Diplomacy and Dollars: Predicting Exchange Rate Change Using Covert State Department Cables

Jieyu Wang
New York University
jw4937@nyu.edu

Yu-Ping Lin
New York University
ypl238@nyu.edu

Hetian Bai
New York University
hb1500@nyu.edu

Abstract

The goal of this study is to investigate whether Covert State Department Cables from Wikileaks could predict the return of exchange rate over time. By applying random forest models with cable information and time series exchange rate as features, we discovered that although covert state department cables do not demonstrate strong predictive power toward exchange rate in general, for certain countries like Korea and Canada, cable features can help improve the performance of predictive models. In addition, we observed that cables in 2008 has stronger predictive power on predicting the return of exchange rate.

1 Introduction

Exchange rate is a key indicator of trading economies. It represents the confidence level in the economy or currency. The movement of exchange rate sometimes can be explained by risks and benefits to cooperations and countries that are holding the foreign assets. In addition to a country's economic growth, deficit or surplus...etc, a line of economic research Reinhart and Rogoff (2004) gives solid evidence conclude that political and economic events also have significant impacts on exchange rate. Thus, we believed that confidential government documents could also disclose information that implies politic and economic risk. In this project, we investigated whether confidential cables from the Public Library of United States Diplomacy (PlusD) effect the return of currency exchange rate over time. Besides, we conducted country-wise and year-wise analysis to fully understand the effects. In terms of research methodology, we applied natural language processing techniques to extract information from textual data and utilize machine learning algorithms to capture the patterns and relationships between PlusD cables and the change of exchange rates. In the end, we try to answer whether confidential cables, PlusD, can bring more predictive power over the exchange rate data.

2 Related Works

Some previous works devoted to combine text and financial data. Gidofalvi and Elkan (2001) and Fung et al. (2003) used news articles to predict stock price. It transformed stock price movement as variable "up", "down", and "unchanged" with Bayesian text classifier. This article discovers that, between 20 minutes interval before and after the news becomes publicly available, news articles have definite predictive power. Mittermayer (2004) further worked on state of art model "SVM" to improve the predictive model's performance. With prior domain knowledge of financial market, the article classified news as "Good" and "Bad" by the words showed in the news. And they developed intra day trading strategy with prediction of stock market movement direction. To further study how investors affected by news, Zhai et al. (2007) applied WordNet with "SVM" model and focused on showing text data improving predictions. Although these previous works gave significant results, all of them were limited on classification task.

To generalize the prediction on numerical value, Schumaker and Chen (2009) deployed "SVM" models to predict stock numerical price instead of movement direction. It also explored different methods to represent text data, ex: named entities and bag of words. In terms of performance, the named entities methods works better than bag of words due to the correlation between stock price and news articles. Bollen et al. (2011) and Ruiz et al. (2012) provided simi-

lar works on different sources, micro blogs and twitters respectively, predicting U.S. stock market.

To study how confidential information affects stock price, Dube et al. (2011) used regression and events driven models to estimate the impact of CIA documents on asset price. It showed coups and top-secret coups significantly impacted on the stock price. Berger et al. (2013) also used CIA document to show CIA interventions affected the pattern of trade during cold war. These previous work showed confidential cables have predictive power on various events. Although the coups were well connected to selected corporation, the content of coups was ignored in its methodology.

Similar works on exploring different financial market, bonds, and currency, also indicated that text data have predictive power. Ito and Roley (1986) examine how news from U.S. and Japan moved the Yen/Dollar currency. Peramunetilleke and Wong (2002) made the same statement that by using news data which are text that describing the current status of world financial markets, political, and general economic news, into the system could yield results that are significantly better than random prediction. Balduzzi et al. (2001) investigated the relation among public news and spread of bid and ask of U.S. treasury. Andersen et al. (2007) navigated high frequent trading trading data, and provided evidence that news produced conditional mean jumps over stock, bond and exchange market.

