# PROJECT REPORT

# Machine Learning and Causality: Estimation of Heterogeneous Treatment Effect of Oil Price Shock on Consumer Price Inflation using Causal Forest

*A dissertation submitted to the University of Hyderabad*

*in partial fulfilment of the requirements for the degree of*

*Master of Arts in Economics*


By:

**Mayur Ingole**

**17SEMA07**

Under the guidance of:

**Professor Naresh Kumar Sharma**




SCHOOL OF ECONOMICS

UNIVERSITY OF HYDERABAD

HYDERABAD – 500046

INDIA


June 2020

# Declaration

**School of Economics,**
**University of Hyderabad,**
**Hyderabad – 500046**
**June 2020**

I hereby declare that the work embodied in this dissertation entitled "**Machine Learning and Causality: Estimation of Heterogeneous Treatment Effect of Oil Price Shock on Consumer Price Inflation using Causal Forest"** submitted to the University of Hyderabad in partial fulfilment of requirements for the degree of Master of Arts in Economics is original work carried by me under the supervision of **Professor Naresh Kumar Sharma**, Dean, School of Economics, University of Hyderabad.

I declare that to the best of my knowledge, this dissertation is free from plagiarism and no part of this thesis has previously formed the basis for award of any other degree, diploma, fellowship or any other similar title of recognition of any other University.

(Signature of the Candidate)

Name: Mayur Ingole

Reg. No: 17SEMA07

# Certificate

June 2020

This is to certify that the research work contained in this dissertation entitled "**Machine Learning and Causality: Estimation of Heterogeneous Treatment Effect of Oil Price Shock on Consumer Price Inflation using Causal Forest"** by **Mayur Ingole** bearing Registration No. 17SEMA07 has been carried out under my supervision.

The thesis has not been submitted previously, in part or in full, to this or any other University or Institution for a degree.

Professor Naresh Kumar Sharma
School of Economics
University of Hyderabad

# Acknowledgements

I would like to acknowledge my gratitude to Professor Naresh Kumar Sharma, my project supervisor, for his guidance and encouragement. I thank him for his time and efforts spent on making the completion of this project possible. I would also like to thank him for giving me the freedom to pursue my own research interests and patiently tolerating constant topic changes.

I am also very thankful to my parents and friends for their constant support through the duration of this project work. It has given me the necessary confidence to see this work through.

At last, I am very grateful for the University of Hyderabad and the amenities provided at the Reading Room of the IGM library, which made work in a comfortable environment possible for students.

# Table of Contents



## List of Figures

# Executive Summary

Name of the student: Mayur Ingole
Enrolment No: 17SEMA07
Degree for which submitted: Master of Arts in Economics
School: Economics
Thesis title: 'Machine Learning and Causality: Estimation of Heterogeneous Treatment Effect of Oil Price Shock on Consumer Price Inflation using Causal Forest'
Thesis supervisor: Professor Naresh Kumar Sharma
Month and year of thesis submission: June, 2020

## Abstract:

This dissertation applies the Causal Forest methodology in order to analyse the heterogeneity in treatment effect of oil price shocks on consumer price inflation across countries. For this the use of World Bank's World Development indicators database has been made. The average treatment effect is calculated along with its confidence interval. The variables used as features in the model are also ranked on the basis of their contribution to the heterogeneity in treatment effect. The relation between top four variables and predicted treatment effect is also analysed. Analysis of heterogeneity in treatment is carried out on the basis of variability in the ranked plotting on individual estimated treatment effect. Finally impact of a hypothetical oil price shock in 2019 on inflation in G-20 countries is also estimated. The limitation of the research due to the incomplete nature of the dataset is also acknowledged.

# Introduction

Machine Learning models are a collection of predictive models. These models are inspired from non-parametric statistical methods. They exploit the recent advancement in computation technology and the phenomenal increase in availability of data to greatly enhance their predictive accuracy. Machine Learning may help in increasing the accuracy of forecasting models in economics but major work of economists involves causal inference. For example: evaluating the impact of a particular policy on an economic indicator. Causal inference involves dealing with counterfactuals. For example: What would've happened in absence of a particular policy? This is where the recently expanding field of Causal Machine Learning may prove helpful.

The best way to have any kind of causal inference is to conduct a randomised control trial. In it the treatment is allocated randomly across half the members of a group and then the difference in outcome between the group which got the treatment and the ones which didn't is studied. But this is not always possible in economics. Most economists rely on observational studies for any kind of causal inference. Standard method involves assigning a dummy variable for the treatment and then regressing it on the outcome controlling for all possible confounding variables. Confounders are variables which are related to the outcome and have an impact on the probability of a particular observation belonging to the treatment group. Conditional on the confounders the treatment is as good as randomly assigned. This however requires accurately taking into account all confounders. Such an exercise can also be described as predicting the counterfactuals using the confounders and dummy variable.

For example: in Instrumental Variable regression the set of potential instrumental variables can be large and sometimes even exceed the number of observations. In this case it is a big question as to which variables to select to construct an instrumental variable estimator. Including all variables may lead to over fitting and spurious results can be given interpretation of causality. While excluding some variables may result in biased estimate of treatment effect.

This traditional econometrics model requires specification of model a priori. If the set of independent variables to choose from is large then selecting through them becomes arbitrary. Machine Learning models are good at detecting complex patterns in large datasets. Random Forest is one of the most popular Machine Learning model as it can handle all kinds of variable be it numerical, categorical or binary and also is able to select through a big set of independent variables in order to enhance predictive power of the model.

Machine Learning model are designed to maximize accuracy of out of sample fit. Therefore the optimization function includes a penalty term for in sample over fitting. The accuracy of the model is judged based on the holdout validation set which is not used in training the parameters of the model. Traditional econometrics focus on in sample fit. Traditionally predictions and causal inference are treated as two different things. It is mirrored by the difference between econometrics and Machine Learning about explanation and prediction.

But many causal inference problems can be broken up into different step some of which include predictions. As we have already noted, Causal Inference can also be described as predicting the counterfactual. Also most of the causal machine learning models estimate average treatment effect however what is needed is to account for heterogeneity between subjects which leads to heterogeneous treatment effect.

One such model which harnesses the predictive power of Machines Learning to the use of Causal Inference is Causal Forest. Causal Forest is based on a popular Machine Learning model called as Random Forest. Random Forest is based on decision trees. The decision tree divide the feature space into different group based on binary yes/no questions. Every node of this tree is a group over which the outcome is averaged and considered as output of for any input feature getting classified into that group. The loss function for the decision trees in Random Forest usually minimizes the error calculated from the predicted outcome and the actual outcome. This error is usually RMSE i.e. Root Mean Square Error. But in Causal Forest the loss function and the objective function is modified. Here the objective is to maximize the difference in the outcome for the untreated and treated observation within groups while also ensuring that the observations belonging to same group are as similar to each other as possible.

In this paper we will use the Causal Forests model to analyse the impact of oil price shocks on consumer price inflation across countries. Our treatment will be binary variable for oil price shock where we will define a price shock in the case when for any month in a year the price increase is more than fifty percent from the price for the corresponding month in previous year. Our outcome variable is the annual consumer price inflation. Our feature matrix will consist of all the possible indicators from the World Development Indicator Database of the World Bank.

