Prediction of Corruption Perception Index via Machine Learning Approach.

Qiteng Feng (fengqite) qiteng.feng@mail.utoronto.ca

Yuanhan Peng (pengyua9) yuanhan.peng@mail.utoronto.ca

Zikun Xu(xuzi9) zikun.xu@mail.utoronto.ca

Introduction:

A country's corruption level is one of the most significant threats to a country's economic and financial stability. Corruption erodes trust in public institutions, exploits financially vulnerable individuals, and further exacerbates social inequality, social divisions, and environmental issues. The World Bank, in 2016, estimated that $2.6 trillion was the cost of corruption globally, which also accounted for 5% of the global gross domestic product (World Bank).

To improve corruption measures, Transparency International created an organization based in Germany to combat corruption and promote transparency publicly across all sectors of society. They noted that 131 countries had made no significant progress in lowering corruption levels in the last decade, according to their Corruption Perception Index. They also mentioned that 40% of business executives paid bribes to public institutions when dealing with business in developing countries (Transparency International). The International Monetary Fund mentioned that the act of bribery alone accounts for around $2 trillion US dollars globally, and bribery is only one of the different forms of corruption that contribute to corruption levels. These facts give insight into the levels of corruption that exist globally and do not affect individuals at a micro-level as much as it impacts a country's economy at a macro level. While it is hard to evaluate absolute levels of corruption for each country since data on this topic is often missing and unreleased to the public, we can analyze the socio-economic indicators of most countries to predict the Corruption Perception Index (CPI). Analysis of these indicators as well as prediction of a given country's Corruption Perception

Index level is necessary for our main motivation, which is to make appropriate suggestions for governments to take action to reduce their levels of CPI ultimately.

This paper aims to contribute to corruption studies globally by providing analysis through machine learning practices. Machine learning algorithms would allow us to reveal and analyze the significant predictors for the corruption perception index, ultimately allowing us to make proper suggestions to the government about anti-corruption policies.

Corruption Perception Index (CPI) is an index that ranks a country's perceived levels of public sector corruption, which economics experts and business institutions assess through various public opinion surveys, such as the World Bank, Economist Intelligence Unit, and Global Insight. This is considered as this paper's main outcome variable and needs to be validated repeatedly in order to represent a country's CPI accurately. As mentioned before, absolute levels of corruption measures are extremely complicated to measure which is because corruption can only be measured either through direct observation, for example, through law enforcement or government interventions; or in our case, from perception and opinion surveys which are assessed by various sources and institutions.

How does CPI affect global economics and or life quality? High levels of corruption perception imply illegal financial or unreported financial activities hidden from the public. Corruption is seen as one of the greatest threats to a country's financial stability, and the negative effects of corruption have a major impact on financially vulnerable people.

To make reasonable suggestions that can reduce corruption perception levels, we will determine which features significantly affect a country's CPI and rank those features from the highest to lowest in terms of how much it affects CPI. Using multiple prediction models allows us to have a better prediction at the end, which also allows for more accurate variable assessments. Finding the most significant machine learning model also allows us to answer other questions, such as what the government can do to reduce corruption perception, as well as which factors contribute the most towards the prediction of CPI.

The main contributions of this paper include the following: Firstly, the analysis target extends to the entire world and multiple index socio-economics index. Secondly, building high-performance machine learning models to predict a country's CPI and compare model accuracies to find the best model that fits our data. Thirdly, rank the top 5 indicators contributing the most towards CPI prediction. Lastly, from the results, make suggestions to the government on the indicators that have the most significant effects on CPI to improve anti-corruption policy formulations. Suggestions may include increase export levels to improve GNI or decrease in the size of the government. These contributions are different from our literature articles in terms of the size of data, usage of different features, and machine learning model comparisons.

