Machine Learning for Statistical Arbitrage: Using News Media to Predict Currency Exchange Rates

Samaskh Goyal (sagoyal), Hari Sowrirajan (hsowrira), Teja Veeramacheneni (tejav)

***Abstract*— We explore the application of Machine Learning for predicting bilateral Foreign Exchange Rates utilizing the sentiment from news articles and prominent macroeconomic indicators. Using a random forest regressor, we were able to predict foreign exchange rates with an average error of 7.8%. In addition, news articles were an important feature in the majority of these random forest regressions. The novelty of our project relies on utilizing the Latent Dirichlet Allocation to cluster news articles into topics and further understand the semantic meanings behind each topic and applying it to foreign exchange rates.**

I. INTRODUCTION

Artificial intelligence is being adopted by the financial industry at breakneck speed - primarily through FinTech firms breaking down financial services traditionally operated by institutional banks. While initially this caused a major rivalry due to the obvious conflict of interest, recently both parties have recognized that in the long run these enhancement provide better client service and financial security which benefits both individuals and businesses[1].

Machine learning is being integrated into the financial sector through process automation, security enhancement, underwriting and credit scoring, algorithmic trading and portfolio advisement.[2] The most lucrative of these is algorithmic trading: a process by which trades are conducted at high speeds and volume based on a number of preset criteria. In the literature this type of financial forecasting is heavily focused on stock prediction and uses several market indicators[3][4] .

In our paper, we attempt to investigate a less studied asset class: foreign exchange rate. The literature suggests using data on macroeconomic variables such as GDP, interest rates, or more granular metrics such as currency-pairings and short term time horizons.[5] In our model, we take a novel approach and use news articles to predict the foreign exchange rate of 6 major country pairs. To this end we use the Latent Dirichlet Allocation (LDA) algorithm to cluster articles into topics and then use the associated weights as the features for our model. We hypothesize that this text-based framework will be able to effectively capture the factors driving the supply and demand of currency and provide accurate predictions on the exchange rate.

Specifically, the input to our algorithm is a year's news articles and macroeconomic indicators. We utilize several regression models, including Linear Regression, Ridge Regression, Random Forests, and Support Vector Machines, to then predict the next year's average currency exchange rate for a specific country pair.

II. RELATED WORK

Predicting foreign exchange rates is a well defined problem. Kamruzzaman et al. utilized artificial neural networks (ANN) and support vector machines (SVM) to predict next week's foreign exchange rate from the current week's rate. They are able to predict the directional change in exchange rate with around 80.0% and 83.0% test accuracy for the ANN and SVM, respectively[6].[7] In addition the National Bureau of Economic Research created a model that accounted for major economic constraints such as: "Behavioral equilibrium exchange rate", "Yield curve", and "Taylor's Rule". Their average error was 42%.[15]

However, such approaches rely solely on data that is directly connected with foreign exchange rates and do not explore whether the sentiment present in news articles is predictive of currency exchange rates. Utilizing news articles to make societal predictions is currently being studied in other contexts, namely conflict and stock price prediction. Mueller was able to summarize news text with topic models to predict military conflict one or two years before outbreak. He found that news data had a comparative advantage in predicting conflict and should be added as a feature to current conflict prediction models.[10] In addition, Schumaker and Chen show that utilizing breaking financial news has a statistically significant impact on predicting stock prices.[11]

III. DATASET AND FEATURES

A. Data Collection

There are three datasets that we utilized in this project: foreign exchange rates, news articles, and macroeconomic indicator data. For the foreign exchange data, we downloaded the yearly average foreign exchange rate from 1981 to 2016 for the Dollar-Yuan (China), Dollar-Rupee (India), Dollar-Yen (Japan), Dollar-Pound (Great Britain), Dollar-Franc (Switzerland), and Dollar-Canadian Dollar (Canada) from the Organization of Economic Cooperation and Development.[20]

For the news articles, we utilized the New York Times Article Search API to collect metadata from around 2000 articles per year (most relevant) from 1981-2015. We specifically searched for articles in the following economic-related sections: Your Money, Job Market, Business, World, Business Day, and Technology. We chose these sections because we hypothesized that articles from these sections would be most related to currency exchange rates. We then wrote a Python web scraper using Beautiful Soup to harvest 72,078 full-text articles corresponding to the metadata collected in the first step.[21]

For the macroeconomic indicators about a country, we collected data about the Gross Domestic Product (GDP) and Power Purchasing Parity (PPP) for the six countries whose exchange rates we were predicting from the World Bank website.[19]

Both the news articles and macroeconomic indicators would be used as features for predicting currency exchange rates.