After estimating the Causal Forest we will estimate the average treatment effect. We will also rank variables based on their contribution to the heterogeneity in treatment effect and select and display top 20 variables. We will try to infer something about the hypothesis of heterogeneity in treatment effect across the observations based on the variability in the graph of estimated treatment effect plotted rank wise. Then we will estimate the treatment effect for G20 countries for the year 2019.

# Literature Review

The literature referred to for determining the methodologies and research objective for this project are as follows.

The primary inspiration for this study from which it heavily borrows is Andrew Tiffin (2019) IMF working paper "Machine Learning and Causality: The Impact of Financial Crises on Growth". The paper begins by explaining the basic premise for applying Machine Learning to the problem of causal inference. It defines the problem of causal inference as the problem of predicting the counter factual. This paper makes pioneering use of the Causal Forests model in order to estimate the impact of a hypothetical financial crisis on a country's GDP growth rate. This study uses IMF's Vulnerability Exercise Database. The dataset includes data on 46 variables ranging from 1985 to 2017. The database covers 107 countries from both emerging markets and OECD economies. It has over 3364 observations. The study also analyses the impact of different variables on treatment effect using Shapley values. Exchange rate flexibility is found out to be the most important confounding variable determining the impact of a financial crisis on growth rate. This is essential in order to determine which variable is to be targeted for intervention in order to reduce the vulnerability of economic growth to a financial crisis.

Susan Athey and Stefan Wager in their 2019 paper "Estimating Treatment Effects with Causal Forests: An Application" applied the Causal Forest model to a dataset simulated from the National Study of Learning Mindset database in US. It is primarily a demonstration of the applicability of Causal Forest models to real world problem. It tried to study the impact of a policy intervention on learning achievements in schools in US. It found out that detection of treatment heterogeneity through the application of Causal Forest is best when the data has a clustered behaviour i.e. the observations form clusters. The application of Causal Forest model in order to determine effect of intervention on learning outcomes in schools in US based on the simulated data doesn't reveal any substantial heterogeneity in treatment effect.

Zhao, J. and others (201)7 in their paper analyse the impact of World Bank projects on forest cover using the Causal Forest model. They use a geolocated dataset of World Bank projects, and expand this dataset with satellite mapped characteristics of their geographic features. The treatment in their paper was World Bank Projects and the outcome was forest cover. They found that the impact for most projects was nearly zero but there were few exceptions. They compared the Causal Forest model with other methods and concluded that the method selected can have substantial effect on the estimation of treatment effect.

# Causal Forest

A Causal Forest is based on Random Forest models therefore we need to understand them first before moving on to Causal Forest.

Random Forests are machine learning models which are based on decision trees. A decision tree works by the repeated partitioning of the feature matrix into two sets, starting with an initial split that decreases the error the most. The algorithm considers every possible split on every possible variable. It then chooses the one split on the one variable that best separates the sample into the two most dissimilar subsamples which is based on the predicted outcome. These binary partitions then continue until the termination of the tree. They are recursive—i.e., each subsequent split only considers the sub-sample under which it falls, rather than the whole dataset. The result is an efficient set of yes or no type questions that can sort any individual instance into an appropriate group of observations which are similar.

A regression tree is a type of decision tree. The purpose of its design is to approximate a continuous real-valued function. Essentially, by sorting the dataset into groups of observations which are similar, it gives a non-parametric estimate of the expected outcome for any individual within that group.

Decision trees are computationally more efficient as compared to other Machine Learning models. They work well when there are nonlinearities and interactions between independent variables. They do not work as well if the relationship is linear. Even then they can discover aspects of the data that are not obvious if we use traditional linear econometric approach.

The Random Forest is an ensemble of the decision-trees. The purpose is to minimize the problem of overfitting. Decision trees have problem of overfitting. They give models that fit the training sample extremely well. But they often perform very poorly when making out-of-sample predictions. Shortening the tree by imposing a penalty for an overly long or complex tree is one of the solutions. Another solution is to not focus on a single tree and use the same data to generate numerous trees and then aggregate the results.

One way of aggregating the results is bootstrap aggregation i.e. “bagging”. In it we build an individual tree on a random sample of the dataset which can be between 1/3$^{rd}$ and 3/4$^{th}$ of the total number of observation. The remaining observations are referred to as out-of-bag samples. They are used to gauge the accuracy of the tree. This is repeated thousands of times. When the random forest model is asked to predict the most likely outcome for new input then it will feed that input through each of these

thousands of individual trees and it will then aggregate their predictions by taking an average.

