CS680 Project Report: ML Application to Model Heterogeneous Treatment Effect

Kyu Min Shim
School of Computer Science
University of Waterloo
kmshim @uwaterloo.ca

May 11, 2023

Abstract

Conditional Average Treatment Effect (CATE) measures different treatment effects for various subgroups within a population. When there exist covariates with unknown and potentially complex relationships to the outcome, CATE is able to provide more sensible results than when the population is considered homogeneous with respect to the treatment. Machine learning is an option for measuring CATE due to its flexibility and lack of necessary assumptions. In this paper, we implement a machine learning algorithm to quantify CATE on a diamond pricing data set, where the interest lies in determining the increase in the price of a diamond when its cut is improved from premium to ideal, conditional on its carat, color, and clarity, all of which are important factors that determine the price of the diamond. The resulting CATE demonstrated weak results, where the CATE seems to improve slightly with the overall quality of the diamonds. That is, the increase in diamond price when its cut is improved from premium to ideal is increased further for diamonds with high carat, good color, and good clarity.

1 Introduction

An A/B test is a type of experiment that is heavily utilized in both industry and academia to draw conclusions about causal relationships. In this setting, experimental units (a sample from the population of interest) are randomly assigned to one of two groups: the units in the control group (A) remain as they are, while the units in the treatment group (B) experience some change (treatment). A metric of interest is measured in both groups to quantify the causal impact of the treatment, and such causal impact may be generalized to the general population. Companies such as Google, LinkedIn, and Microsoft each run hundreds of online A/B tests on millions of users every day, quantifying the impact of changes to their platforms on key business metrics [1].

It is common to assume homogeneous treatment effects in A/B tests. That is, the treatment effect is the same across every unit in the population. In this case, the interest of the A/B test lies in measuring the average change in the metric of interest when the treatment is applied. This is commonly referred to as the Average Treatment Effect (ATE). However, the treatment effect may vary across different subgroups or segments within the population. When interest lies in estimating a heterogeneous treatment effect, the goal of the A/B test shifts to measuring the conditional average treatment effect (CATE), which is the average change in the metric of interest conditional on the covariates of the unit. If we can predict the outcome (metric of interest) given the covariates (subgroups or segments within the population), we may isolate the change in the outcome resulting solely from applying the treatment. Such a relationship can be highly complex with a very large number of covariates, hence it is an appropriate target for machine learning models [2].

In this paper, we aim to apply machine learning to model heterogeneous treatment effects on a data set of diamond pricing. The data set contains information on over 50,000 different diamonds, including their price, 4C's, and dimensions [3]. The 4C's of a diamond refers to its carat, clarity, cut, and color. The 4C's play a large role in determining the price of the diamond. The carat, the clarity, and the color of a diamond are inherent to the diamond, and they cannot be altered. However, the cut of a diamond is determined by the cutter's skill in the fashioning of the diamond [4], hence is considered to be a treatment in this paper.

The heterogeneous treatment effect for this data set is defined as follows. The experimental units are the diamonds in the data set. The metric of interest is the price of diamonds. Treatment is defined as improvement in the level of cut from premium (the second highest level of cut) to ideal (the highest level of cut). The covariates are the remaining 3C's. The interest lies in measuring the CATE, which is the average change in the price of diamonds when the cut is improved from Premium to Ideal, conditional on the carat, color, and clarity of diamonds. The data set is highly structured and covariates such as color and clarity are categorical variables. Hence, it is suspected that tree-based machine learning algorithms will do a fine job in modeling the relationship between the diamond price and the 3C's, which will provide the basis to predict the CATE.

2 Related Works

Larsen et al. [2] describes a framework for the conditional treatment model which will be elaborated in detail in Section 3. Using the semi-parametric model from Robinson [5], Larsen et al. [2] considers measuring heterogeneous treatment effect to be ripe target for machine learning methods. Jacob [6] has applied this method to study the effect of microcredit availability on the amount of money borrowed using meta-learners such as Doubly-Robust and generalized random forest, and found positive treatment effect for all observations. Fan et al. [7] adapted this idea to model the effect of smoking during pregnancy on birth weight as a function of mother's age, and identified some, but weak, age-related heterogeneity. These studies were unable to identify strong evidence of heterogeneity in treatment effects.

