ORIE 5740 Final Project Report

# CDS Spread Analysis

Yash Ganatra (ybg3); Harshavardhan Bapat (hsb57)

## Abstract

This project employs machine learning techniques using financial and economic data to attempt to predict sovereign CDS spreads of 3 emerging markets. Prediction is attempted using linear models with lasso and ridge regularization with time series cross validation and non-linear ensemble methods such as Random Forest, Boosting and XGBoost. Lagged terms and other engineered features are also added to account for auto-correlations. At the end, model limitations and possible improvements are discussed.

## 1 Data Analysis

### 1.1 Background

The purpose of this project is to train a machine learning model to predict sovereign CDS spreads of emerging market countries based on various financial and economic data. A CDS (Credit Default Swap) is a financial agreement on an underlying credit security that act as a type of 'insurance' protecting bondholders in the case of default. It is not necessary to hold the underlying bond to purchase a CDS, and there exists a vast secondary market for credit default swaps. Intuitively, the CDS spread is an indication of the riskiness of the underlying bond. CDS spreads on sovereign bonds would therefore indicate the implied riskiness of government bonds of different countries. There are likely various global and domestic economic factors that cause fluctuations in sovereign CDS spreads, making it prudent to build a model to predict these spreads using the various factors. Initially, we intended to analyse the spreads of 3 developed and 3 emerging markets, but due to data restrictions on Bloomberg for developed nations, we have based our study on the sovereign spreads of 3 emerging markets (Brazil, China and Russia).

### 1.2 Data Description

We base our feature set on economic and financial variables that we expect to have significant impact on sovereign CDS spreads. For each emerging market, the target variable is the 5Y Sovereign CDS spread, while our features are

a combination of domestic and global financial data. Domestic factors include the nation's most prominent stock market index, which acts as a proxy for the country's financial and economic health. We expect there to be an inverse correlation between the domestic stock index and CDS spreads, as worsening financial performance would increase the perceived risk, and result in widening spreads. We also consider the domestic currency FX rate, since appreciation/depreciation in the country's domestic currency directly impacts investors' cost and return on investment. We consider the following global factors:

1. S&P 500 Implied Volatility Index (VIX): The VIX tracks the expected volatility in the S&P 500 implied by options on the index. The index is a global indicator of financial health, so the VIX behaves as a proxy for global uncertainty and volatility. We expect this to have a positive correlation to sovereign spreads, as an increase in volatility results in an increase in perceived risk.

2. WTI Crude Oil Prices: Crude oil prices have a significant impact on economic health of nations. Higher oil prices benefit oil exporting nations vs importing nations and vice versa. Therefore, we include oil prices as one of the features in our model, expecting some correlation with CDS spreads, especially for nations with significant oil exposure (eg. Russia).

3. Crude Oil Volatility Index: Volatility in oil prices is especially influential in impacting financial metrics, and therefore we include the Oil Volatility Index (OVX) to capture information regarding periods of high volatility.

We also consider the following two features to account for cost and liquidity of financing, which significantly impacts investor decisions.

1. 10-Year US Treasury Rate: The interest rate on the current 10-yr US treasury note can be used as a proxy for global developed risk free interest rate. While rates on emerging market notes are higher than this, it can be used as a good benchmark for cost of financing.

2. TED Spread: The TED spread is the difference between inter-bank loan interest rates and the US short rate. This is a proxy for aggregate liquidity in the market, as lower liquidity results in a widening of this spread.

### 1.3 Data Cleaning and Processing

We consider data between 2007-2018, which includes the global financial crisis of 2007-08 as well as a transition in commodities cycle. We face some data loss due to missing data, but the loss is not significant and majority of the information is retained. Initially, we had also intended to consider the US Economic Policy Uncertainty Index as a proxy for global economic uncertainty, but the data available was very erratic and not reliable, so we chose to drop it from consideration. Finally, we are left with 7 initial features for each sovereign CDS spread analysed. We have

2,769 observations for Brazil, 2,505 observations for China, and 2,741 observations for Russia.