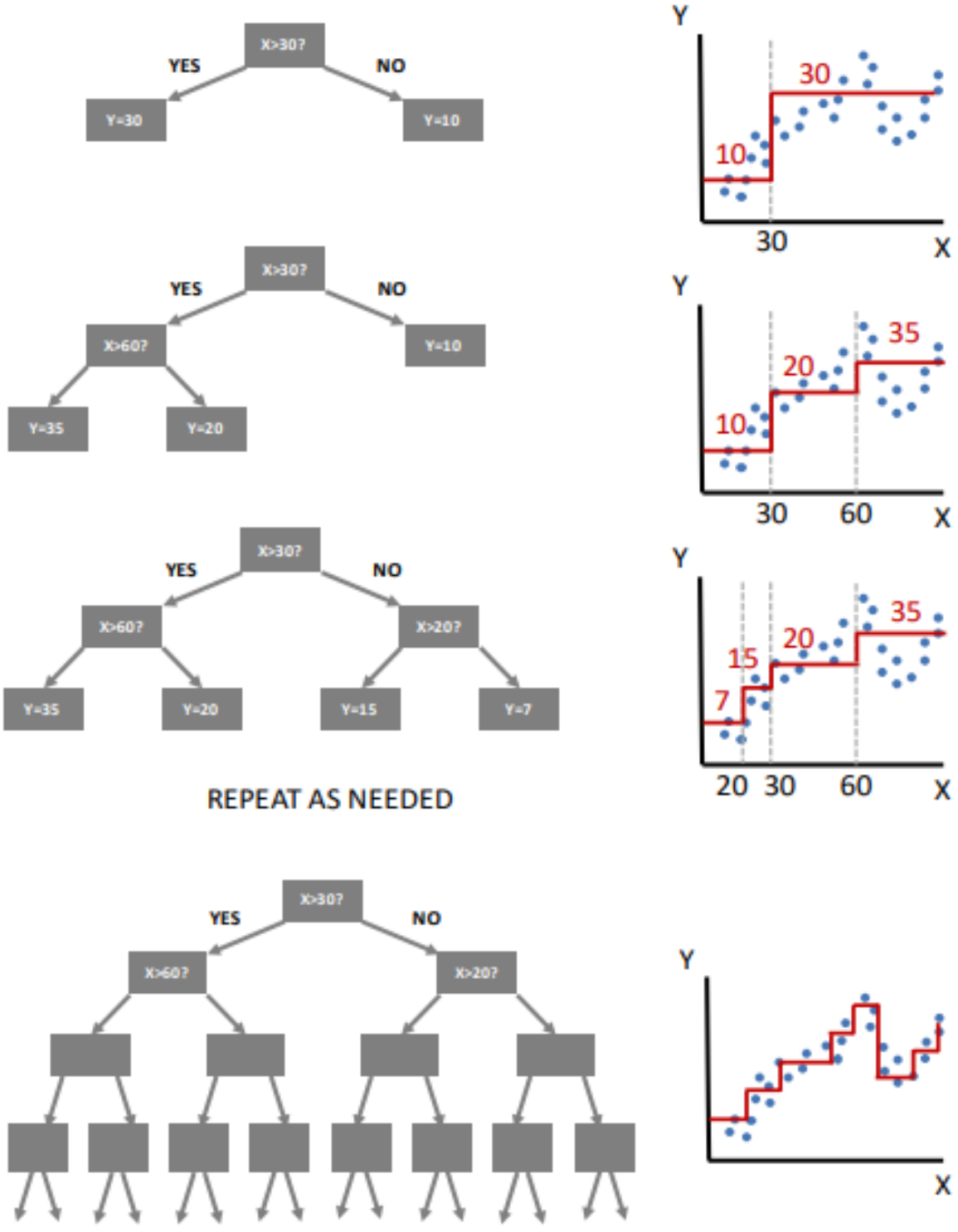


Figure 1: Evolution of a Decision Tree

As none of the tree is pruned it means that each individual tree is a weak model. It will have difficulty in distinguishing signal from statistical noise. On the other hand a large ensemble of individual trees will exploit the law of large numbers which will average out the noise and thus help the model in deciphering the signal from the data. Another modification which can be made is to take a random sample of the set of features at each split. In the case of highly correlated features, and particularly in the event of a single driving feature, bagging by itself can be insufficient, as it may simply produce multiple versions of the same tree. To get around this problem, random forest introduces an added element of randomization—at each split, the algorithm only considers a random subset of the available set of features. By randomizing the features, the random forest model effectively ensures that diverse trees will go into the final collection. As it is grown on a limited dataset, each tree on its own will be a weak model. But by combining a large number of weak models, we can end up with an aggregate prediction that is accurate.

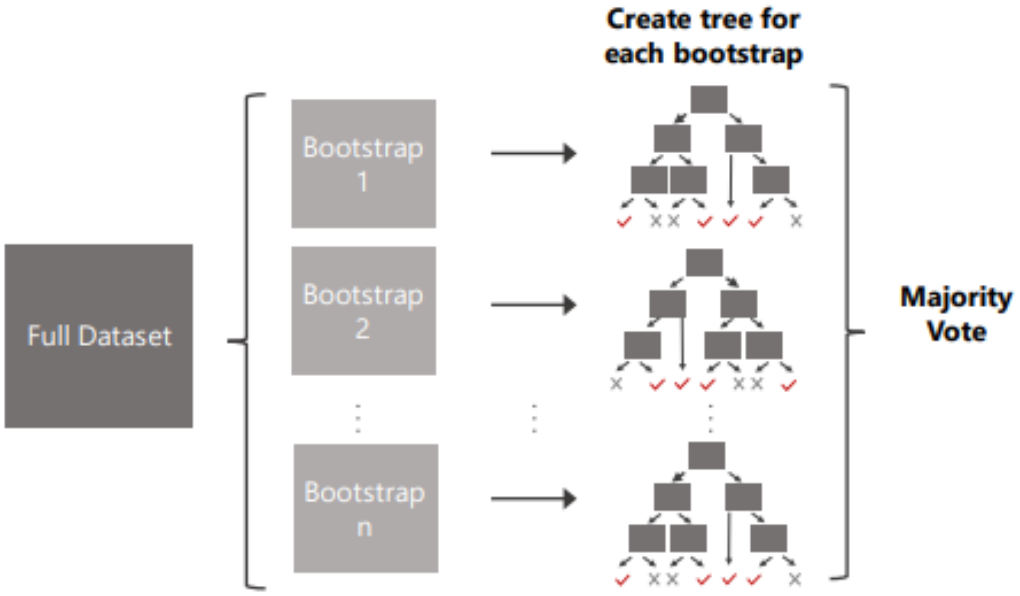


Figure 2: Bootstrap Aggregation

A Causal Forest is a Random Forest with a modified optimization function i.e. the loss function. It is made up of honest causal trees. The loss function in Random Forest usually minimizes the root mean square error of the predicted outcome. But in Causal Forest the focus is on the impact of a treatment on outcome for an individual observation. This impact can't be estimated directly as a single observation cannot both be treated and untreated. Therefore the Causal Tree focuses on the average difference in outcomes between treated and untreated observations within each node of the tree. The splitting rule of Causal Tree has two objectives:

1. To fine the splits where the difference in treatment effect is maximum.

2. To maximize the accuracy of the estimate of the treatment effect.

Each node of a causal tree can be considered as an artificial experiment. In it the observations of the experiment are as similar as possible. On the other hand the average treatment effect for that group predicts the individual effect for future observations with the same feature and being classified into the same group by the model.

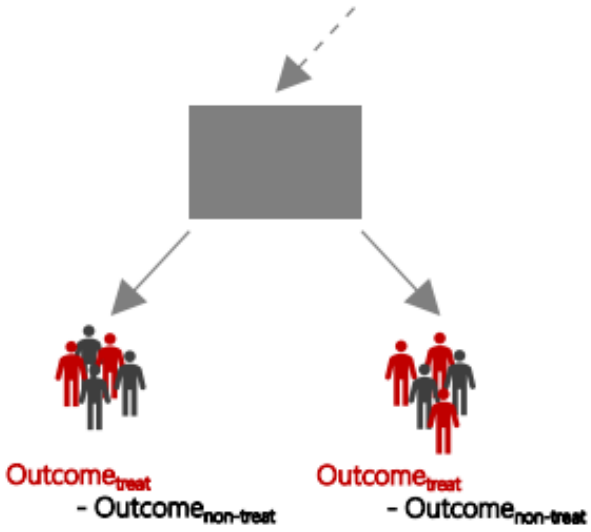


Figure 3: Leaf of a Causal Tree

Estimated Causal Trees honest is made honest by using different data to partition the tree and different data to estimate the average impact. Half of the data is used for determining the splits for the tree, and other half for filling that tree with observations and then estimating the treatment effect for each group. Random Forest model builds thousands of individual trees, and uses a new bootstrap sample for each tree; so ultimately the model uses all the data for both splitting and estimation.

# Data & Methodology

For our analysis we use the World Development Indicator Database of World Bank. This database originally contains 1,600 time series indicators for 217 countries and almost 40 groups of countries, with data for many indicators having range of more than 50 years. Our dataset contains 16104 observations of 1433 variables ranging from the year 1989 to 2019. This dataset has lot of missing values. Therefore we remove all the variables where the proportion of missing values is more than half. We also remove all the inflation related variables except consumer price inflation which is output i.e. predicted variable of our model. We use the linear interpolation method to impute missing values in our dataset. After cleaning and imputation of missing values our final dataset has 7392 observations of 332 variables. In all our dataset contains 2454144 data points.

For our treatment which is oil price shock we use historical monthly price data of Brent crude which is used as a benchmark to set the price of two-thirds of the worlds internationally traded crude oil supplies. This data is sourced from the Energy Information Administration, US. We define our treatment = 1 if in a year the increase in monthly price for any month is more than 50% for the corresponding month in previous year otherwise we define treatment = 0. We include on positive shocks and exclude all negative shocks. The range of price data is from 1987-2020. Treated observations make nearly 7 % of total observations.