3 Data Description

There are two data sources that are used in this study, one is cable data, Public Library of US Diplomacy (PlusD) and another is financial data, Federal Reserve Bank reports. The first dataset was retried by crawling from PlusD website and the second was retried from the Wharton Research Data Service (WRDS). These two datasets are stored on the Azure server. You can access the PlusD dataset through directory `\WorkData\wikileaks\Crawled_Data`, whereas the directory for exchange rate (Federal Reserve Bank Reports) is `\WorkData\wikileaks\crsp_data`

3.1 Public Library of US Diplomacy

The Public Library of US Diplomacy (PlusD) covers US involvements in diplomatic and intelligence reporting on many countries and most of them are confidential. Cables in PlusD include e-mails, informal or formal reports, government documents, etc. In our study, the PlusD dataset crawled is from 2000/01/01 to 2010/12/31. The total number of cables is 250,254 and the average length of cables is 561.24 words. There are only two cables that do not contain any content in 2006. The table1 shows the exact number of cables and the average length of cable for each year between 2000 to 2010:

Statistics for Wikileaks Cables			
Years	Numbers	Average Length	Missing
2000	432	560.67	0
2001	878	557.55	0
2002	2,570	568.63	0
2003	8,083	570.15	0
2004	11,812	593.41	0
2005	24,131	538.95	0
2006	42,391	545.87	2
2007	44,693	568.63	0
2008	49,446	542.98	0
2009	56,813	553.74	0
2010	9,005	545.9	0
Total	250,254	561.24	2

Table 1: PlusD statistics: covers period from 2000 to 2010. Total cables numbers are 250,254. Total average length of each cable is 561.24

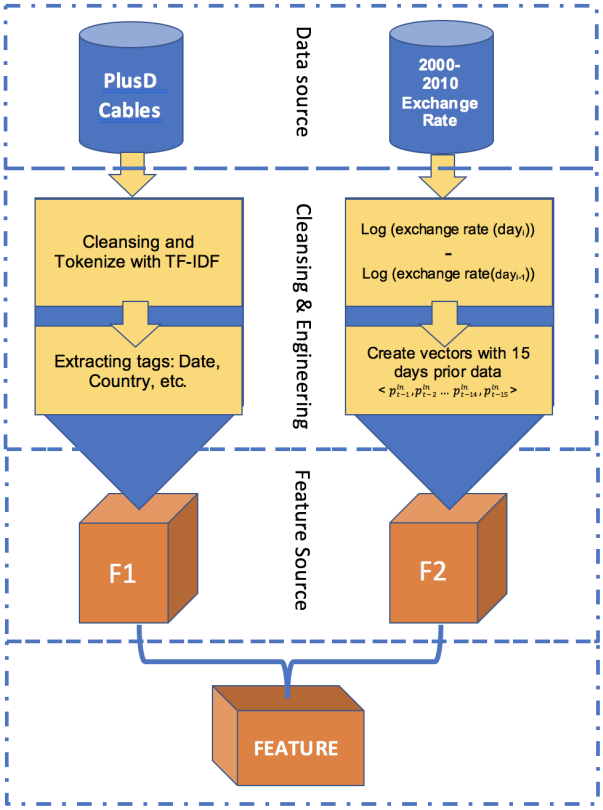
3.2 Federal Reserve Bank Reports

Exchange rate is calculated by comparing world currency rates with the US dollar. The detailed exchange rate data is obtained from the Federal Reserve Bank Reports by Wharton Research Data Services. The dataset covers twenty-two countries or regions' exchange rate data between 2000/01/03 and 2010/12/31. The twenty-two currencies are from Australia, Brazil, Canada, China, Denmark, Hong Kong, India, Japan, Korea, Malaysia, Mexico, New Zealand, Norway, Sweden, South Africa, Singapore, Sri Lanka, Switzerland, Taiwan, Thailand, United Kingdom, and Venezuela. The total number instances is 2,768 trading days during the period. The following table 2 show the statistics:

Statistics for Exchange Rate		
Currency	Mean	Std
Australia	1.4363	0.2865
Brazil	2.2953	0.5205
Canada	1.2691	0.2001
China	7.7914	0.6134
Denmark	6.3917	1.1561
Hong Kong	7.7838	0.0205
India	45.4879	2.5429
Japan	109.7948	11.6077
Korea	1131.2774	133.1647
Malaysia	3.6324	0.2189
Mexico	10.9582	1.2887
New Zealand	1.7001	0.3732
Norway	6.9339	1.1769
Sweden	1.6118	0.1518
South Africa	7.6577	1.3554
Singapore	101.0353	10.8086
Sri Lanka	7.9708	1.2661
Switzerland	1.3119	0.2296
Taiwan	32.8109	1.3063
Thailand	38.0737	4.5330
United Kingdom	0.6004	0.0686
Venezuela	1.9073	0.9251