B. Topic Modeling

To generate features from the news articles, we decided to use unsupervised learning to cluster the news articles. Ideally, each of these clusters would correspond to some type of generic topic. The proportion of yearly articles in each cluster would be the features for that year. To cluster the news articles, we utilized Latent Dirichlet Allocation (LDA).[16] LDA is a generative statistical model that clusters documents that are similar to each other. LDA can be described by the hyper-parameters α, β, and K, where α represents the Dirichlet prior for the topic distribution for a certain document, β represents the Dirichlet prior for the word distribution for a certain topic, and K represents the number of topics.

To cluster the news articles, we chose $\alpha = 0.2$ and $\beta = 0.2$ after experimenting with several values. We let K be $5, 10, 20, 25, 50$ to see which number of clusters would be optimal for the currency exchange prediction. After clustering the news articles, we discovered that the clusters had semantic meaning. For example, when $K = 5$, some of the words most associated with cluster 3 were "American", "military", "Russia", "Iraq", and "Israel", indicating that cluster 3 were news articles associated with the military and foreign policy. Some words most associated with cluster 2 were "European", "China", "trade", "industry", "oil", "Japan", indicating that cluster 2 were news articles associated with international trade and the monetary system. Figure 1 shows what the results of the clustering are.

Cluster	Possible Cluster Topic	Key Words
1	Corporate Earnings and Success	Company, Million, Sales, Shares, Revenue
2	Trade, International Economic Ties	European, China, Oil, Trade, Industry, Japan
3	War/Foreign Policy	American, Military, Russia, Iraq, Israel
4	Society	Family, Home, People Work, Women
5	Financial Institutions	Bank, Market, Rates, Tax, Bonds, Fed, Debt

Fig. 1. Semantic Meaning of the Clusters when K = 5

IV. METHODS

We attempted to predict the average next year's exchange rate with four different regression models; linear regression, ridge regression, support vector regression, and random forest regression. We utilized Scikit-Learn to implement these three models.

A. Linear Regression

Linear regression works well as an initial baseline measurement. Given an input vector x 2 R^{K+1}, where x contains the yearly proportion of the topics and current year's exchange rate, linear regression finds a θ such that $\theta^T x$ is the predicted next year's exchange rate. To find the θ, we choose the θ that minimizes the least-square cost function

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} \quad y^{(i)})^2$$

where n is the number of training examples, $y^{(i)}$ is the true label for examples i, and $x^{(i)}$ is the data for example i.

B. Ridge Regression

We also experiment with adding an L2 regularization term. This means we choose the θ that minimizes the following loss function

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} \quad y^{(i)})^2 + \lambda \sum_{i=1}^{K+1} (\theta_i)^2$$

where λ is a constant that determines the weights on L2 regularization term. We set $\lambda = 0.01$ after experimenting with several different lambda values.

C. Support Vector Regression

Support vector regression is another good model that might be able to capture non-linear trends in the data with certain kernels. We aim to find a w such that $w^T x + b$ is a prediction for the next year's foreign exchange rate. This w is found by minimizing

$$\frac{1}{2} jjwjj^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

where n is the number of training examples. This is subject to the constraints

$$\begin{aligned} y_i \quad wx_i \quad b \quad \epsilon + \xi_i \\ wx_i + b \quad y_i \quad \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \quad 0 \end{aligned}$$

After experimenting with polynomial kernel, linear kernel, and the gaussian radial basis function kernel, the radial basis kernel performed best. We use the radial basis function as a kernel for our data where for an $x^{(i)}$ and $x^{(j)}$

$$K(x^{(i)}, x^{(j)}) = \exp(\quad \frac{1}{2} jjx^{(i)} \quad x^{(j)} jj^2)$$

We set $C = 1.0$ and $\epsilon = 0.1$ after conducting tuning regularization.