Now Methodology:

First we divide our dataset into two parts. The train set which will be used to train the causal forest and the test set which will be used to judge the accuracy of the model. We include nearly 60% of our observation in train set and remaining 40 % of observation in test set. Our Causal Forest will be an ensemble of 5000 Causal decision trees. After estimating the Causal Forest first we calculate the average treatment effect along with its standard error which will allow us to judge its accuracy. Then we calculate the contribution of each variable to the heterogeneity of treatment effect in the outcome and then select the top 20 variables. We then judge the heterogeneity of the treatment effect across observations. We also predict the effect of a hypothetical oil price shock in 2019 on consumer price inflation in the countries included in G20 group.

We use R software for the purpose of data analysis and visualization the code for which can be found in the appendix. We use the “tidyverse” packages to clean the dataset and prepare it for analysis and for visualization. We use the “grf” package (Generalized Random Forest) to estimate the Causal Forest. All the visualizations to be followed are made using R.

# Results

First let us have a look at picture of one of the 5000 trees generated in the ensemble of our Causal Forest model.

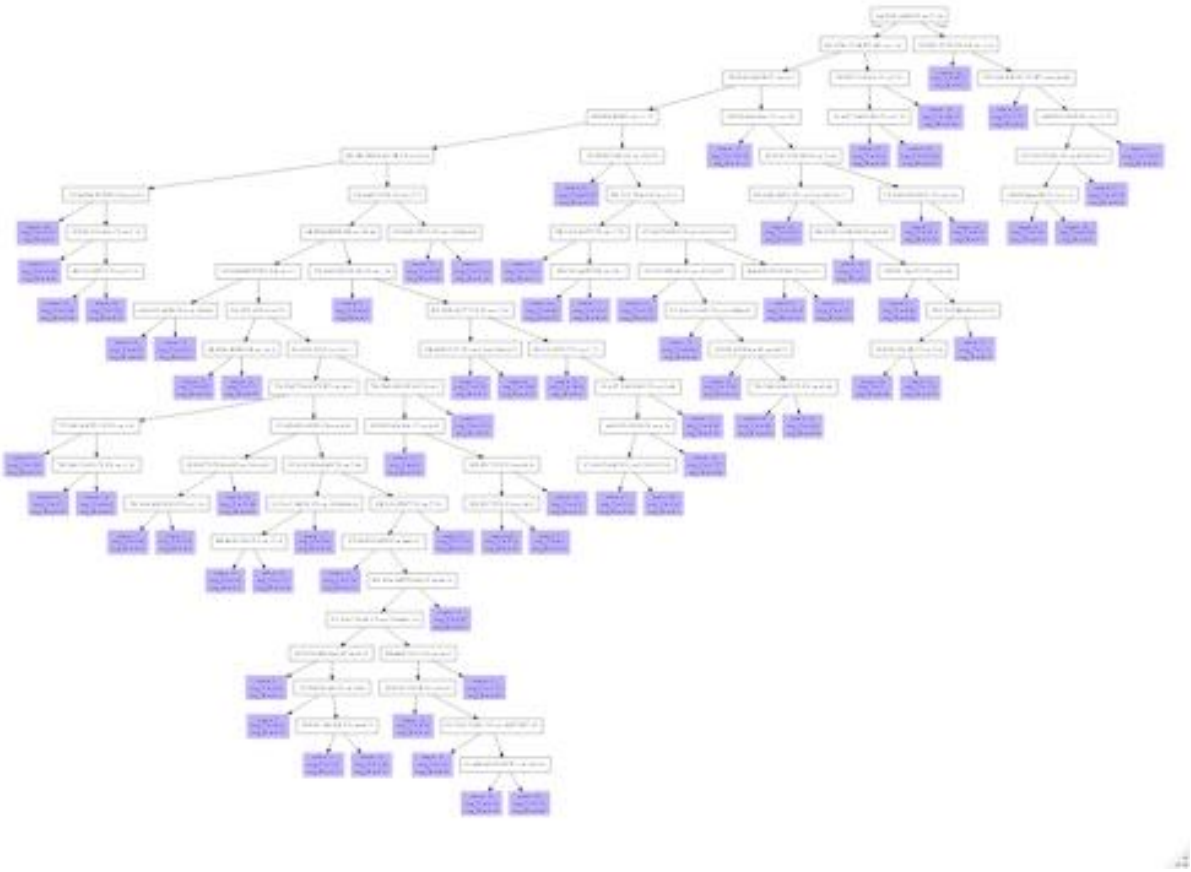

Figure 4: Generated Decision Tree

The calculated absolute average treatment effect for our model along with standard error is:

| Estimate | std.err |
|---|---|
| 9.440614 | 6.572544 |

Therefore 95% Confidence Interval for the average treatment effect is: 9.441 +/ - 12.882.

As we can observe the standard is very big and consequently the confidence interval of the estimate of average treatment effect. This is an indication that our model is inaccurate in terms of predicting the average treatment effect.

Now we will have a look at the table of top 20 variables in terms of their importance in determining the heterogeneity of the treatment effect.

The variable importance is a simple weighted sum of how many times the variable was split on at each depth in the forest.

| Sr. | Variable | Contribution |
|---|---|---|
| 1 | Domestic credit to private sector (% of GDP) | 0.0306471 |
| 2 | GNI per capita (current LCU) | 0.0298757 |
| 3 | GDP per capita (current LCU) | 0.0277108 |
| 4 | Adjusted savings: carbon dioxide damage (% of GNI) | 0.0252142 |
| 5 | Domestic credit to private sector by banks (% of GDP) | 0.0229724 |
| 6 | Merchandise trade (% of GDP) | 0.0227717 |
| 7 | Imports of goods and services (current LCU) | 0.0226300 |
| 8 | Manufactures imports (% of merchandise imports) | 0.0223684 |
| 9 | CO2 emissions (kg per 2010 US$ of GDP) | 0.0189828 |
| 10 | Exports of goods and services (current LCU) | 0.0172995 |
| 11 | Industry, value added (annual % growth) | 0.0154223 |
| 12 | Transport services (% of commercial service imports) | 0.0146977 |
| 13 | GDP per capita (current US$) | 0.0141336 |
| 14 | GDP (current LCU) | 0.0140198 |
| 15 | GNI (current LCU) | 0.0140136 |
| 16 | GNI per capita, Atlas method (current US$) | 0.0134743 |
| 17 | Population, female (% of total population) | 0.0122307 |
| 18 | General government final consumption expenditure (% of GDP) | 0.0116054 |
| 19 | GDP per capita growth (annual %) | 0.0114466 |
| 20 | Population, male (% of total population) | 0.0112436 |

Now let us have a look at the World Bank's definition of top five variables in the following table.

