Predicting Banking Crises

By: Andrew Furlong, Zhenxin Zhang, Kyle Wagner, Amy Tang

Project Supervisor: Professor Freeman

Background & Introduction

The banking crisis breaks people's trust in the banking system and damages the finance of a country. Hence, it is crucial to learn from the banking crisis in history and detect whether it would happen. Various economic (international and domestic) and political factors may contribute to the breakout of a banking crisis. Utilizing a global dataset on banking crises containing relative contributions, we would like to explore which variables help predict banking crisis and create a model that could accurately predict whether a banking crisis will occur.

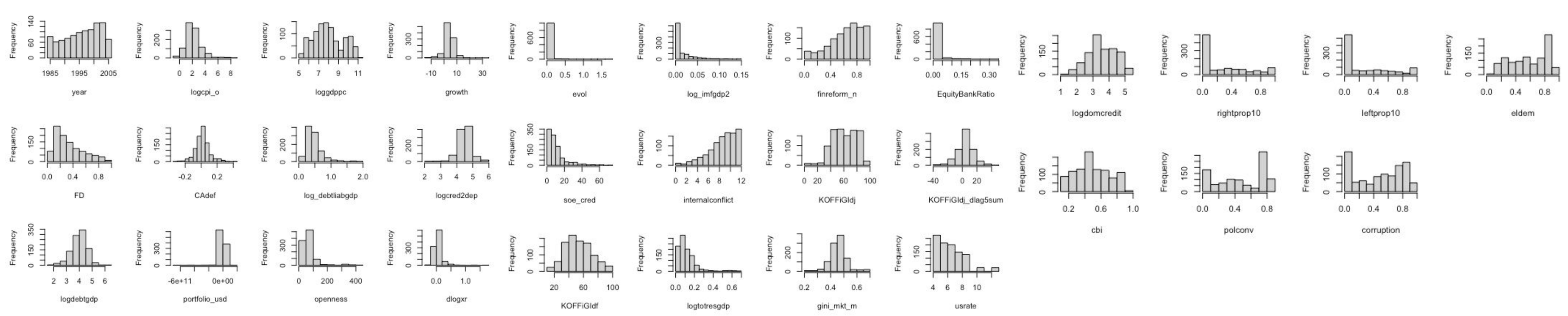
Variable Category	Region	Economic Growth	International Factors	Bank	Political Factors
Example Variables	country, regionun	growth, FD, loggdppc	dlogxr, logdebtgdp	EquityBankRatio, usrate	Internalconflict, corruption, elections

Data Pre-Processing

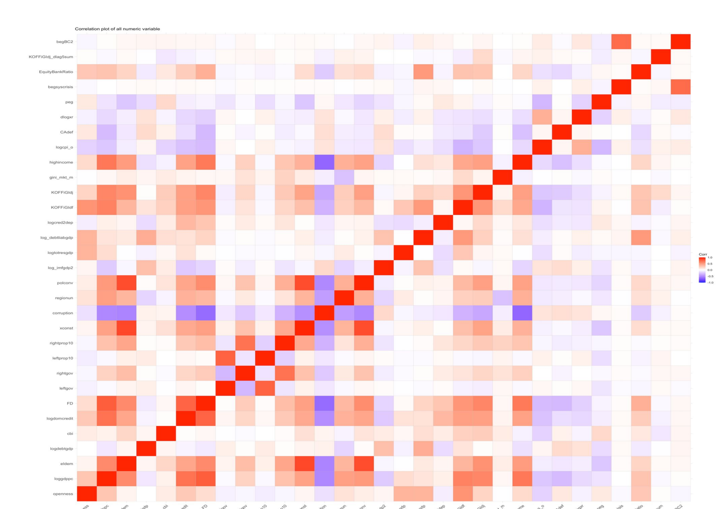
- **Dataset**

The original dataset has 11560 rows and 44 columns. Each row represents a particular year in a country's history with associated economic variables. The response variable that we are focusing on is `begsyscrisis`. In total there are 148 crises, and 9734 non-crises, with 1678 missing values. We decided to use a subset without any missing data because the statistical models we used could not handle missing data, and missing data would make inference difficult.