D. Random Forest Regression

Decision tree regression is also another approach that can capture non-linear trends in a data set. However, because decision trees are prone to high variance that is dependent on random splits, random forests are often used as they aggregate the results from several decision trees. To train our model, we utilized a 100 decision trees with the splitting criterion

$$\sum_{i \in L} (y^{(i)} \quad y^{(L)})^2 + \sum_{i \in R} (y^{(i)} \quad y^{(R)})^2$$

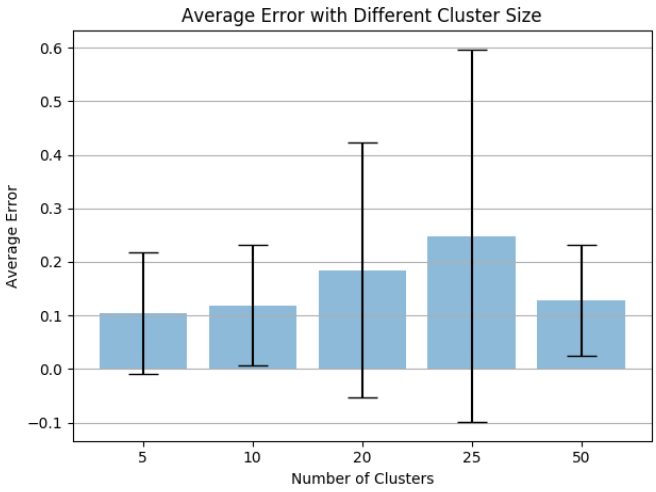

Fig. 2. Cluster Size vs Error

where $y^{(L)}$ is the average currency exchange rate in the left node and $y^{(R)}$ is the average currency exchange rate in the right node. Hyper parameters that were turned were minimum leaf samples and minimum samples for a split to reduce over-fitting.

V. RESULTS

To measure the accuracy of our models we find total percentage error of the foreign exchange rate prediction for each currency. For each currency and for ever year i between 2000 and 2015, we train the model on all years from 1981 to $i \quad 1$ and test the model for year i. For example, one of the predictions would be training on news articles from 1981-2000 and predicting on 2001. The percentage error for a year i is then defined as

$$\epsilon^i = \frac{\hat{y}^{(i)} \quad y^{(i)}}{y^{(i)}} \quad 100$$

where $\hat{y}^{(i)}$ is the predicted exchange rate and $y^{(i)}$ is the actual exchange rate. The average percentage error for a specific currency pair is then defined as the average of these ϵ values for that currency pair.

A. Optimal Number of Clusters

To determine the optimal number of cluster/topics we should use in our regression model, we compare the average total error of all the models when using news article weights when $K = 5, 10, 20, 25, 50$. The lowest average error was when $K = 5$. See Figure 2.