| Sr. | Variable | Topic | Definition |
|---|---|---|---|
| 1 | Domestic credit to private sector (% of GDP) | Financial Sector: Assets | Domestic credit to private sector refers to financial resources provided to the private sector by financial corporations, such as through loans, purchases of non-equity securities, and trade credits and other accounts receivable, that establish a claim for repayment. For some countries these claims include credit to public enterprises. The financial corporations include monetary authorities and deposit money banks, as well as other financial corporations where data are available (including corporations that do not accept transferable deposits but do incur such liabilities as time and savings deposits). Examples of other financial corporations are finance and leasing companies, money lenders, insurance corporations, pension funds, and foreign exchange companies. |
| 2 | GNI per capita (current LCU) | Economic Policy & Debt: National accounts: Local currency at current prices: Aggregate indicators | GNI per capita is gross national income divided by midyear population. GNI (formerly GNP) is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad. Data are in current local currency. |
| 3 | GDP per capita (current LCU) | Economic Policy & Debt: National accounts: Local currency at current prices: Aggregate indicators | GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current local currency. |
| 4 | Adjusted savings: carbon dioxide damage (% of GNI) | Economic Policy & Debt: National accounts: Adjusted savings & income | Cost of damage due to carbon dioxide emissions from fossil fuel use and the manufacture of cement, estimated to be US$30 per ton of CO2 (the unit damage in 2014 US dollars for CO2 emitted in 2015) times the number of tons of CO2 emitted. |
| 5 | Domestic credit to private sector by banks (% of GDP) | Financial Sector: Assets | Domestic credit to private sector by banks refers to financial resources provided to the private sector by other depository corporations (deposit taking corporations except central banks), such as through loans, purchases of nonequity securities, and trade credits and other accounts receivable, that establish a claim for repayment. For some countries these claims include credit to public enterprises. |

As we can observe the maximum contribution is of Domestic credit to private sector (% of GDP) which is almost 3 %. All of the variables' contribution is in single digit percentage which means none of the variables contributes in any significant way to the heterogeneity in treatment effect.

Now let us plot and analyse the top four variables which contributed most to the heterogeneity in treatment effect against the predicted values. The top four variables being: Domestic credit to private sector (% of GDP), GNI per capita (current LCU), GDP per capita (current LCU), and Adjusted savings: carbon dioxide damage (% of GNI).

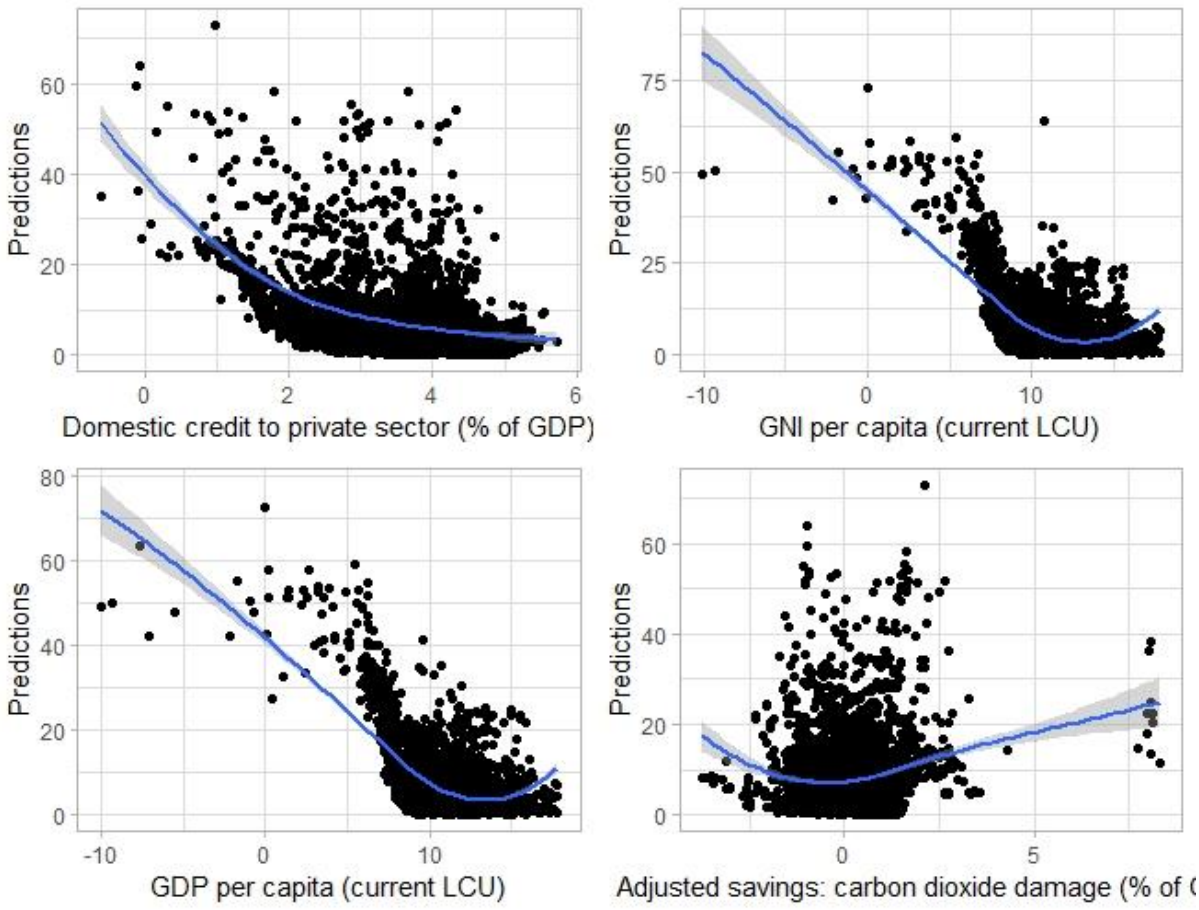


Figure 5: Top four Predictor Variables vs. Predictions

As we can observe the first three variables namely: Domestic credit to private sector (% of GDP), GNI per capita (current LCU), and GDP per capita (current LCU) have a negative impact on the treatment effect whereas the fourth variable Adjusted savings: carbon dioxide damage (% of GNI) has a positive impact on the treatment effect.

Now we will analyse the heterogeneity in treatment effect i.e. whether oil price shocks have heterogeneity in treatment effect on consumer price inflation across countries or not. For this we will plot the predicted treatment rank wise with the treatment effect on Y axis and ranks on X axis. (Figure 6)