- **EDA**
  - *Univariate Analysis*:
    Below are the histograms of all of the numeric variables.



  - *Bivariate Analysis*:
    The graph to the right is the correlation plot of all of our variables. We noticed that some variables are highly correlated with one another, which is an important information to keep in mind during the modeling process.



- **Data Imputation**
  - In order to resolve some of the issues in this dataset, we will use two different methods:
    - Multiple Imputation: This method analyzes the distributions of all covariates in our dataset and predicts missing values, thereby reducing the number of NA observations.
    - SMOTE: This method artificially increases the size of the minority class (banking crises) by building examples that are similar to those already in the feature space.

Methods

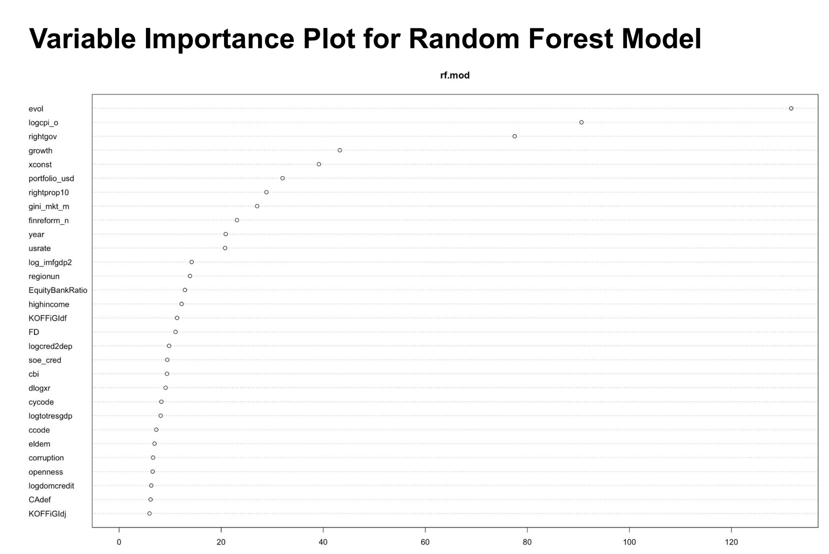
- We fit models using seven methods, including logistic regression, decision trees, ridge regression, lasso regression, XGboost, and random forests.
- Random forest - a machine learning method used for classification and regression that involves constructing and aggregating multiple decision trees - yielded our lowest misclassification rate when run on the smote dataset (2.9%).
- We decided to use our random forest model to generate final predictions, because it minimized misclassification rate while maximizing AUC.