This paper will apply similar concepts, but with a more structured and simpler data set of diamond pricing. The relationship between market price and product quality should be more direct (especially for a luxury good like a diamond) than the relationships modelled in the above papers. Hence, I hope to find strong evidence of heterogeneous treatment effect in this data set with the help of the highly predictive machine learning models. The result of this experiment is easily applicable in real-life as opposed to the above studies which are in the context of highly regulated and sensitive areas such as finance and health care. Note that the carat, the color, and the clarity of the diamond cannot be changed. Hence, given the 3C's of the diamond, the cutter can determine the expected increase in price when an ideal cut is applied to the diamond as opposed to a premium cut. Since an ideal cut is more precise and intricate in design, it would take additional resources such as time and effort compared to a premium cut. If the expected increase in price is small relative to the additional resources required for an ideal cut, then it would be optimal for the cutter to apply a premium cut instead.

3 Problem Formulation

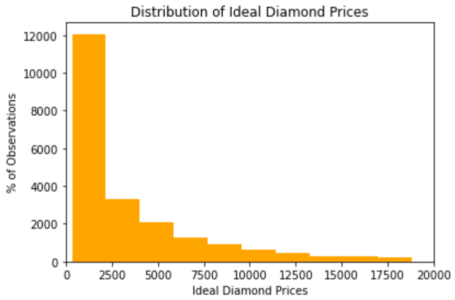
The problem formulation in this paper will follow the potential outcomes model from Robinson [5]. Let $Y_i(0)$ represent the outcome of unit i if i does not receive treatment and $Y_i(1)$ represent the potential outcome if i receives treatment. We define W_i as the binary treatment indicator for unit i. The expected outcome is $E[\frac{1}{N}\sum_{i=1}^{N} Y_i(W_i)] = \mu(\mathbf{W})$. The Average Treatment Effect (ATE) can be defined as $\tau = \mu(1) - \mu(0) = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1) - Y_i(0)]$. When interest lies in estimating the heterogeneous treatment effect, the inference is made on the Conditional Average Treatment Effect (CATE): $\tau(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$, where $\mathbf{X}_i$ represents a vector of covariates for unit i. Using the semi-parametric model from Robinson [5], we can model the relationship between the outcome and covariates as such: $Y_i = \tau(\mathbf{X}_i)W_i + g(\mathbf{X}_i) + \epsilon_i$, where ϵ_i is error term with mean 0 and no assumptions are made on the form of the functions τ and g. Conditional on the covariates $\mathbf{X}_i$, we define $m(\mathbf{X}_i) = E[Y_i|\mathbf{X}_i] = \tau(\mathbf{X}_i)E[W_i|\mathbf{X}_i] + g(\mathbf{X}_i)$. Finally, subtracting $m(\mathbf{X}_i)$ from Y_i yields $Y_i - m(\mathbf{X}_i) = \tau(\mathbf{X}_i)(W_i - E[W_i|\mathbf{X}_i]) + \epsilon_i$. Double machine learning from Chernozhukov et al. [8] used non-parametric regression methods to estimate $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$, but this paper will also use machine learning algorithms on a set of hold-out samples to estimate $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$. The estimate of τ is then $\hat{\tau}$ that minimizes squared error: $\frac{1}{N}\sum_{i=1}^{N}[Y_i - m(\mathbf{X}_i) - \hat{\tau}(\mathbf{X}_i)(W_i - E[W_i|\mathbf{X}_i])]^2$, which will also be found using machine learning algorithms. As mentioned in the Section 1, I propose the use of tree-based methods, specifically random forest, to model the above relationships since the diamond pricing data is highly structured.

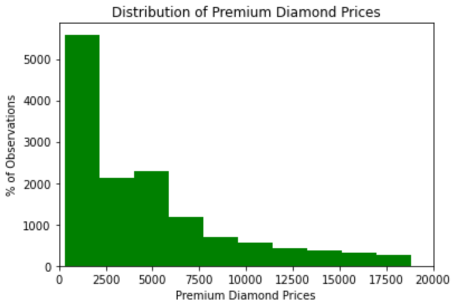
There are two assumptions that must be satisfied first [2]. The first assumption is the Stable Unit Treatment Value Assumption (SUTVA), which says that the outcome of each unit does not depend on other unit's treatment assignment. The second assumption is unconfoundedness; the potential outcomes and treatment assignment conditional on the covariate are independent: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i|\mathbf{X}_i$. The SUTVA assumption is satisfied by the diamond pricing data set, as the price of a diamond is based on its own 4C's and does not depend on whether the other diamonds have ideal or premium cut . The unconfoundedness assumption is also satisfied, since the potential diamond price for each level of cut $\{Y_i(1), Y_i(0)\}$ does not change based on the treatment assignment. Note that this is the *potential diamond price*, not the observed diamond price defined as $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$, which is dependent on the treatment assignment W_i.