Table 2: Exchange Rate statistics: covers period from 2000 to 2010. Total trading days is 2,768. It shows the mean and standard deviation of exchange rate per country

Figure 1: Data Preprocessing Work-flow Diagram



4 Data Preprocessing

4.1 Feature Generation

Data Preparation contains three main procedures 1: data cleansing, feature engineering, and feature merging.

4.1.1 Data Processing on PlusD Cables

To tokenized text data in each document of PlusD Cables, we utilized TF-IDF (Pedregosa et al., 2011) from Scikit-Learn in Python. TF-IDF would help us with data cleansing and convert raw documents into n-grams sparse matrix[1]. At the same time, we saved document tags including Cable Date, Country, etc from raw Cabel data. Thus, we were able to use tags to filter out what countries the cable was related to. With this relationship, it is easy to match the cable with exchange rate corresponded to each country.

4.1.2 Data Processing on Exchange Rate

Exchange rates p are time series data. We applied domain knowledge that exchange rate could be correlated to the rates in previous days. Thus, for each day t, we have a feature $(p_{t-1}, p_{t-2}, \ldots, p_{t-15})$ which represents exchange rates in previous 15 days. Moreover, for financial time series data, logarithmic returns are much stable over time. Thus, we defined the first order logarithmic difference p_t^{ln} is $\ln p_t - \ln p_{t-1}$, and the feature is $(p_{t-1}^{\ln}, p_{t-2}^{\ln}, \ldots, p_{t-15}^{\ln})$. We concatenated these two vectors for each day t as our exchange rate feature.

4.2 Feature Engineering

Features are generated from Wikileaks cables and exchange rates by bridging text data and financial data together. We projected these data into high dimensional vector space and concatenated them. For each day t, if there is no corresponded cable data, we simply projected the data as 0 vector. Since our financial data, exchange rate could not be zero, the 0 vector would not conflict to any in-

[1]Hyper-parameter setting of *TfidfVectorizer*: TfidfVectorizer(binary=True, ngram_range=ngram_range = (1,2), stop_words = 'english'

stance. This is one trick how we let the machine do event driven learning.

4.3 Target Variable Generation

Labeling on abnormal change in exchange rate is defined and operated by the following method on exchange rate data:

First, we normalized the exchange rate into the interval $[0, 1]$ over 10 years. This method prevents machine favoring the vector with large quantitative size. As a result, we have the exchange rate normalized by each currency.

Second, we took 15 days prior and 15 days posterior (a 30-day window) to calculate the change percentage of exchange rate. Calculation followed by:

$$\text{Exchange Change Rate} = \frac{\text{avg}(p_{t-15} + p_{t+15}) - p_t}{\text{avg}(p_{t-15} + p_{t+15})}$$

Third, in the task of regression, the return of exchange rate is the target variable, which is continuous. In the task of classification, we defined that the top 10 percentile as abnormal return, that is, the top 10 percentile were labeled as 1, otherwise, 0. Thus, we generated the binary outcomes as target variables. In the model, we utilized features we generated to make predictions on the continuous and binary target variables.

5 Models

Based on the feature we extracted from PlusD Cables and historical currency exchange rate, we built two models: binary random forest classifier and random forest regression. We first built a binary classification model with random forest algorithm to predict whether there was an abnormal change in exchange rate on a day. This model is served as a probe to explore whether it is meaningful to conduct further experiments with regression models using PlusD cable features. After that, we made a series of regression models, which directly predict the change of exchange rate from the perspective of currency and year.

5.1 Random Forest Classifier

The previous works, Gidofalvi and Elkan (2001) and Fung et al. (2003), showed the classifier was good at predicting stock return. With the Public Library of US Diplomacy cable vectors, we trained a Random Forest classier with 60 estimators to predict the existence of an abnormal change in exchange rate [2]. We deployed package `sklearn.ensemble RandomForestClassifier`. The binary predicting labels were described in section 4.3.