### 1.4 Feature Engineering

A correlation plot between the initial variables indicates some correlation between a few of the features. The correlation between VIX and OVX is expected, since both are a measure of volatility albeit on different metrics, but volatility in oil prices is often accompanied with volatility in the stock market. Also, the USDRUB (Russian Ruble) exchange rate is strongly correlated with oil prices, since Russia is a major oil exporter, while the USDBRL (Brazilian Real) exchange rate has an inverse correlation with oil prices since oil is one of Brazil's major imports.

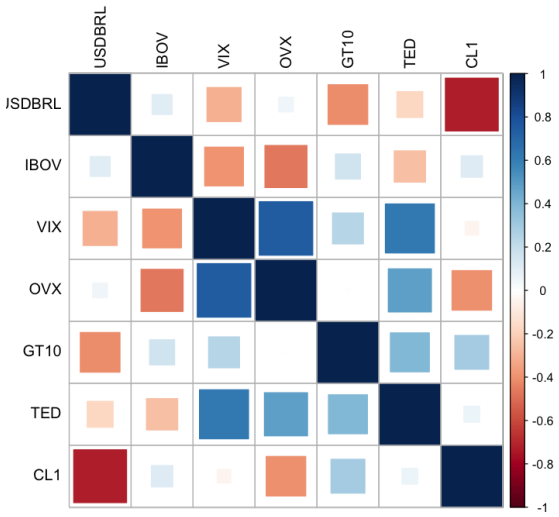


Figure 1: Correlation plot for Brazil

#### 1.4.1 Including changes in features

Often sudden movements in financial indicators are more informative than the value of the index itself. While we attempt to predict the CDS spread, we include percentage changes in the initial indicators considered to our feature set, on the expectation that these changes may provide additional information that is lost in the scale of the value of the indicators.

#### 1.4.2 Including Lagged Terms

It is expected that lagged values of the time series itself hold important information for predicting future values. We verify this by attempting to fit an Auto-regressive (AR) Model on the CDS spreads data. AR models values of the time series as a sum of lagged values multiplied by coefficients, and an error term. Our basis for attempting this is to check how many lagged values of the time series are considered significant for modeling this data. "auto.arima" in R fits an AR(2) model, indicating values of lag T-1 and T-2 are important in modeling the data, therefore, we include these lags as features.

Thus we have a total feature set of 16 variables.

## 2 Model Selection

The purpose of time series forecasting is to produce precise future forecasts. In the case of time-series data, the rapid and powerful methods we rely on in machine learning, such as train-test splits and k-fold cross-validation, do not work. This is due to the fact that they overlook the problem's time aspects. Therefore, we use time-series CV to choose different training and validation datasets that are contiguous to tune hyper-

parameters and use them for training the entire train dataset. We use this model and test it on the test dataset to get an unbiased evaluation of the model.

### 2.1 Preliminary Linear Regression

We first attempted linear regression on our dataset to get an idea of the fit, how significant each feature is in the prediction, and check if there is any over or underfitting, which would help us build a road map on which future models need can be implemented to address each problem that may arise.

### 2.2 Regularization

We found that some of the features from our dataset might be correlated which can affect the model's interpretability adversely. To avoid the problem of multicollinearity, we use regularization to produce a unique solution.

#### 2.2.1 Ridge Regression

We first use Ridge regression, which uses $l_2$ regularization. Since Ridge regression does not set coefficients to 0 but instead shrinks coefficients of unimportant variables, we believe it was a good idea to regularize using ridge to get a full picture of every feature importance.

#### 2.2.2 Lasso Regression

We also use Lasso regression which uses $l_1$ regularization. The idea is to investigate if our model can perform better with lasso regularization, and then to see if eliminating some variables will make it easier for us to comprehend the variables, since lasso can set some coefficients to 0.