Note that our interest lies specifically on diamonds with either premium or ideal cuts. About 40% of the data set (∼22,000) consists of ideal cut diamonds and about 26% (∼14,000) of the data set consists of premium cut diamonds. The data set will be randomly separated into three equal parts, where each part consists of one-third of ideal cuts and one-third of premium cuts in the data set. The hold-out samples to estimate $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$ will use the first part, training of $\hat{\tau}$ that minimizes squared error will use the second part, and finally testing of $\hat{\tau}$ will use the third part.

4 Main Results

We begin our analysis by visualizing the data. Below are the histograms of ideal and premium cut diamonds, and some summary statistics of diamond prices. If we only consider the cut as the factor that determines the price of the diamond, the mean, median, and histogram suggests that ideal diamonds have generally lower prices than premium diamonds. This contradicts the fact that ideal cut is of higher grade than premium cut.

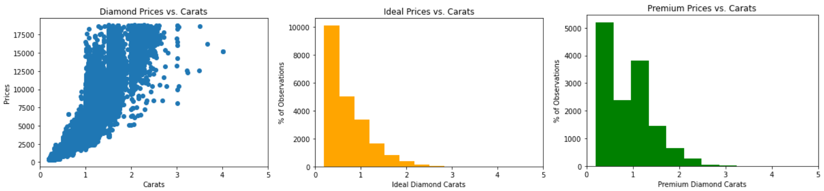




Statistic\Cut	Ideal	Premium
Mean	3457.54	4584.25
Median	1810	3185
Variance	3808.40	4349.20

Under homogeneous treatment effect, we assume that improving the cut level from premium to ideal will change the price of the every diamond by approximately equal amount. A/B test can be used to test the hypothesis: $H_0 : \mu_{Ideal} \leq \mu_{Premium}$ vs. $H_A : \mu_{Ideal} > \mu_{Premium}$, using a t-test to compare means of two samples. This t-test result in p-value of 1, thus there is basically no evidence against H_0 that states the average price of ideal diamonds are less than or equal to the average price of premium diamonds. Again, this result does not reflect the fact that ideal diamonds are of higher quality cut than premium diamonds, hence should be priced higher. This is likely due to the fact that there is a much larger sample size for ideal diamonds where most of ideal diamonds fall under low price range which drags down the average prices of ideal diamonds, and more importantly, we are not considering other factors that impact the price such as carat, clarity, and color.

For instance, a diamond's weight is measured by carat. If a diamond is high in carat, this implies that the diamond is large and heavy, hence it would result in higher price. From Diamond Prices vs. Carats plot below, there is a clear positive correlation between carats and prices. Observing the below histograms, it is also clear to see that there are, proportionally, more premium diamonds with greater than 1 carat than ideal diamonds. Therefore, it seems crucial to understand the relationship between the price of the diamonds and their 3C's in order to reliably determine the true change in price resulting from increase in cut level from premium to ideal.



We first prepare the data. From the original set of diamond prices, only the premium and ideal level diamonds are selected. We separate the ideal diamonds and premium diamonds, randomly divide each set into three equal parts, and put them together to create three equally sized sets with the same ratio of ideal to premium diamonds in each set as the original set. Since the color and clarity variables are categorical, one-hot encoding is used represent the different levels in color and clarity.

The first set is used as a hold-out samples to train the functions $m(\mathbf{X}_i) = E[Y_i|\mathbf{X}_i]$ and $E[W_i|\mathbf{X}_i]$. Using RandomForestRegressor and 80% of the hold-out samples, we train a random forest model to predict the price of diamonds based on its carat, color, and clarity. 10-fold cross validation is used to identify the number of trees and tree depth that maximizes R^2, which measures the proportion of variability in diamond prices that can be explained by the model. $0 \leq R^2 \leq 1$, and R^2 close to 1 means the model does a very good job at capturing variability in diamond prices.