5.2 Random Forest Regression

Besides classification analysis, we did regression to analyze how Wikileaks cables affect the continuous rate of change of currency exchange rates in currency and year levels. We set up log return for the last 15 days as our features and log return as the target value. We deployed package `sklearn.ensemble RandomForestRegressor` functions to implement Random Forest model with the 30 estimators [3]. To study country-wise and year-wise effects of cables data on exchange rate, we split training and testing set and trained regression model respectively.

6 Evaluation Metrics

1. For binary classification models, results predicted are binary, thus we have AUC (Area Under an ROC Curve) as the evaluation metrics. AUC is measured by the area under the Receiver operating characteristic curve. An area of 1.0 represents a perfect test; an area of 0.5 represents a worthless test.

2. For regression models, we have Median Absolute Error and R^2 coefficients as evaluation metrics. MedAE 1 is short for Median Absolute Error, which measures the median of the magnitude of the errors in a set of predictions. A small MedAE indicates a good prediction results from regression model.

$$\text{MedAE}(y, \hat{y}) = \text{Median}(|\hat{y_1} - y_1|, ..., |\hat{y_n} - y_n|) \quad (1)$$

R^2 2 known as the coefficient of determination, is the second metric we have to evaluate regression model. The best possible score is 1.0 and it can be negative (because the model

[2] Hyper-parameter of Random Forest Regressor model: RandomForestClassifier(n_estimators=60, verbose=100)

[3] Hyper-parameter of Random Forest Regressor model: RandomForestRegressor(n_estimators=30, criterion=mse, max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=auto, max_leaf_nodes=None, min_impurity_decrease=0.0)

can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \text{Mean}(\hat{y}))^2} \quad (2)$$

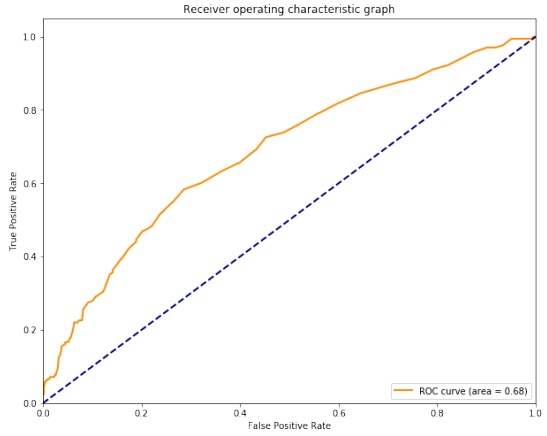
MedAE is a good indicator for showing the distribution of magnitude prediction error, whereas R^2 represents the general information of the model.

7 Random Forest Classifier

7.1 Results & Analysis

In this section, we checked the predictive power of pure text feature that we extracted from PlusD cables. From Figure 2, we observed that the binary classification model has an AUC of 0.68, which means, by giving the text features only, the classifier shows an average predictive power. This indicates cable features is a predictor to forecast an abnormal change in exchange rate, but the predictive power is not strong. We infer that PlusD Cable information does show predictive power, even though very weak.

Figure 2: ROC curve with pure cable text features



8 Random Forest Regression

In this section, we conducted experiments with three levels: first, using full dataset to study the overall relationship between PlusD cables and exchange rate; second, by slicing dataset by years to see the variation on the influences by year; third, by slicing dataset by country to study the how counties' currency exchange rate react to PlusD cables. In each experiment, we trained and tested three regression tree models 5.2 by using different sets of features, which are pure Cable text features, pure numerical exchange rate features, and cable-exchange-rate combined features.

We initially experimented on full dataset and found that data is more sensitive to negative return. It is believed that financial market is more sensitive to bad news (Koutmos and Booth (1995) and Veronesi (1999)). Thus, this section showed and discussed the experiment results for instances with negative return. The following are results and analysis of three experiments.