### 2.3 PCR

Principal component regression (PCR) is a technique used for dimensionality reduction. The basic idea is to perform principal component analysis and using the principal components in linear regression. Since we observed mild correlations between a couple of features as can be seen in the correlation plots, we tried using PCR as it helps tackle multicollinearity, as well as helps with reducing overfitting.

### 2.4 Random Forest

We also tried fitting non-linear models to our dataset. The intuition behind employing a non-linear model was to capture non-linear relations between features and the target. One of the known non-linear models is the decision tree. Random Forests build uncorrelated decision trees by performing feature selection implicitly. This makes it an excellent model for dealing with data with a large number of features. Also, Random Forests are not influenced by outliers to a fair degree by binning the variables. Random Forests are known for their high accuracy and ability to manage the bias-variance trade-off. The variance is averaged as well because the model's premise is to average the results over the various decision trees it creates.

### 2.5 Boosting

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added until the training set is predicted perfectly or a maximum number of models are added. Boosting technique helps when we are dealing with bias or underfitting in the data set. Here, we use the standard boosting, as well as XGBoost that builds trees using Gradient boosting, a technique that involves creating new models that forecast the residuals or errors of previous models, which are then combined to form the final prediction.

## 3 Results

Refer **Figure 8** (Appendix) for full tabulated results.

### 3.1 Linear Regression

#### 3.1.1 Preliminary Regression

After data cleaning and feature preparation, we have a final feature set of size 14, and around 2500 data entries for each of the countries. We also use walk-forward validation to find the tuning hyper-parameter $\lambda$ that gives lowest MSE for the model.

*Results:* Initially the model seems to perform extremely well for Brazil (Train MSE = 49.352, Test MSE = 51.121), but upon deeper inspection we realise that the model is memorising the previous day data, and does not capture any other economical moves (Russia Train MSE = 38.118, Test MSE = 405.681). We can also observe memorisation in **Figure 2** for test predictions from linear regression of China's dataset.

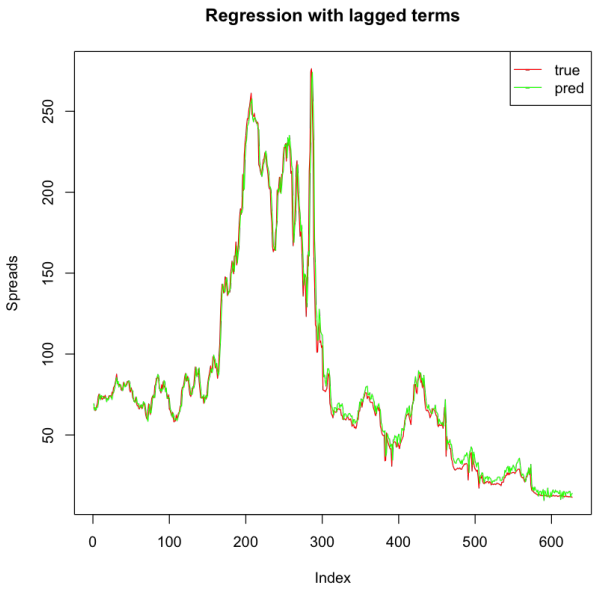


Figure 2: Linear regression (China)

#### 3.1.2 Ridge Regression

Since linear regression is unable to accurately capture the economic moves, we use Ridge regression with regularization next since it allows us to use complex models and avoid over-fitting at the same time. Despite OLS being the best linear unbiased estimator, ridge can demonstrably achieve a lower MSE than OLS by being a biased estimator.

*Results:* Ridge regression gives a good fit for China and Brazil (**Figure 3**), but gives a high test error for Russia suggesting overfitting.

