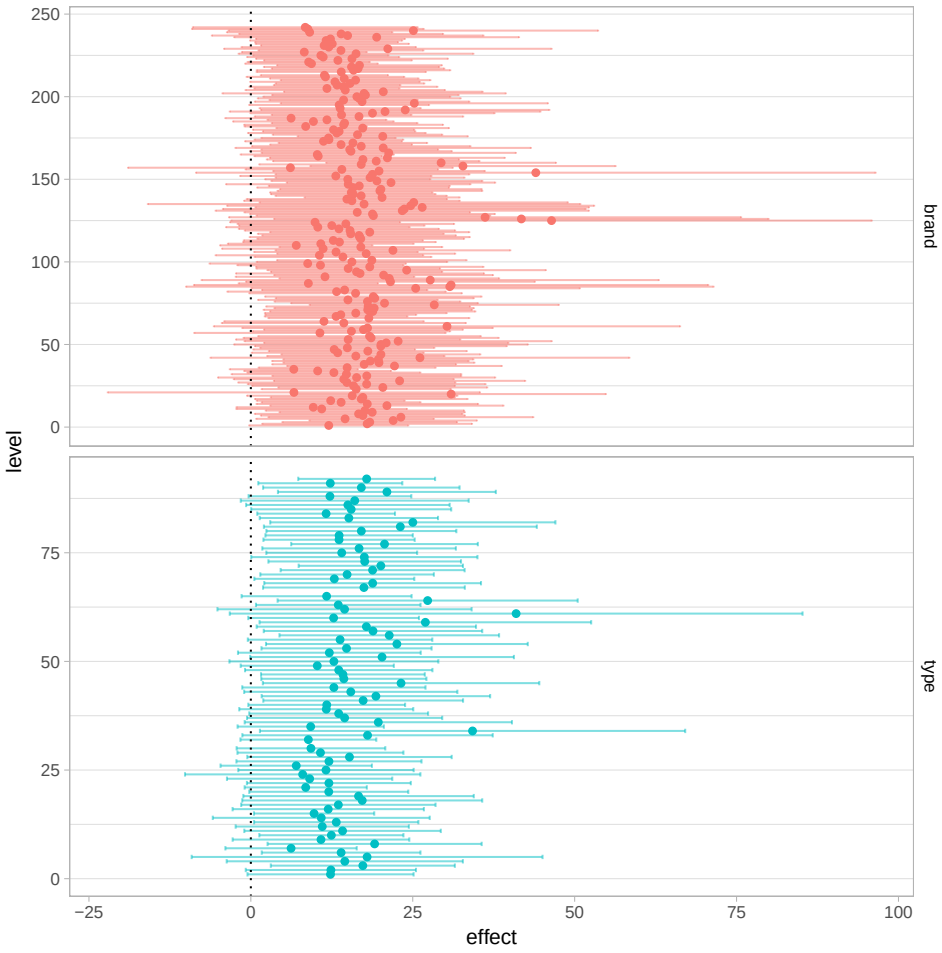

Figure 10: Treatment Effect over Unit Types

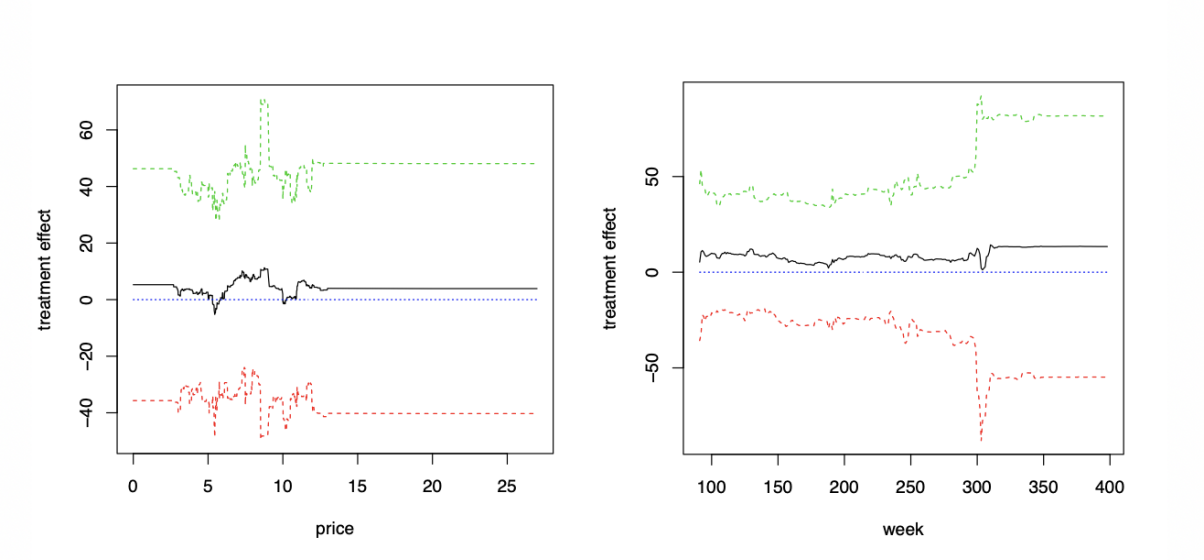

Figure 11: Treatment Effect by Random Forest Prices and Time

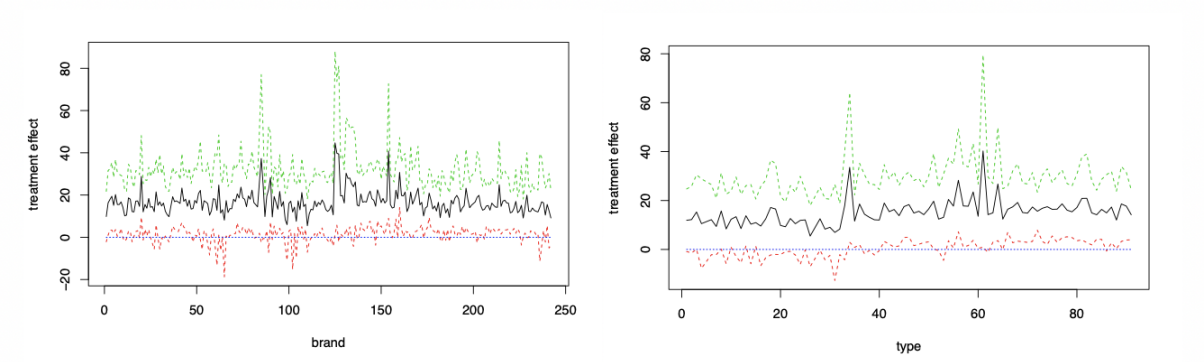

Figure 12: Treatment Effect by Random Forest Brands and Types

Supplement Results by Random Forest Method

The Random Forest Method generated roughly the same results as the Causal Forest on our data. However, the Random Forest generated a fuzzy treatment effects over the price and time (as showed in figure 11). The treatment effect over brands and types are similar to that of the Causal Forest method, but less significant (showed in figure 12). So, the Causal Forest might be superior.

Conclusion

By the regular linear regression, we could only find a coefficient for variables and it does not describe the trend for some local change. From the Causal Forest method, we could conclude that more promotions should be used during the summer, on the brown ale and black beers. Also, it is better to apply promotions on those beers priced between 6 to 8 dollars per unit. We should avoid to add promotions on festivals, on Lager and light beers, and those beers priced over 8 dollars is not suitable for any promotions.

Bibliography

[1] "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", Stefan Wager, Susan Arthey, July, 2017

[2] "Classification and Regression Trees", Friedman, Olshen, Stone, 1984

[3] "Gene selection and classification of microarray data using random forest", Ramon Diaz-Uriarte, Sara Alvarez de Andres, 2006

[4] "An assessment of the effectiveness of a random forest classifier for land-cover classification" V.F. Rodriguez-Galiano et.al, 2011