Literature Review

Three different pieces of literature that relate to corruption perception analysis using statistical models were studied and they all have the same research objective of identifying the most influential factors that define a country's corruption levels. The first article, "The Human Development Index (HDI) and the Corruption Perception Index (CPI) 2013-2017" (Sarabia, 2020) analyzes populism in Europe by using linear regression model with CPI as the dependent variable and aspects of human development index as predicting variables, which is also partially used in our prediction model. However, they only focus on European countries as well as the relationship between CPI and HDI, but in this paper, many more countries were included, and predictors other than HDI were used for the statistical model. The second literature, "Predicting and explaining corruption across countries: A machine learning approach" (Lima, 2020), use a random forest model and cross-validation method in their study to predict corruption, which is very similar to this study, but in addition to the random forest model, this study also uses linear regression to explore linear relationships between CPI and predictors. In the third paper, "The corruption perception index: Analysis of dependence on socio-economic indicators" (Domashova, 2020), both random forest and linear regression models were analyzed and modeled for the prediction of CPI. However, they evaluated variable importance by using clusters such as K-means to divide countries into four groups and find common features for each group, while we ranked the predictors according to their performance. The main reason these three literatures were chosen is to compare the results of different machine learning methods with different predictors and countries and see if there is commonality or trends within the results.

Data and Methodology:

In this study, the dataset contains 13 variables: Corruption Perception Index(CPI) from the Transparency International website; Size of Government, Legal System & Property Rights, Sound Money, Freedom to Trade Internationally and Regulation from Fraser Institute Economic Freedom; Life Expectancy, Gross National Income per Capita(GNI per capita), Expected Years of Schooling and Mean Years of Schooling from Human Development Reports; Unemployment Rate, Government Primary Net Loans/Borrowing, and Export Volume of goods and services from Our World in Data website, where CPI is used as the dependent variable and remaining 12 variables as predictors. These 12 predictors were chosen because they were all found to be associated with CPI (Sarabia, 2020; Lima, 2020; Domashova, 2020), but they had not been analyzed together for a comparison of influential capacity on CPI.

The dataset contains 196 countries in the world. Due to the limitation of country numbers, using only one year of data as data sets may lead to an overfitting problem of the models since the sample size is too small. Therefore, the dataset collects nine years of data from 2012 to 2020, including a total of 1764 observations. The timing was chosen because the data for each variable within the nine years is more complete and avoids a large amount of missing value.

As the dependent variable, the scale range of CPI is from 0 to 100, where a higher CPI score indicates less corruption perceived by the masses in this country. The heatmap in figure 1 displays the correlation within each variable. The top 5 features with significant correlation with CPI are legal system & property rights, Gross

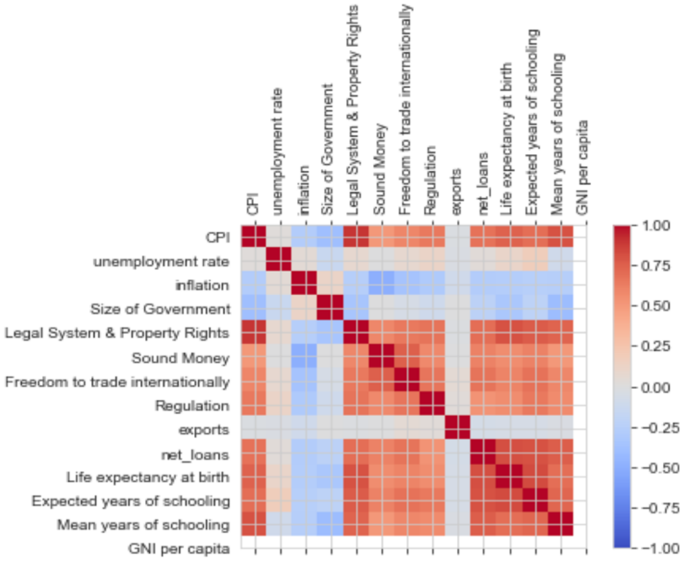

Figure 1: Correlation heatmap of within variables

National Income (GNI) per capita, life expectancy at birth, expected years of schooling, and the size of the government.

When the further model training process was performed, as different from the usual way of randomly and proportionally splitting the data into training and test sets, the allocation of the data into training or test sets was manipulated via a random split of the countries. 80% of the countries' data were divided into a training set and the remaining 20 percent into a test set. Therefore, countries that appear in the training set will not appear in the test set, and the purpose of it is to avoid a potential correlation between the training and test sets. Since the data in this study was collected from 2012 to 2020, the data set contains a time series structure, which implies the data for adjacent years of the same variable in a country were correlated. When data from this country are present in both the training and test sets, the accuracy of the tests will be upwardly estimated and invalid, despite the different years of data collection.