Pure Cables Text Features In this experiment, we only included text features and did n-gram featuring on text with package `TfidfVectorizer`, the parameter n gram is 1 to 2. Mapping words to vectors and did random forest regression. (refer to Data-Prep Figure 1)

Pure Numerical Exchange Rate Features – last 15 days log return In this part, we only included last 15 days log return as features. The log return, first order logarithmic difference, is p_t^{ln} is $\ln p_t - \ln p_{t-1}$. The numerical feature is $(p_{t-1}^{\ln}, p_{t-2}^{\ln}, \ldots, p_{t-15}^{\ln})$. (refer to Data-Prep Figure 1)

Combined Cable-Exchange-rate Features We then mixed these two features together with `scipy.sparse.hstack`. So, our random forest regression model would have log returns and text features.

8.1 Analysis All Countries and Years

We first examined on the whole feature set we had. The setting is same as section 5.2. However, we found the sparse matrix of text features contains a high degree of noise, and it diluted the predictive power of the model. Thus, we selected top 50 important features and dropped the rest. It has significant improved model performance for text data featured model. In this way, the Wikileaks cables features are much informative and representative. The sample feature importance was showed in figure 3

Since we are interested in how Wikileaks could effect the exchange rate, we focus on study the performance of models with mix features. The result was showed in table 3. With only text features,

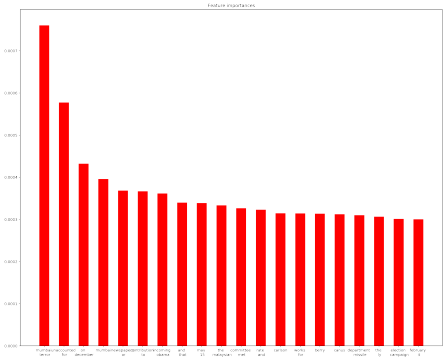


Figure 3: the sample of feature importance

the prediction was poor. However, if we dropped features and only keep top 50 important features, we improved the model with mix features on both Median Absolute Error and R^2 score. With this observation, we created the modeling procedure as follows:

– Building the model with random forest regression
– Selecting top 50 features
– Retraining the model

The next section follows this procedure and using sliced data by year and by country.

8.2 Analysis by Year

We sliced the data by years from 2000 to 2010. The number of cables with negative log return (corresponded to its currency) are various cross years. From 2000 to 2004 and 2010, we have comparatively small amount of cables each year. In contrast, the rest years have more than a thousand cables that means our models could learn more from Wikileaks cables over these periods. The result was showed in table 4.

Even though we kept top 50 important features, the models with mixed feature were generally perform worse than models only with past 15 days exchange rate features. However, for the year 2008, we found models with mixed features perform better than the other two settings. The R^2 score was improved from 0.8311 to 0.8331. The Median Absolute Error were improved as well. We initially believe that the negative log return would related to 2008 global financial crisis. Indeed, the model could capture information from the Wikileaks cables related to the financial crisis. These potential relations are worth to explore in the future work.

8.3 Analysis by Country

In addition to the analysis in year level, we conducted regression experiments on 22 currencies to study the variation on effects from Cables information. We are interested in to find out which country is most sensitive to Wikileaks cables. Results shown in table 5.

We observed that R^2 of Korea by using mixed feature is 0.8150, without Cable text feature, R^2 is merely 0.3769. This indicates that by adding cable information into regression model helps to capture the pattern of change of exchange rate better and makes predictions more accurately. Similarly scenario shows on Canada, where there is a rise in R^2 by adding cable features. However, for rest currency, cable features dilute the performance of regression model. Additionally, we observed that for majority countries or regions, history exchange rate change is a strong predictor, which could give perfect regression model with high R^2. For example, the R^2 of Japan with pure Exchange rate feature is 0.9837. However, there are few in the list showing that both Exchange rate feature and pure cable features are not good to make predictions, such as Hong Kong and Singapore.