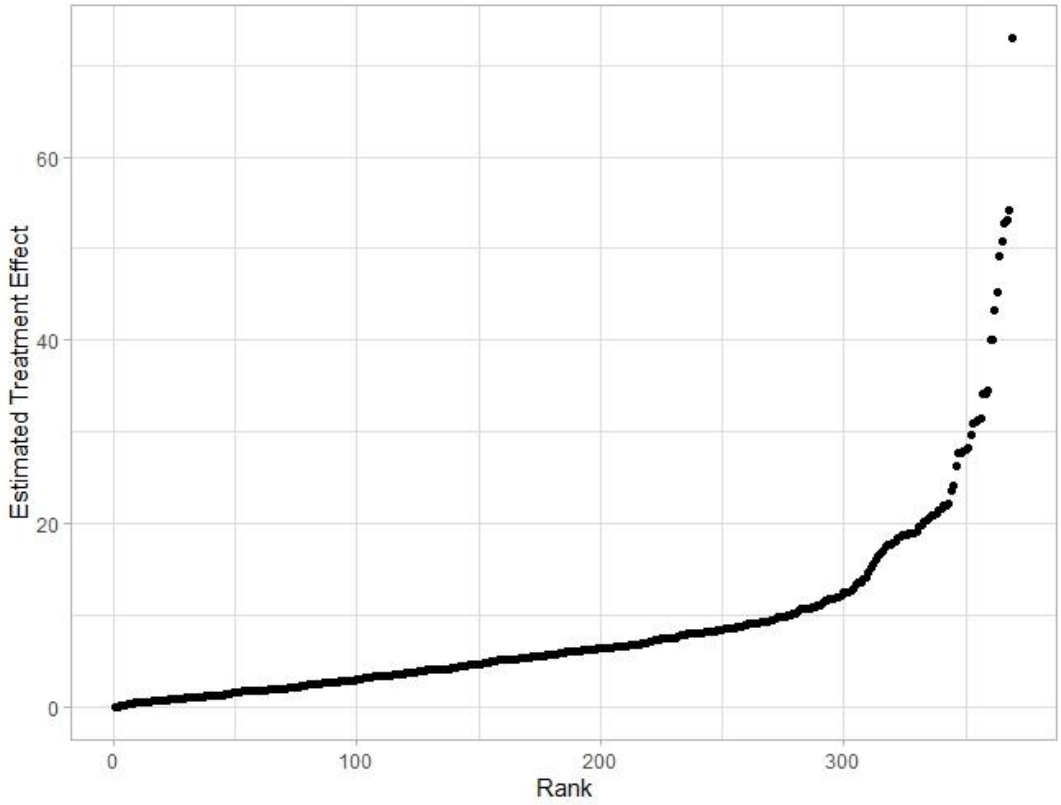


Figure 6: Rank wise Treatment effect

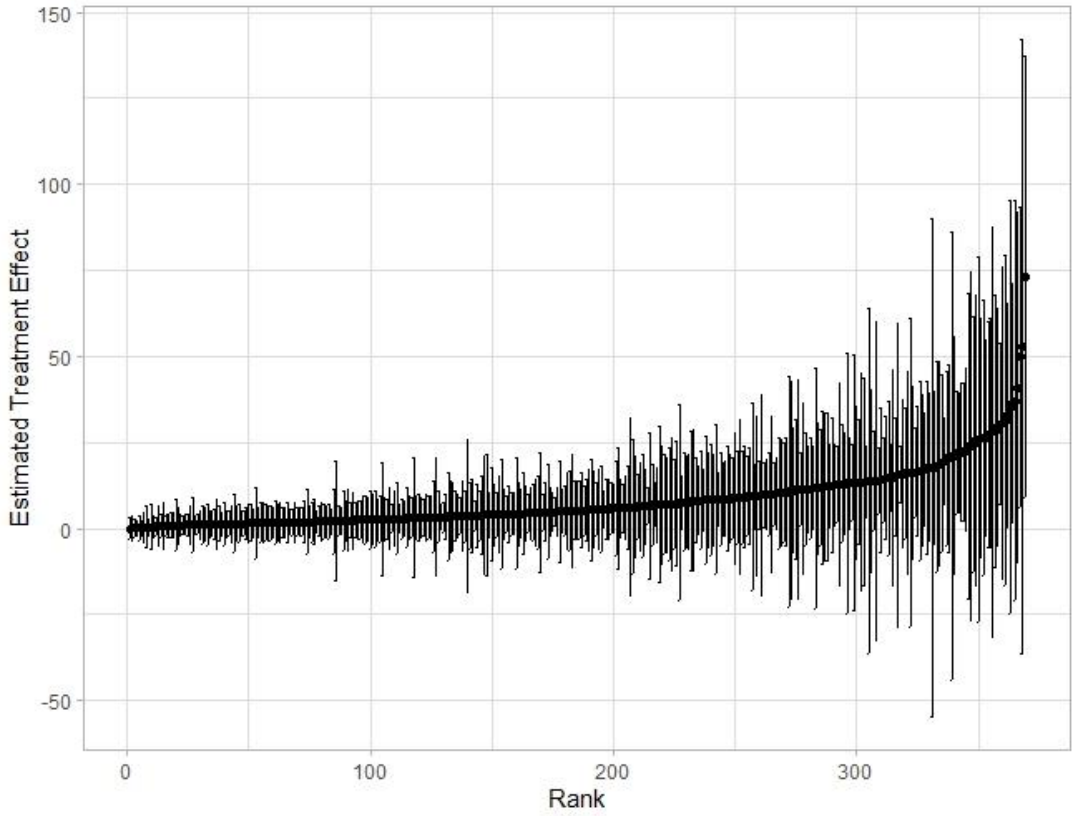


Figure 7: Rank wise Treatment Effect with Confidence Interval

There is not a lot of variability in estimated treatment effect for lower level ranks and the variability in treatment effect increases as the rank increases after 300.

Since the estimated treatment effects in Causal Forest are asymptotically normally distributed we will plot these treatment effects along with their 95% Confidence Interval in order to have a batter grasp of the uncertainty surrounding their estimation.(Figure 7)

As we can observe from the figure the uncertainty surrounding the estimation of treatment effect increases as we increase the ranking of the effect i.e. the absolute value of the treatment effect increases.

Now we can have a look at the estimated effect on consumer price inflation of a hypothetical oil price shock on G20 countries.

For this we pass the feature vector for the year 2019 for each of this country to our estimated Causal Forest. We will then plot the estimated effect in a bar plot to compare them.