#### 3.1.3 Lasso Regression

Lasso is used as both a regularization as well as feature selection technique as it shrinks coeffi-

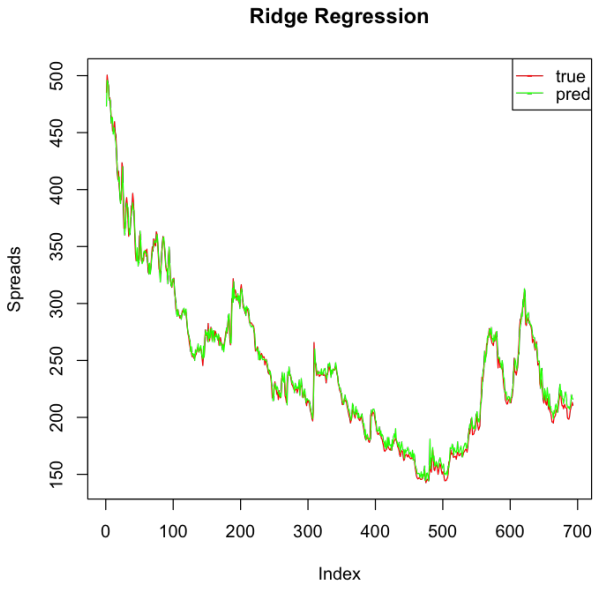


Figure 3: Ridge regression (Brazil)

cients of unimportant variables to 0.

*Results:* Lasso regression also gives a good fit for Brazil and China (slight overfit), but gives a high test error for Russia suggesting overfitting (**Figure 4**).

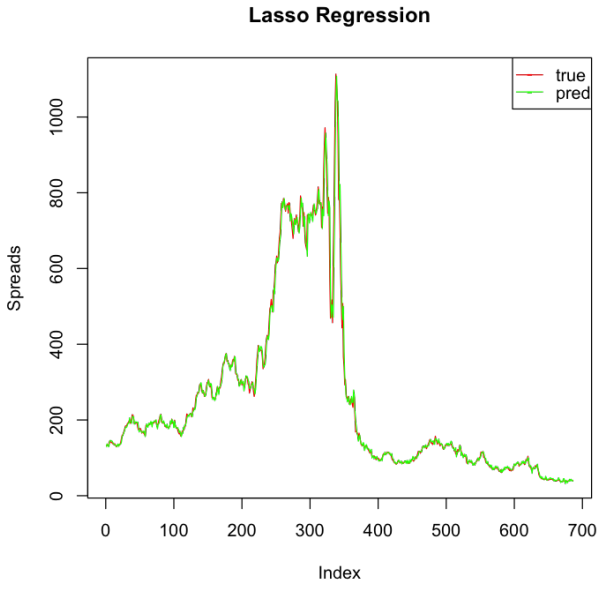


Figure 4: Lasso regression (Russia)

#### 3.1.4 Principal Component Regression

Since Lasso gives promising results suggesting some features might be unimportant, we try a dimensionality reduction using PCR.

*Results:* PCR gives a poor test error for all countries, and selects all 16 principal components as it's unable to capture enough variability to reduce dimensions.

### 3.2 Non-Linear Models

#### 3.2.1 Random Forest Regressor

We now try an ensemble method to predict the price as it implicitly performs feature selection to generate uncorrelated random trees. We attempt bagging using RandomForest as a variance reduction technique.

*Results:* Random Forests is unable to capture the variability in China and Russia's spreads, and results in an overfit. It performs extremely well on the train sets and poorly on tests proving the overfit. It performs fairly well on Brazil (**Figure 5**) (Train MSE = 12.795, Test MSE = 60.815).

#### 3.2.2 Boosting and XGBoost

We now employ another ensemble method following the Random Forest Regressor. Boosting is used as a bias reduction method. XGBoost makes the model slightly more complex and reduces the bias of the model significantly.