9 Conclusion and Future Work

In this article, we explored different models and attempted to answer whether there is predictive power of Wikileaks cables. We firstly did random forest classifier model and inferred that PlusD cables does show predictive power, even though very weak. However, we are more interested about predicting the change of exchange rate values. It is believed that to predict exchange rate, historical rate was an valuable predictor. Thus, we took the time series exchange rate model as baseline model. We designed our experiments with three types of features:

– Pure text feature
– Pure exchange rate features (baseline model)
– Combined text feature and exchange rate features

We want to explore that if we add text features to the baseline model, which only contained exchange rate, the predictive power could be enhanced or not. Since text feature was created into sparse matrix by `TF-IDF`, we only kept top 50

Result on Test Set						
	Text		EX Rate		MIX	
Drop features	MedAE	R2	MedAE	R2	MedAE	R2
Not Drop	0.0010	0.0665	0.0001	0.8992	0.0006	0.5681
Drop	0.0011	-0.2165	0.0001	0.8290	0.0002	0.7620

Table 3: All countries and years statistic. Random forest regression model on test set with three types of features. $MedAE$ is Median Absolute Error, $R2$ is R-square. The scale is %.

Result on Test Set							
		Text		EX Rate		MIX	
Year	Cables number	MedAE	R2	MedAE	R2	MedAE	R2
2000	9	0.0014	0.0	0.0004	0.0	0.0012	0.0
2001	30	0.0001	0.9071	0.0001	0.8511	0.0001	0.9295
2002	106	0.0009	-0.2936	0.0003	0.6822	0.0005	0.2350
2003	667	0.0013	-0.0096	0.0004	0.3780	0.0005	0.2956
2004	861	0.0011	-0.1457	0.0003	0.2683	0.0007	0.4617
2005	2059	0.0009	-0.1148	0.0002	0.7733	0.0003	0.6153
2006	4420	0.0009	-0.3802	0.0001	0.7457	0.0002	0.7189
2007	5085	0.0008	-0.1347	0.0003	0.8704	0.0005	0.6989
2008	5522	0.0014	-0.3094	0.0001	0.8311	0.0002	0.8331
2009	6681	0.0014	-0.5390	0.0001	0.8047	0.0001	0.7747
2010	858	0.0011	0.0453	0.0001	0.9072	0.0001	0.8704

Table 4: The year by year statistic for random forest regression model on test set with three types of features. $MedAE$ is Median Absolute Error, $R2$ is R-square. The scale is %.

important features in order to improve the predictive power. We then sliced the data by year and by country to see which model is sensitive to text features. In other words, the Wikileaks cables might improve the prediction for specific years or currencies. We examined the quality of data and excluded the one with only a few cables. Through experiments, we obtained some interesting finding.

For models by year:

- the performance of model only with text feature is poor
- the performance of model only with exchange rate feature is good (time series data)
- the performance of model with mix feature is poor
- however, for the year 2008, the predictive power was improved.

For models by country:

- the performance of model only with text feature is poor
- the performance of model only with exchange rate feature is good (time series data)
- the performance of model with mix feature is poor
- however, for Korea and Canada, the predictive power was significantly improved.

These findings show that although PlusD cables might not have strong predictive power toward exchange rate in general, by combing both cables and exchange rate features, for some countries and years, PlusD cables can still have predictive power.

9.1 Future Work

At the initial stage of our project, we planned to use company stock price and exchange rate to discuss the effect of confidential cables from the Public Library of United States Diplomacy on financial markets across the world. However, we found that it is difficult to obtain a robust list of companies which could represent each country, and access its corresponding historical stock data. Therefore, in this report, we only worked with on exchange rate, whose dataset is easily to access. Thus, for the future work, if the required stock datasets are available, we can investigate how the PlusD cables affects companies' stock price.