To optimize the performance of models, a cross-validation method was applied in the model training process at the training set to adjust the hyperparameters of the models. The training set is divided into five folds, and each fold became the validation set when the model was trained based on the other four groups. The cross-validation estimates were the average of 5 criterions calculated when each fold as a validation set. The models' hyperparameters were adjusted based on the result of cross-validation estimates. The three machine learning models were proposed: linear regression, decision tree, and random forest. The model performance evaluation proceeded by comparing their mean squared error (MSE) and coefficient of determination (R^2).

Linear regression model is widely used in data analysis, and it is appropriate in this study since the dependent variable CPI is a numerical and continuous variable. Its mathematical expression is: $Y = \beta_0 + \sum_{\mathrm{j}=1}^{12} \beta_{\mathrm{j}} X_{\mathrm{ij}} + \epsilon$, where Y is CPI, β_0 is the intersection, β_j is coefficient of predictors, X_{ij} is the predictor, and ϵ is the error term. Since there is no variable selection process in this study, too many predictors may cause overfitting problems. A regularization method, LASSO regression was used to penalize the coefficients of predators via minimizing equation: $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{\mathrm{j}=1}^{12} \beta_{\mathrm{j}} X_{\mathrm{ij}}\right)^2 + \lambda \sum_{i=1}^{12}\left|\beta_j\right|$, where λ is a hyperparameter in LASSO regression. When λ increases, the coefficients will approach to 0, which helps to find the most influential predictors in linear regression model (James, Witten, Hastie, Tibshirani, 2017).

For regression tree model, unlike linear regression model emphasizing the linear relationship between dependent variable and predictors, it can distinguish the non-linear and complex relationship between CPI and 12 predictors. Regression tree model makes prediction by recursively splitting the data based on different predictors. The expression equation of regression tree model is: $Y = \sum_{m=1}^{M} c_m I_{x \in R_m} + \epsilon$, where R_m is the partition of feature space. However, the regression tree model has two main disadvantages: overfitting and non-robust. The predictive power may not be as good as the other two proposed models, and a subtle data change may lead to a drastic change in the model (James, Witten, Hastie, Tibshirani, 2017).

To address the shortcomings of the regression tree model, a random forest model was applied in this study. It reduces the regression tree model's variance by training various regression trees based on sub-training sets from the training data set and taking an average of their results. In regression, the predicted value is calculated by$\widehat{f_{rf}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$, where T_b is the tree (James, Witten, Hastie, Tibshirani, 2017).

Result

A 5-fold cross-validation technique is used on the training dataset to tune models to find the most appropriate hyperparameters and prevent overfitting. For lasso regression, the optimal alpha value is approximately 0.017. By plotting, shown in figure 2, alpha shows an exponential relationship with the mean squared error on each fold. For decision trees, ccp_alpha is a significant hyperparameter to prevent the tree from overfitting and control the size of a tree. Its optimal value after 5-fold cross-validation is 0.5. For random forest, each hyperparameter is optimized by using randomized search. The best combination of hyperparameters for the random forest is when the maximum depth of the tree is 40, the minimum number of samples needed to be split is 7 and the number of trees is 600.

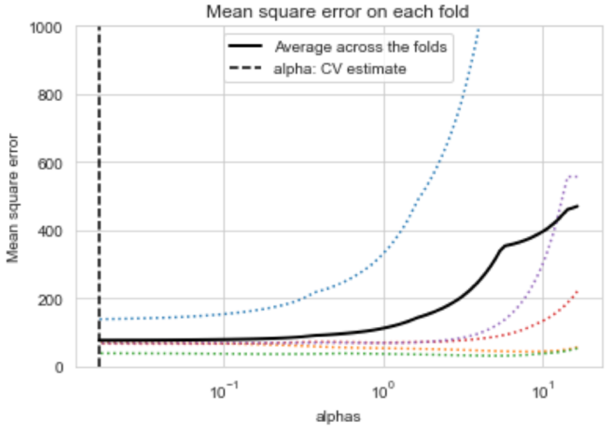

Figure 2: Mean square error on each fold

The results of the prediction performance of three different machine learning models: lasso regression, decision tree, and random forest are shown in table 1. The Lasso regression algorithm is the best model for the prediction of CPI. It has the lowest mean squared error and highest r-squared, which are 77.65 and 0.84 respectively.