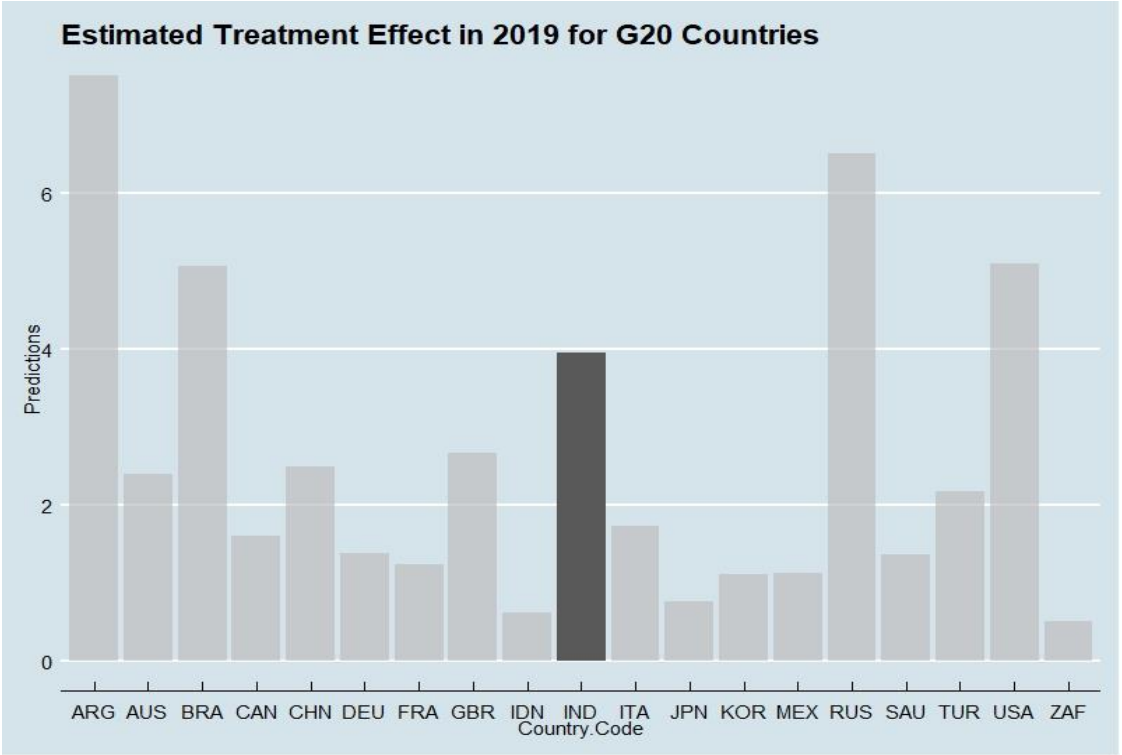


Figure 8: Treatment Effect for G20 Countries in 2019

As we can observe Argentina has highest effect on its inflation of oil price shock at 7.499599. It means if the increase in oil price was more than 50% for any month in that year. Then Argentina is predicted to have an increase in consumer price inflation by almost 7.5. For India the effect is 3.952644 which is above 2.595646 i.e. the mean effect for G20 countries. The minimum estimated effect is for South Africa which is 0.5106252.

# Conclusion

We applied the Causal Forest model in order to study the heterogeneous treatment effect of oil price shock on consumer price inflation. For the feature matrix which was used to classify countries in similar groups in order to calculate group wise treatment effect we used the World Bank's World Development Indicator database. We estimated a Causal Forest with 5000 decision trees using the "grf" package in R. Then we calculated the average treatment effect. The average treatment effect had big standard error therefore its accuracy was deemed to be unreliable. Then we calculated the contribution of variables to the heterogeneity in treatment effect. The contribution of all the variables was well below in single digit of percentage leading us to the conclusion that no one variable contributed significantly to the heterogeneity in treatment effect. The most important variable in terms of contribution to heterogeneity came out to be Domestic Credit to Private sector (% of GDP). Then we plotted the predicted value against four most important variables in terms of contribution to heterogeneity. We found out that Domestic credit to private sector (% of GDP), GNI per capita (current LCU), and GDP per capita (current LCU) have negative effect on the treatment effect of oil price shock on inflation while Carbon-dioxide Damage (% of GNI has a positive effect on the treatment effect of oil price shock on inflation. This is according to expectation and thus doesn't indicate any anomaly. Then we tried to analyse the heterogeneity in treatment effect. When we plotted the estimated treatment effect rank wise we found out less variability which makes it hard for us to conclude any heterogeneity in treatment effect. The plot with confidence interval surrounding the treatment effect also shows that as the quantum of treatment effect increases the variability and the uncertainty surrounding the estimate also increases.

At last we estimated the impact of oil price shock in 2019 on G20 countries. We found out that Argentina has the biggest effect of an oil price shock on inflation. This result is plausible considering the turbulent macroeconomic history of Argentina. The effect on inflation for India was above average which is also plausible considering India's dependence on imports for majority of its crude oil requirement.

The reason for uncertainty regarding the estimates of treatment effect and also lack of evidence for heterogeneity can be detected in the nature of the dataset. The dataset has big proportion of missing values and even after removing more than half of variables the proportion of missing values remains to be around 48%. A better method for imputing missing values or a more complete dataset can improve the quality of estimates.

# References

Athey, S; Imbens, G; Kong, Y & Ramchandra, V (2016). An Introduction to Recursive Partitioning for Heterogeneous Causal Effects Estimation Using causalTree package. arXiv:1902.07409v1.

Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences.

Athey, S., Tibshirani, J., & Wager, S. (2018). Generalized random forests. Annals of Statistics.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. arXiv:1810.13237v2.

Tiffin, A (2019). Machine Learning and Causality: The Impact of Financial Crises on Growth. IMF Working Papers. WP/19/228.

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association.

Zhao J., Runfola D.M., Kemper P. (2017) Quantifying Heterogeneous Causal Treatment Effects in World Bank Development Finance Projects. In: Altun Y. et al. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science, vol 10536. Springer, Cham.

## Internet Sources

https://cran.r-project.org/web/packages/grf/index.html

https://cran.r-project.org/web/packages/grf/grf.pdf

https://www.markhw.com/blog/causalforestintro

https://www.statworx.com/ch/blog/machine-learning-goes-causal-ii-meet-the-random-forests-causal-brother/