*Results:* Boosting and XGBoost performs the best for Brazil (**Figure 7**), with a Train and Test RMSE of 48.842 and 31.313 respectively for boosting. For Russia and China unfortunately

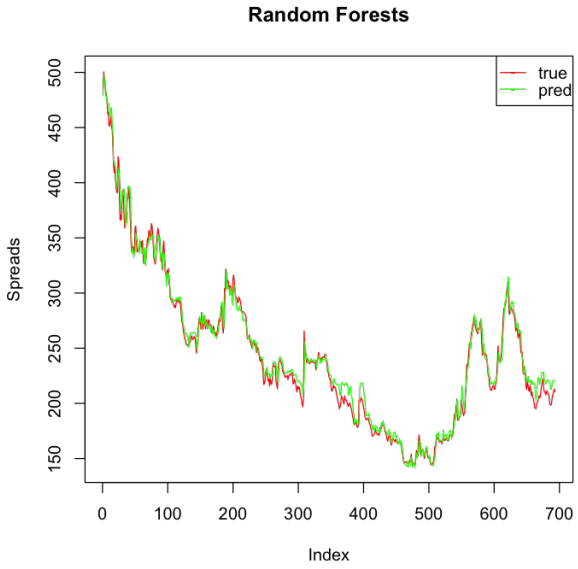


Figure 5: Random Forest Regression (Brazil)

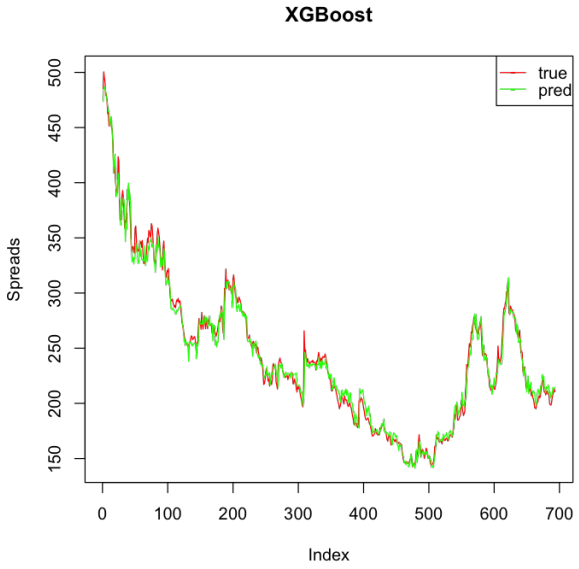


Figure 7: XGBoost (Brazil)

these methods also overfit giving low training and high test errors (**Figure 6**).

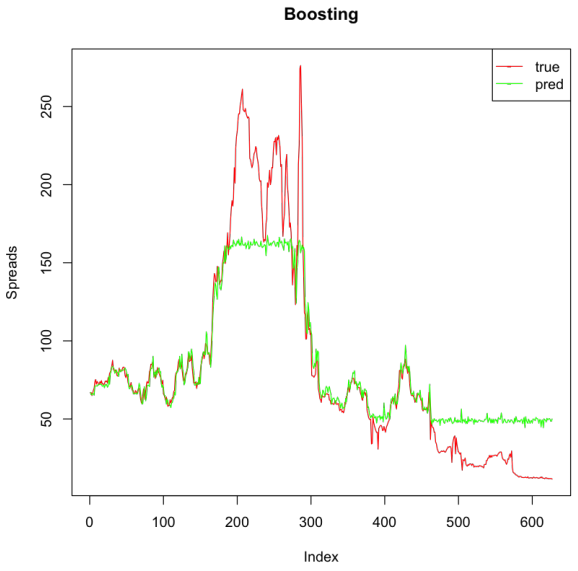


Figure 6: Boosting (China)

# 4 Future Scope

### Forward Looking Models

An approach to solving the problem of using lagged data to predict CDS Spreads could be to model out the future (i.e current) values of features based on previous trends and using these predictions to model CDS Spreads. While this adds a layer of estimation to the target variable, we felt this may be an interesting approach to consider if short term and long term trends in the feature set could be captured.