Result on Test Set							
		Text		EX Rate		MIX	
Currency	Cables number	MedAE	R2	MedAE	R2 rate	MedAE	R2
Australia	676	0.0025	0.4399	0.0003	0.8605	0.0008	0.6571
Brazil	1607	0.0017	-0.1972	0.0002	0.9202	0.0005	0.6504
Canada	1388	0.0002	-0.4101	0.0001	0.5572	0.0002	0.4930
China	4163	0.0002	-0.0303	0.0001	0.6339	0.0002	0.0632
Denmark	184	0.0015	-0.2246	0.0007	0.0463	0.0014	-0.4428
Hong Kong	442	0.0001	-0.0109	0.0001	-0.0192	0.0001	-0.0119
India	2361	0.0009	-0.1700	0.0001	0.9303	0.0002	0.7325
Japan	3186	0.0012	-0.1077	0.0001	0.9837	0.0001	0.8564
Korea	1369	0.0016	-0.5129	0.0002	0.3769	0.0002	0.8150
Malaysia	579	0.0010	-0.7178	0.0003	0.5743	0.0006	0.2448
Mexico	1446	0.0011	-0.1475	0.0001	0.8701	0.0002	0.7497
New Zealand	735	0.0016	-0.2524	0.0003	0.7841	0.0007	0.5530
Norway	515	0.0022	-0.2931	0.0010	0.1003	0.0014	0.0417
Singapore	312	0.0005	-3.8112	0.0002	-0.5842	0.0002	-0.9066
South Africa	956	0.0021	-0.9765	0.0005	0.6886	0.0010	0.4410
Sri Lanka	1576	0.0020	-0.3300	0.0004	0.8396	0.0007	0.4788
Sweden	391	0.0009	-0.3982	0.0003	0.4195	0.0005	0.2887
Taiwan	1846	0.0006	-0.3492	0.0002	0.7357	0.0002	0.5578
Thailand	1576	0.0010	-0.4801	0.0001	0.8820	0.0002	0.5790
United Kingdom	962	0.0013	-0.1495	0.0003	0.8952	0.0005	0.7813
Venezuela	28	0.0008	-0.0468	0.0003	0.4177	0.0007	-2.1995

Table 5: The country by country statistic for random forest regression model with three types of features on test set. $MedAE$ means Median Absolute Error, $R2$ is R-square.

Code and Data

Readers are encouraged to visit the team's Github directory which contains all the implementations. The README.md shows the list of scripts.

To see the list of data preprocessing and other scrips, visit here

Data path

We organized the path of data in file in `data_description.md`. Please see the link: Github

Acknowledgment

We thanks Dr. Elliot Ash and Dr. Daniel Chan for their mentorship and providing Azure server for this project to store data and do computation. NYU provided library CRSP and high performance computer Prince provided cloud space and computation.

References

Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Clara Vega. 2007. Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of international Economics*, 73(2):251–277.

Pierluigi Balduzzi, Edwin J Elton, and T Clifton Green. 2001. Economic news and bond prices: Evidence from the us treasury market. *Journal of financial and Quantitative analysis*, 36(4):523–543.

Daniel Berger, William Easterly, Nathan Nunn, and Shanker Satyanath. 2013. Commercial imperialism? political influence and trade during the cold war. *American Economic Review*, 103(2):863–96.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Arindrajit Dube, Ethan Kaplan, and Suresh Naidu. 2011. Coups, corporations, and classified information. *The Quarterly Journal of Economics*, 126(3):1375–1409.

G Pui Cheong Fung, J Xu Yu, and Wai Lam. 2003. Stock prediction: Integrating text mining approach

using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402. IEEE.

Gyozo Gidofalvi and Charles Elkan. 2001. Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*.

Takatoshi Ito and V Vance Roley. 1986. News from the us and japan: which moves the yen/dollar exchange rate?

Gregory Koutmos and G Geoffrey Booth. 1995. Asymmetric volatility transmission in international stock markets. *Journal of international Money and Finance*, 14(6):747–762.

M-A Mittermayer. 2004. Forecasting intraday stock price trends with text mining techniques. In *system sciences, 2004. proceedings of the 37th annual hawaii international conference on*, pages 10–pp. IEEE.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Desh Peramunetilleke and Raymond K. Wong. 2002. Currency exchange rate forecasting from news headlines. *Aust. Comput. Sci. Commun.*, 24(2):131–139.

Carmen M Reinhart and Kenneth S Rogoff. 2004. The modern history of exchange rate arrangements: a reinterpretation. *the Quarterly Journal of economics*, 119(1):1–48.

Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM.

Robert P Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12.

Pietro Veronesi. 1999. Stock market overreactions to bad news in good times: a rational expectations equilibrium model. *The Review of Financial Studies*, 12(5):975–1007.

Yuzheng Zhai, Arthur Hsu, and Saman K Halgamuge. 2007. Combining news and technical indicators in daily stock price trends prediction. In *International symposium on neural networks*, pages 1087–1096. Springer.