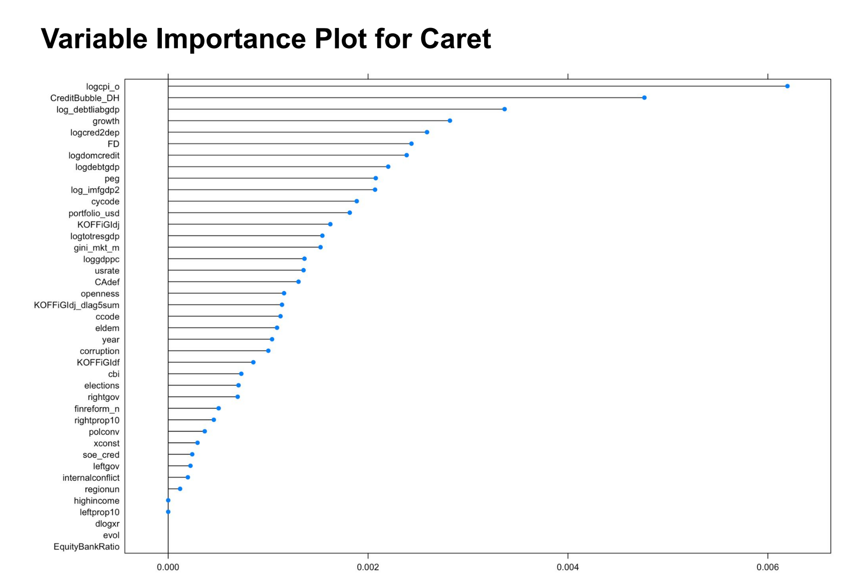
Analysis and Results

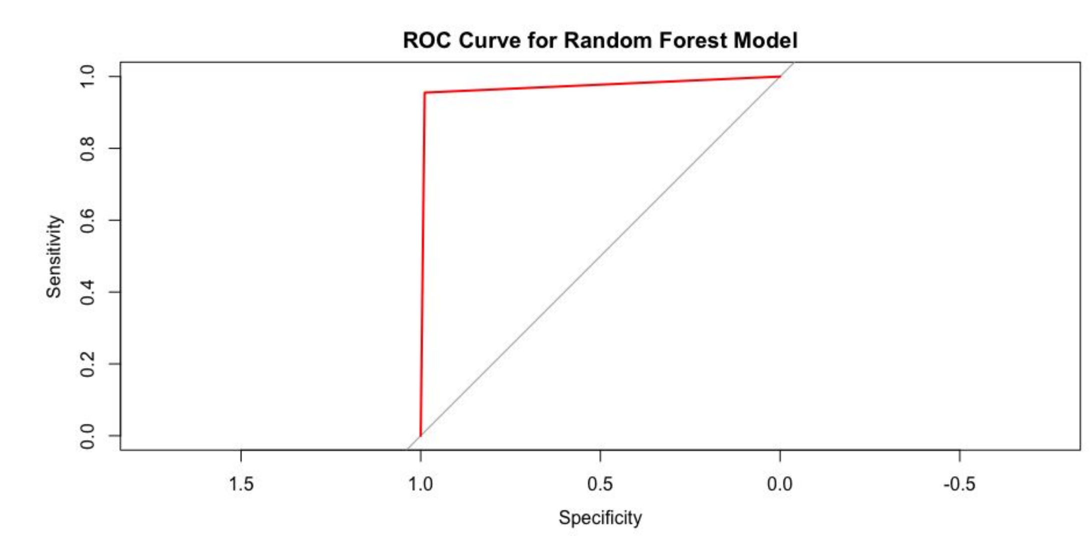
Predicted Value	No Banking Crisis	Banking Crisis
Predicted No Banking Crisis	179	9
Predicted Banking Crisis	2	192

Model Name	Misclassification Rate	Area Under Curve
Logistic Regression	0.051	0.5
Decision Tree	0.0314	0.5
Ridge Regression	0.034	0.5
Lasso Regression	0.0341	0.5
Random Forest	0.0294	0.5
MI Logistic Regression	0.0104	0.5
MI Logistic Regression, Weighted	0.234	0.799
MI Decision Tree	0.0104	0.5
MI Decision Tree, Weighted	0.228	0.782
MI XGBoost	0	1
MI Random Forest	0.0104	0.5
MI Random Forest, Weighted	0.0104	0.5
SMOTE Logistic Regression	0.118	0.883
SMOTE Logistic Regression, Weighted	0.0681	0.934
SMOTE Decision Tree	0.113	0.888
SMOTE Decision Tree, Weighted	0.0707	0.931
SMOTE Random Forest	0.0262	0.975

- The table above shows the performance of our best model.
- The table to the right shows the performance of all models built.
- Below are the variable importance plot obtained from caret and our best model, the Random Forest Model obtained from SMOTE
- Using the ROC curve below, we derived a Youden's J Statistic value of 0.9441742, which indicates that our model had good performance.







Conclusions

- From our two variable importance plots, we can observe that there are many variables that are considered significant in terms of predicting banking crises
- We determined that our Random Forest Model was superior when we wished to classify our data. This was due to a low misclassification rate and a high AUC value
- In the future, we can try to explore the relationship between banking crises and political instability in a country, in order to identify economic factors that are correlated with increased unrest.