### Natural Language Processing (Text input)

While long term trends in CDS Spreads are likely more heavily influenced by economic and fundamental data, short term fluctuations and daily movements are governed more by headlines and news statements, which either take time to be reflected in data, or are not reflected at all. Using natural language processing techniques, this information can be quantified and included in the feature set, and possibly trained to capture short term fluctuations better in the target variable.

### Neural Networks (LSTM)

Recurrent neural networks such as Long Short-Term Memory networks are capable of learning order dependence in prediction problems, by giving importance to the information about past inputs for a variable amount of time depending on the weights. Such a network may be trained to capture short & long-term trends, which could vastly improve the prediction capabilities. This may be used to improve the previous method.

## 5 Conclusion

Initially we attempted to predict CDS spreads using linear regression and other linear techniques such as Lasso & Ridge regression. The linear regression likely memorized data in the manner of allocating too much weight to previous date's spreads. Regularization was able to provide good results for Brazil & China, but overfits for Russia, prompting the use of Non-Linear models to attempt variance & bias reduction. PCR underperformed severely with high overfit for all countries. Ensemble methods like Random Forests & Boosting gave promising results for Brazil, but underperformed for China & Russia. The models are unable to fit noise well, suggesting that the data is not enough to capture the volatility in these countries' spreads. There is plenty of room for improvement in this model before it can be utilized in a market strategy. The complexity in CDS spread movements is difficult to model using purely regression techniques & numerical data.

## 6 Appendix

| | Brazil | | China | | Russia | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| Linear Regression | 49.352 | 51.121 | 9.784 | 52.014 | 38.118 | 405.681 |
| Ridge Regression | 51.509 | 26.838 | 10.33 | 62.928 | 41.598 | 473.872 |
| Lasso Regression | 55.684 | 23.205 | 9.866 | 46.335 | 38.721 | 425.828 |
| PCR | 125.343 | 333.104 | 12.822 | 152.561 | 80.706 | 1422.498 |
| RandomForest | 12.795 | 60.815 | 1.952 | 571.113 | 9.302 | 7971.681 |
| Boosting | 48.842 | 31.313 | 7.264 | 714.942 | 30.712 | 5865.04 |
| XGBoost | 13.499 | 49.006 | 3.637 | 552.603 | 16.655 | 9066.199 |

Figure 8: Tabulated results

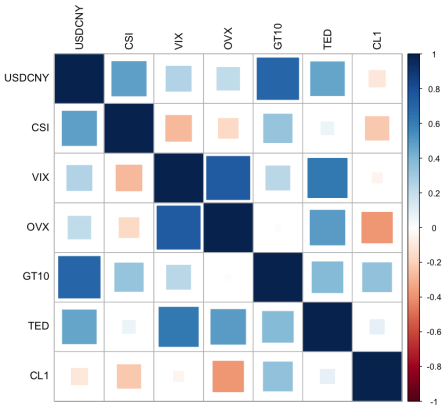


Figure 9: Correlation plot for China

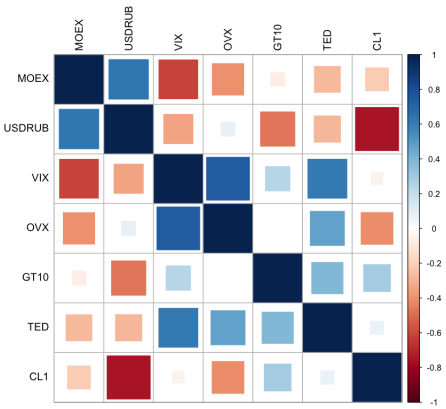


Figure 10: Correlation plot for Russia

## References


[1] Naifar, N., 2020. What explains the sovereign credit default swap spreads changes in the GCC region?. *Journal of Risk and Financial Management*, 13(10), p.245.

[2] Mercadier, M. and Lardy, J.P., 2019. Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research*, 277(1), pp.351-365.

[3] http://xgboost.readthedocs.io/en/latest/