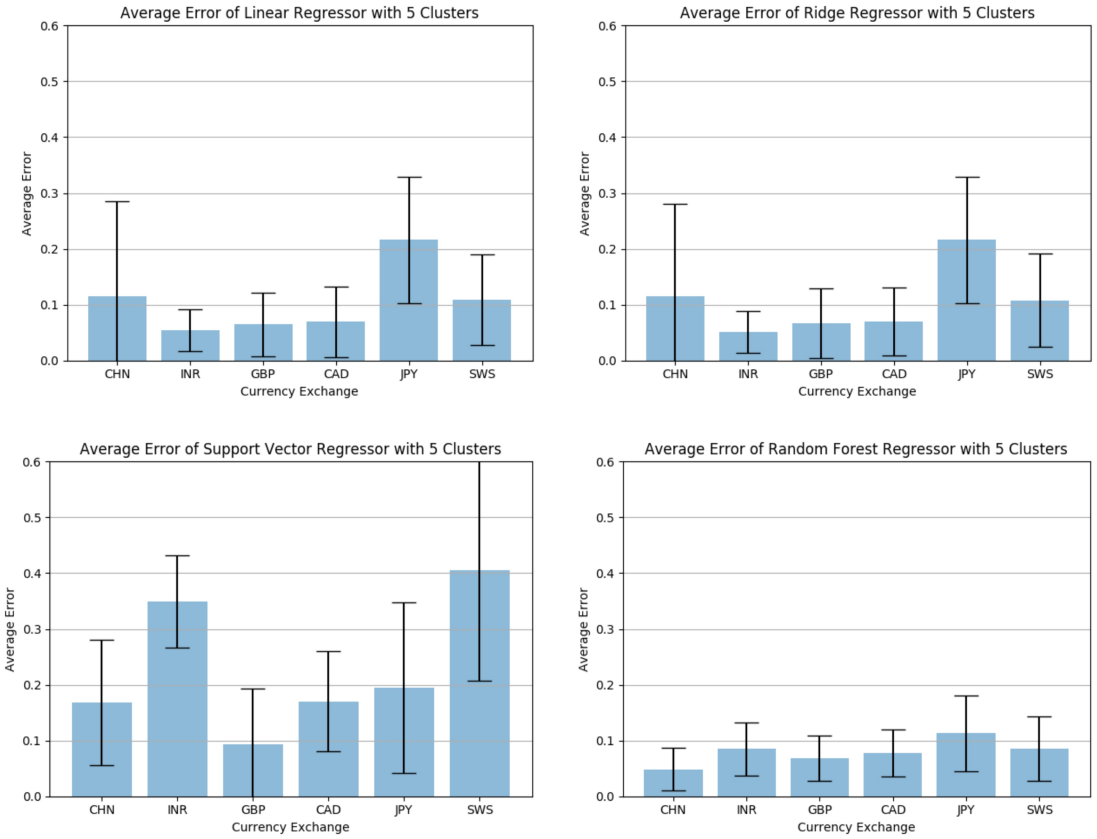

Fig. 3. Regression Models: Linear Regression, Ridge Regression, Support Vector Regressor, Random Forest Regressor

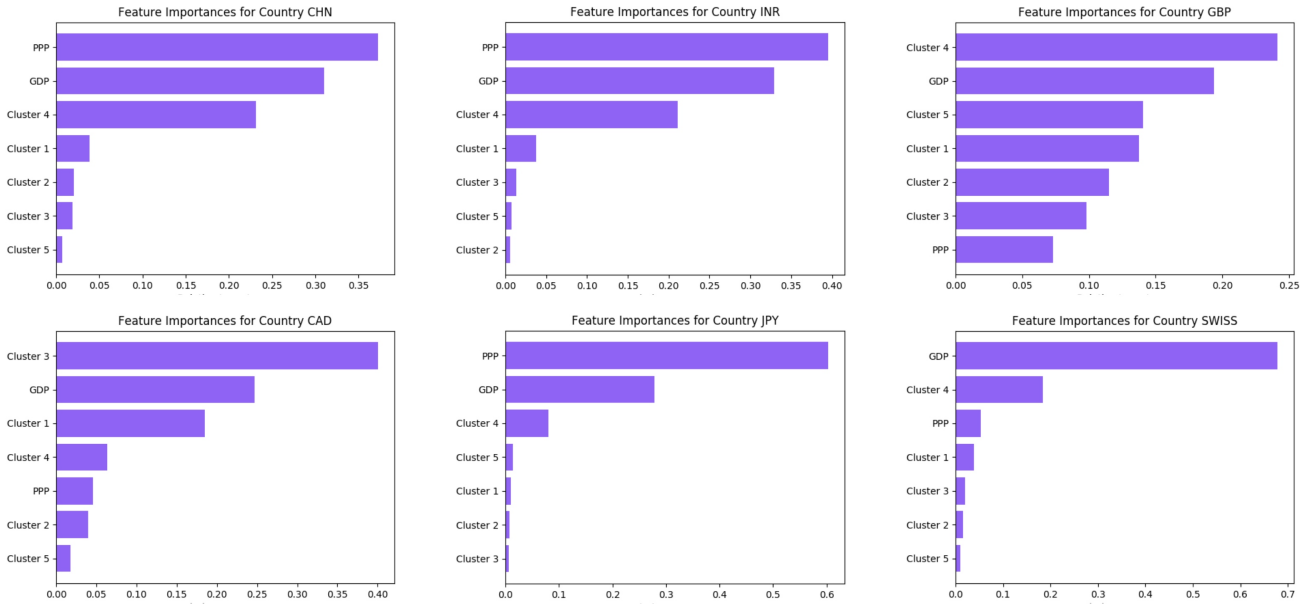

Fig. 4. Feature Importance per country

B. Model Comparison

We compared the four regression models when $K = 5$. We see that Random Forest Classifier performed the best (had the lowest average error and lowest standard deviation) across 5 of the currency exchange rates with an average error of 7.9%. Support Vector Regressor had the poorest performance having the highest average error across all 6 currencies with an average error of 23.0%. Linear and Ridge Regression produced average accuracies in between these with average errors of 10.6% and 10.4%, respectively. See Figure 3.

C. Feature Importance

Feature importance is a metric used in random forest regressors to determine which features are important to the prediction[17].[18] To determine the feature importance, we utilized the mean decrease impurity of a feature. Using this metric, a country's GDP per capita consistently remained a important feature for its foreign exchange rate. Purchasing Power Parity was a significant indicator for China, India and Japan. Cluster 4 was the most important news article topic for all countries except Canada. See Figure 4.