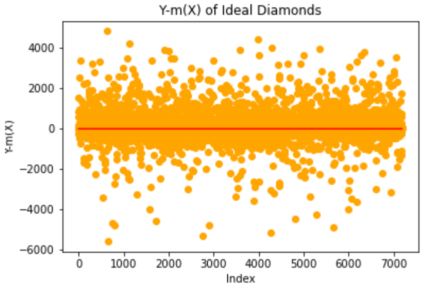
NumTrees\Depth	Max Depth	1	2	3	4	5
50	0.9752	0.6464	0.8396	0.8848	0.9119	0.9277
100	0.9753	0.6464	0.8396	0.8849	0.9120	0.9273
200	0.9752	0.6464	0.8396	0.8847	0.9120	0.9274
400	0.9753	0.6464	0.8397	0.8847	0.9120	0.9276
1000	0.9752	0.6464	0.8397	0.8848	0.9121	0.9277

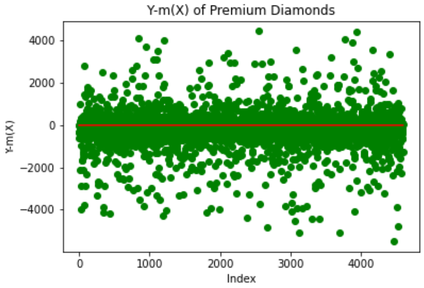
Using the optimal tree depth of maximum depth (trees are split until all leaf nodes contain one observation) and number of trees (100), a random forest model is trained on the 80% of the hold-out samples. 100 was chosen as the optimal number of trees instead of 400 as it had better R^2 at the fifth decimal place which is not shown in the above table. Then, the model is tested against the remaining 20% of the hold-out samples to test its accuracy in predicting diamond prices based on 3C's, which resulted in the test R^2 of 0.9757. This is very slightly higher than the R^2 observed in training.

However, considering that the data set is highly structured and the train set was made to be a very good representation of the test set through randomization, this result does not raise concerns.

The hold-out samples is then used to estimate $E[W_i|\mathbf{X}_i]$. Again, using 10-fold cross validation, we can find the optimal tree depth (5) and number of trees (100) result in R^2 value of 0.0669. This implies whether the diamond receives ideal or premium cut is not dependent on the value of 3C's. It is reasonable to consider the 3C's of a diamond and its cut as independent, as the carat or color or clarity of a diamond does not determine the cut it receives from the cutter. It is also reassuring to see that treatment assignment (even though we did not technically assign the level of cut) was randomized well across covariates, which will allow us to infer treatment effect conditional on the value of covariates. Now, with both the estimate of $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$, we may proceed to apply machine learning in finding the CATE τ.

From Section 3, we defined the estimate of $\tau(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1)-Y_i(0)|\mathbf{X}_i = \mathbf{x}]$ as $\hat{\tau}$ that minimizes the squared error: $\frac{1}{N}\sum_{i=1}^{N}[Y_i - m(\mathbf{X}_i) - \hat{\tau}(\mathbf{X}_i)(W_i - E[W_i|\mathbf{X}_i])]^2$. Such $\hat{\tau}$ also minimizes the following squared error $\frac{1}{N}\sum_{i=1}^{N}[(Y_i - m(\mathbf{X}_i))/(W_i - E[W_i|\mathbf{X}_i]) - \hat{\tau}(\mathbf{X}_i)]^2 = \frac{1}{N}\sum_{i=1}^{N}[(Y_i^* - \hat{\tau}(\mathbf{X}_i)]^2$. Using the training data, we first compute Y_i^*, find optimal parameters with 10-fold cross validation, and build a random forest model for $\hat{\tau}$. Unexpectedly, the best R^2 value from cross validation was less than 0.02. We can find explanation of this in the scatter plots of $Y_i - m(\mathbf{X}_i)$ below.





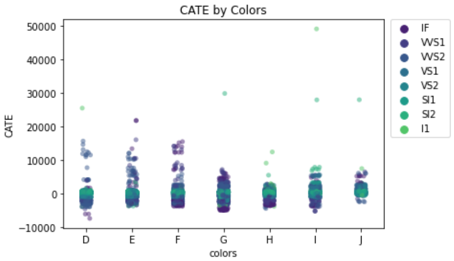
The $Y_i - m(\mathbf{X}_i)$ looks like random noise! Explanation of this can be found in our estimate of $m(\mathbf{X}_i)$. Recall that this model had R^2 value of 0.97 on our hold-out samples. Calculating R^2 on this training sample also results in the value of 0.97. That is, $m(\mathbf{X})_i$ explains so much of variability in Y_i, that $Y_i - m(\mathbf{X}_i)$ is practically random noise, and we are unable to fit a useful model for $\hat{\tau}$. In Chernozhukov et al. [8], non-parametric regression models were used to estimate $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$ instead of machine learning models. This may be due to the fact that machine learning models are too flexible, hence may extract too much information out of the data with $m(\mathbf{X}_i)$ and not leaving any behind for τ.