Model	Test_MSE	R^2
Lasso Regression	77.65	0.84
Decision Tree	93.93	0.8
Random Forest	82.46	0.83

Table 1: Model performance

In addition, the ranking of features that significantly affect the prediction of corruption score is another focus of this research. Since the lasso regression has been revealed as the most appropriate model to predict CPI, its feature importance, shown in figure 2, indicates the top 5 important features for prediction. The legal system & property rights feature is the most important feature, followed by GNI per capita, meaning years of schooling, regulation, and sound money.

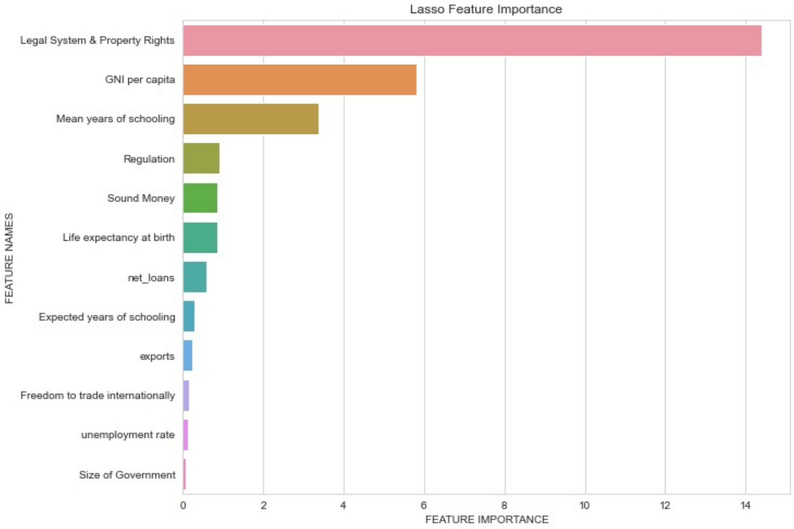

Figure 2: Lasso regression feature importance

Conclusion

The data on the corruption perception index and 12 potential predictors from 2012 to 2020 is downloaded from four different sources, such as Transparency International, Fraser Institute Economic Freedom, Human Development Reports, and Our World in Data. In order to predict CPI, three regression models are chosen, and by the application of K-fold cross-validation, the linear regression model with lasso regularization is the best model for the prediction of CPI, compared with decision tree and random forest models. Furthermore, the feature importance of the lasso regression indicates that Legal system & property rights, GNI per capita, meaning years of schooling, regulation, and sound money are significant features for the prediction of CPI score. In addition to a high-accuracy prediction model, changes in those five indicators play a key role in formulation of government anti-corruption policies. In general, an increase in those indicators leads to an increasing CPI and a less corrupt country. For example, promoting an independent and unbiased judiciary to protect property rights and increasing surveillance on property transactions can help improve the legal system & property rights feature, resulting in less corruptive government. Moreover, government policies that improve access to education and decrease inflation are helpful to prevent corruption, by increasing mean years of schooling and sound money respectively.

Although the study can provide the government with proper model to predict CPI and significant features to help make anti-corruption policies, there are some limitations in this study. The number of features and overall variables used to train and test models are limited.

Reference:

Domashova, J., & Politova, A. (2021). The corruption perception index: Analysis of dependence on socio-economic indicators. *Procedia Computer Science*, *190*, 193–203. https://doi.org/10.1016/j.procs.2021.06.024

Lima, M. S., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, *37*(1), 101407. https://doi.org/10.1016/j.giq.2019.101407

Sarabia, M., Crecente, F., del Val, M. T., & Giménez, M. (2020). The human development index (HDI) and the corruption perception index (CPI) 2013-2017: Analysis of social conflict and populism in europe: Znanstveno-strucni casopis*.Ekonomska Istrazivanja, 33*(1), 2943-2955. doi:https://doi.org/10.1080/1331677X.2019.1697721

World Bank. "C02 Emissions (Metric Tons Per Capita)." *World Development Indicators*, The World Bank Group, 2015, data.worldbank.org/indicator/EN.ATM.CO2E.PC.

IMF. (2016, May 11). IMF survey : Fighting Corruption Critical for growth and macroeconomic stability-IMF paper. IMF. Retrieved December 15, 2022, from https://www.imf.org/en/News/Articles/2015/09/28/04/53/sores051116a

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*.