http://themlbook.com/wiki/doku.php?id=start

# Appendix

## R Codes

```
> # LIBRARIES
> library(tidyverse)
> library(lubridate)
> library(grf) # Generalized Random Forest for CausalTree estimation of he
terogenous treatent effect
> library(imputeTS)
> library(gghighlight)
> library(DiagrammeR)
> library(ggthemes)
> library(knitr)
>
> # DATA IMPORT
> WDIData <- read.csv("WDIData 1960-2019.csv", stringsAsFactors = F)
> WDISeries <- read.csv("WDISeries.csv", stringsAsFactors = F)
> brent_month <- read.csv("brent-month_csv 1987-2020.csv", stringsAsFactor
s = F)
> WDICountry <- read.csv("WDICountry.csv", stringsAsFactors = F)
>
> inflation_variables <- c("Inflation, GDP deflator (annual %)", "GDP defl
ator (base year varies by country)", "Inflation, consumer prices (annual %
)")
> inflation_Indicator.Code <- c("NY.GDP.DEFL.KD.ZG", "NY.GDP.DEFL.ZS", "FP
.CPI.TOTL.ZG")
>
> G20_countries <- c("ARG", "AUS", "BRA", "CAN", "CHN", "DEU", "FRA", "IND
", "IDN", "ITA", "JPN", "MEX", "RUS", "SAU", "ZAF", "KOR", "TUR", "GBR" ,"
USA")
>
>
> # DATA WRANGLING
> WDIData$ï..Country.Name <- NULL
> WDIData$Indicator.Name <- NULL
> WDIData_1 <- gather(WDIData, key = "Year", value = "Value", -Country.Cod
e, -Indicator.Code)
> WDIData_2 <- spread(WDIData_1, key = "Indicator.Code", value = "Value")
> WDIData_2$Year <- str_replace(WDIData_2$Year, "X", "") %>% as.numeric()
>
> # removing inflation related variables' columns
> WDIData_2A <- WDIData_2 %>% select(-NY.GDP.DEFL.KD.ZG, -NY.GDP.DEFL.ZS,
-FP.CPI.TOTL.ZG)
>
> # removing missing columns where more than 1/2 of observations are missi
ng values
> WDIData_col_NA <- map_int(WDIData_2A, ~sum(is.na(.x)))
```

```
> nrow_WDIData_2A <- nrow(WDIData_2A)
> WDIData_col_NA_which <- ifelse(WDIData_col_NA <= nrow_WDIData_2A*0.50, T
, F) %>% which() # nrow_WDIData_2/2 because only 50% i.e. 1/2 missing valu
es are allowed
> WDIData_3 <- WDIData_2A[, WDIData_col_NA_which]
>
> # removing rows where year is missing and truncating it to 1986-2016
> WDIData_4 <- WDIData_3[which(!is.na(WDIData_3$Year)),]
>
> # imputing missing values column wise using linear interpolation
> WDIData_5 <- WDIData_4 %>% arrange(Country.Code, Year)
> for(i in 1:ncol(WDIData_5)){
+   WDIData_5[, i] <- na_interpolation(WDIData_5[, i], option = "linear")
+ }
>
>
>
> # left joining outcome variable inflation
> inflation_consumer <- WDIData_2 %>% select(Country.Code, Year, Inflation
= FP.CPI.TOTL.ZG)
> inflation_consumer <- inflation_consumer[which(!is.na(inflation_consumer
$Year)),]
> inflation_consumer <- inflation_consumer %>% arrange(Country.Code, Year)
> inflation_consumer$Inflation <- na_interpolation(inflation_consumer$Infl
ation, option = "linear")
> WDIData_6 <- left_join(WDIData_5, inflation_consumer, by = c("Country.Co
de", "Year"))
>
> # binary treatment variable for oil price shock
> # if increase in price is more than 50 % from 12 month lag then a price
shock, only positive shocks included
> brent_month$Date <- brent_month$Date %>% as.Date()
> brent_month$Year <- year(brent_month$Date)
> brent_month$price_change <- (brent_month$Price - lag(brent_month$Price,
n = 12))/brent_month$Price
> brent_month <- brent_month %>% filter(price_change > 0)
> brent_month$Shock <- ifelse(brent_month$price_change >= 0.5, 1, 0)
> brent_month$price_change <- NULL
>
> treatment <- brent_month %>%
+   na.omit() %>%
+   group_by(Year) %>%
+   summarise(treatment = sum(Shock)) %>%
+   mutate(treatment = ifelse(treatment > 0, 1, 0))
>
> # we will ad one year in treatment so that feature is one year back and
is not affect by treatment
> treatment$Year <- treatment$Year + 1
>
> # joining treatment to our final dataset
> WDIData_6 <- left_join(WDIData_6, treatment, by = "Year")
> final_dataset <- WDIData_6 %>% na.omit()
>
```

```
> # proportion of treatment variables
> sum(final_dataset$treatment)/nrow(final_dataset) # 0.07142857
>
> # arranging by country and year
> final_dataset <- final_dataset %>% arrange(Country.Code, Year)
>
> # DATA ANALYSIS
> # estimating heterogenous treatment effect of oil price shock on inflati
on
>
> # dividing data into train and test
> set.seed(1996)
> cases <- sample(seq_len(nrow(final_dataset)), round(nrow(final_dataset)*
0.6))
> train <- final_dataset[cases,]
> test <- final_dataset[-cases,]
>
> # estimating the causal forest
> CF <- causal_forest(
+   X = train[, -c(1, 2, ncol(final_dataset)-1, ncol(final_dataset))],
+   Y = train$Inflation,
+   W = train$treatment,
+   num.trees = 5000,
+   seed = 1996
+ )
>
> preds <- predict(
+   object = CF,
+   newdata = test[, -c(1, 2, ncol(final_dataset)-1, ncol(final_dataset))]
,
+   estimate.variance = TRUE
+ ) %>% abs()
>
> test <- cbind(test, preds)
>
> variable_importance <- CF %>%
+   variable_importance() %>%
+   as.data.frame() %>%
+   mutate(variable = colnames(CF$X.orig)) %>%
+   arrange(desc(V1))
>
> colnames(WDISeries)[1] <- colnames(variable_importance)[2]
> variable_importance <- left_join(variable_importance, WDISeries, by = "v
ariable")
>
> # saving variable importance
>
> write.csv(variable_importance, file = "variable_importanceA.csv")
>
> # Estimating individual effect for G20 countries
> G20_dataset <- final_dataset %>% filter(Country.Code %in% G20_countries,
Year == 2019)
>
```

```
> G20_preds <- predict(
+   object = CF,
+   newdata = G20_dataset[, -c(1, 2, ncol(G20_dataset)-1, ncol(G20_dataset
))],
+   estimate.variance = TRUE) %>%
+   abs()
>
> G20_dataset <- cbind(G20_dataset, G20_preds)
>
> # Average treatment effect
> average_treatment_effect <- average_treatment_effect(CF) %>% abs()
> paste("95% CI for the ATE:", round ( average_treatment_effect[1] , 3) ,"
+/ -", round ( qnorm (0.975) * average_treatment_effect[2] , 3))
>
>
> # PLOTS
>
> # plotting relation between top four important varibles and the predicti
ons
>
> p1 <- ggplot(test, aes(x = log(test[,variable_importance$variable[1]]),
y = predictions)) +
+   geom_point() +
+   geom_smooth(method = "loess", span = 1) +
+   theme_light() +
+   xlab(variable_importance$Indicator.Name[1]) +
+   ylab("Predictions")
> p2 <- ggplot(test, aes(x = log(test[,variable_importance$variable[2]]),
y = predictions)) +
+   geom_point() +
+   geom_smooth(method = "loess", span = 1) +
+   theme_light() +
+   xlab(variable_importance$Indicator.Name[2]) +
+   ylab("Predictions")
> p3 <- ggplot(test, aes(x = log(test[,variable_importance$variable[3]]),
y = predictions)) +
+   geom_point() +
+   geom_smooth(method = "loess", span = 1) +
+   theme_light() +
+   xlab(variable_importance$Indicator.Name[3]) +
+   ylab("Predictions")
> p4 <- ggplot(test, aes(x = log(test[,variable_importance$variable[4]]),
y = predictions)) +
+   geom_point() +
+   geom_smooth(method = "loess", span = 1) +
+   theme_light() +
+   xlab(variable_importance$Indicator.Name[4]) +
+   ylab("Predictions")
>
> cowplot::plot_grid(p1, p2, p3, p4)
>
> # plotting the heterogeneous treatment effect
>
```

```
> plot_htes <- function(cf_preds, ci = FALSE, z = 1.96) {
+   out <- ggplot(
+     mapping = aes(
+       x = rank(cf_preds$predictions),
+       y = cf_preds$predictions
+     )
+   ) +
+     geom_point() +
+     labs(x = "Rank", y = "Estimated Treatment Effect") +
+     theme_light()
+
+   if (ci == 1) {
+     out <- out +
+       geom_errorbar(
+         mapping = aes(
+           ymin = cf_preds$predictions + z * sqrt(cf_preds$variance.estim
ates),
+           ymax = cf_preds$predictions - z * sqrt(cf_preds$variance.estim
ates)
+         )
+       )
+   }
+
+   return(out)
+ }
> hte_plot <- plot_htes(preds[sample(1:nrow(preds), size = nrow(preds)/8),
])
> hte_plot_CI <- plot_htes(preds[sample(1:nrow(preds), size = nrow(preds)/
8),], ci = T)
>
> # G20 countries bar plot
> G20_dataset %>%
+   ggplot(aes(x = Country.Code, y = predictions)) +
+   geom_col() +
+   gghighlight(Country.Code == "IND") +
+   ggtitle("Estimated Treatment Effect in 2019 for G20 Countries") +
+   ylab("Predictions") +
+   theme_economist()
>
>
> # plotting one of the tree in forest
> CF %>% get_tree(222) %>% plot()
>
> # accuracy of predictions
> rmse <- sqrt(sum((test$Inflation - test$predictions)^2)/nrow(test))
>
> # FOR KNITTING RMD DOCUMENTS IN CURRENT SESSION
> rmarkdown::render("causal_tree.Rmd", output_format = "word_document")
```