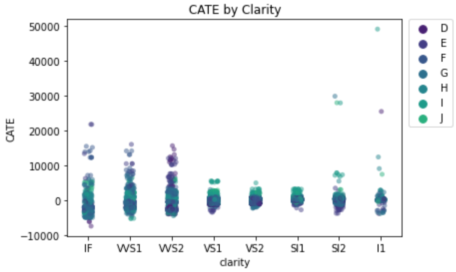
We try estimating $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$ using simple linear regression instead of random forest on the hold-out samples. The goal is to leave behind information for τ to use in estimating the CATE by using a less flexible model to estimate $m(\mathbf{X}_i)$. The R^2 of this model trained on 80% of the hold-out samples is 0.9182. It appears that, since the relationship between price and each of the 3C's is quite straight forward (heavier, clearer, colorless diamonds are more expensive), even a simple linear model is able to capture most of the variability in diamond price. On the remaining 20%, the trained model gives R^2 of 0.9214. Similarly, a linear regression model to estimate $E[W_i|\mathbf{X}_i]$ gives train and test R^2 of 0.0584 and 0.0627 respectively. Again, the train R^2 is better than the test R^2, but this can be explained by the fact that there is a very large number of observations and not as many features, and randomly splitting data into test and training ensures the training data is a very good representation of the test data.

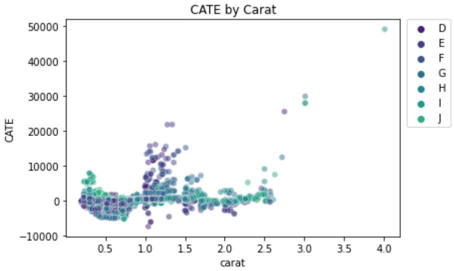
With this new $m(\mathbf{X}_i)$ and $E[W_i|\mathbf{X}_i]$, we can again try to fit a random forest model for $\hat{\tau}$ which minimizes $\frac{1}{N}\sum_{i=1}^{N}[(Y_i - m(\mathbf{X}_i))/(W_i - E[W_i|\mathbf{X}_i]) - \hat{\tau}(\mathbf{X}_i)]^2 = \frac{1}{N}\sum_{i=1}^{N}[(Y_i^* - \hat{\tau}(\mathbf{X}_i)]^2$. 10-fold cross validation is used to find the best parameters for $\hat{\tau}$, and the resulting R^2 is presented in the table below.

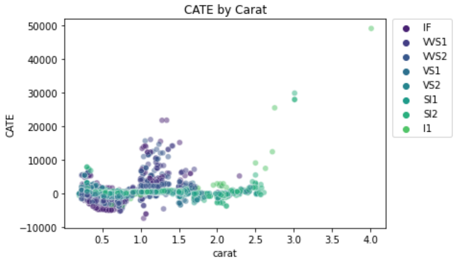
NumTrees\Depth	Max Depth	1	2	5	10	20
50	0.1449	0.0495	0.0746	0.1969	0.2515	0.1745
100	0.1479	0.0491	0.0753	0.1972	0.2514	0.1779
200	0.1510	0.0499	0.0751	0.1954	0.2519	0.1806
400	0.1511	0.0494	0.0747	0.1969	0.2524	0.1809
1000	0.1528	0.0488	0.0745	0.1985	0.2537	0.1826

These are definitely not the most impressive values of R^2, but it is much better than the previous R^2 when $m(\mathbf{X}_i)$ was estimated with random forest. On the full training set, a random forest model with depth of 10 and 1000 trees results in R^2 of 0.4415, and on the test set, the model results in R^2 of 0.2627. Let's visualize $\hat{\tau}$ with various plots.









It should be mentioned that the colors and clarity scales are in the order of best to worst. That is, the color D and J are the best and worst color of diamonds respectively, while the clarity of IF and I1 are the best and worst clarity of diamonds. As mentioned before, the greater the carat, the heavier the diamond hence more valuable. From all the plots, it appears that CATE is generally around 0. This is expected as $\hat{\tau}$ does not have the best predictive accuracy. In "CATE by Colors" plot, CATE appears to be higher for diamonds with good colors (D,E,F) relative to other diamonds, especially for those also with good clarity. Similarly, "CATE by Clarity" plot shows higher CATE for diamonds with good clarity (IF, VVS1, VVS2) compared to other ones, especially for those also with good colors. This implies that, the average increase in price of diamonds when cut is improved from premium to ideal is higher for diamonds with good color or clarity than for those with neither good color or clarity. Comparing the two "CATE by Carat" plots, there is high level of CATE for diamonds between 1 and 1.5 carats, and those diamonds also have good colors or clarity. Recall the distribution of carats in the data set is heavily skewed to the right, where most of the diamonds are below 1-1.5 carats and diamonds of carats greater than 2 are extremely rare. Hence, there are not enough observations to confidently demonstrate CATE in diamonds with high carats. However, from inspection of available data, it appears that CATE and carats have a positive relationship. Overall, it appears that the treatment effect of improving a diamond's cut from premium to ideal level varies across the values of 3C's. From the available data, it appears that CATE is amplified by the overall condition or rarity of diamond. That is, diamonds with high carat, good color, and good clarity increases in value at larger magnitudes compared to diamonds with low carat, poor color, and poor clarity.

5 Conclusion

CATE allows us to quantify treatment effect for various subgroups of a population. This is especially useful when covariates that greatly impact the outcome exist and cannot be controlled for. Machine learning is known to be a useful tool in measuring CATE due to its flexibility and lack of necessary assumptions. Using the diamond pricing data set, we implemented a simplified version of double machine learning technique [8] to quantify the increase in price of a diamond when the cut is improved from premium to ideal, conditional on its carat, color, and clarity, all of which are important factors that determine its price. The resulting CATE demonstrated weak results, where the CATE seems to slightly improve with the overall quality of the diamonds. That is, the increase in diamonds price when cut is improved from premium to ideal is increased further for diamonds with high carat, good color, and good clarity. This result maybe useful for diamond cutters, who may optimize work by prioritizing ideal cuts on good quality diamonds while applying premium cuts to more 'common' diamonds under resource constraints. However, it should be noted that the machine learning model for CATE did not demonstrate good predictive accuracy in training or testing. A problem remains where there is a trade-off between the predictive accuracy of $m(\mathbf{X}_i) = E[Y_i|\mathbf{X}_i]$ and $\tau(\mathbf{X}_i)$. When $m(\mathbf{X}_i)$ is highly predictive, it leaves only random noise for $\tau(\mathbf{X}_i)$ to model. For a data set where relationship between the covariates and the outcome is straight forward, even using a simple linear model for $m(\mathbf{X}_i)$ results in high predictive accuracy, which results in relatively poor predictive accuracy in $\tau(\mathbf{X}_i)$. This area is to be explored further in the future.

References

[1] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments a practical guide to A/B testing*. Cambridge University Press, 2020 (cit. on p. 1).

[2] Nicolas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T. Stevens. *Statistical Challenges in Online Controlled Experiments: A review of A/B testing methodology* (cit. on pp. 2, 3).

[3] Vittorio Giatti. *Diamond prices*. Oct. 2022 (cit. on p. 2).

[4] *4Cs of diamonds*. Oct. 2022 (cit. on p. 2).

[5] P. M. Robinson. "Root-n-consistent semiparametric regression". *Econometrica*, vol. 56, no. 4 (1988), p. 931 (cit. on pp. 2, 3).

[6] Daniel Jacob. "Cate meets ML - conditional average treatment effect and machine learning". *SSRN Electronic Journal* (Apr. 2021) (cit. on p. 2).

[7] Qingliang Fan, Yu-Chin Hsu, Robert P. Lieli, and Yichong Zhang. "Estimation of conditional average treatment effects with high-dimensional data". *Journal of Business amp; Economic Statistics*, vol. 40, no. 1 (2020), pp. 313–327 (cit. on p. 2).

[8] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. "Double/debiased/neyman machine learning of treatment effects". *American Economic Review*, vol. 107, no. 5 (2017), pp. 261–265 (cit. on pp. 3, 6, 8).