VI. DISCUSSION

Our model construction shows that Random Forest Regressors produced the lowest average error, the plausible reason for this is that it was able to catch the non-linear trends in the data, which the linear and ridge regression were not. In addition, the random forest regressor had the lowest standard deviation, indicating that the model was not overfitting and the proper hyperparamters were tuned.

The weights on the features indicate that GDP remained the most important features for the current years FER for the majority of countries tested. This was expected since the GDP captures the economic strength of a country. More interestingly, the PPP was only a significant predictor for non-European developing countries (China, India, Japan). Historically these countries have less stringent working conditions and are known to have a manufacturing heavy economic sector, thus the rate at which a market basket of goods (PPP) is priced is relatively high in their local currency.

Additionally Cluster 4 is the most significant news topic for the majority of the countries (all except Canada). This is interesting since the key words indicate that this cluster codes for the sentiment of "society" relating to foreign exchange rate. Society implies that the domestic perception about foreign exchange rate actually has an influence in driving foreign exchange rate. This not an uncommon economic condition. It has been shown that inflation rates are also heavily dependent on domestic expectation of inflation.[14] Further the sentiment of "society" in news captures the average health associated with the domestic economy on the home front.

Our random forest regression model performed favorably to an analysis conducted by the National Bureau of Economic Research. Their model was able to obtain a 42% average error, whereras our model was able to achieve 7.8% average error. This can be explained by the fact that we used random forests, which are able to capture non-linear trends in data, and news articles, which can capture underlying yearly sentiment, while their paper utilizes a random walk model, which might be susceptible to overfitting and does not utilize news articles.

VII. CONCLUSION/FUTURE WORK

There are multiple steps that can be taken to improve both the accuracy of our predictions and to fully validate if news articles are indeed significant predictors of foreign exchange rate. Primarily to elucidate the connection between news articles and foreign exchange rate we could gather articles from more distributors (apart from NYT), have access to full length articles and use different keywords such as Monetary and Fiscal Policy. Furthermore to better predict the overall accuracy of our model we can capture more granular economic variables like those used in stock price forecasting. These would include: "Sticky Price", "Country's Risk and Liquidity Factors". We could also include other variables such as the average inflation rate - since sustained inflation weakens the currency, as well as document domestic sentiment towards: the central bank, monetary policy and political stability.

The goal of this project was to try to analyze if the news was an appropriate way to predict the foreign exchange rate with many of the US's major trading partners. With our current results it would seem that this is plausible case. While news article sentiment itself does not seem to outweigh established global macroeconomic indicators like GDP and PPP, it does offer major feature importance for certain countries.

VIII. CODE

The github link for the code is https://github.com/HS189/FinancialArbitrage2

IX. CONTRIBUTIONS

Samaksh: Literature Research, Data Collection/Cleaning, Model Debugging, Report

Hari: Literature Research, NYT Collection, Model Debugging, Poster, Report

Teja: Literature Research, LDA Implimentation, Base Line Models, Random Forest, Visualizations, Poster, Report

REFERENCES

[1] World Bank. Global Financial Development Report 2017/2018: Bankers Without Borders. World Bank Group, 2017.

[2] Pavliv, Andrew. Machine Learning in Finance: The Why, What amp; How. Machine Learning in Finance: The Why, What and How, N-IX, 10 July 2018, www.n-ix.com/machine-learning-in-finance-why-what-how/

[3] Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert Systems with Applications 42.1 (2015): 259-268.

[4] Pyo, Sujin, et al. "Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets." PloS one 12.11 (2017): e0188107.

[5] Deng, Shangkun, et al. "Hybrid method of multiple kernel learning and genetic algorithm for forecasting short-term foreign exchange rates." Computational Economics 45.1 (2015): 49-89.

[6] Kamruzzaman, Joarder, Ruhul A. Sarker, and Iftekhar Ahmad. "SVM based models for predicting foreign currency exchange rates." Third IEEE International Conference on Data Mining. IEEE, 2003.

[7] Kamruzzaman, Joarder, and Ruhul A. Sarker. "ANN-based forecasting of foreign currency exchange rates." Neural Information Processing-Letters and Reviews 3.2 (2004): 49-58.

[8] Abreu, Gonalo, Rui Neves, and Nuno Horta. "Currency exchange prediction using machine learning, genetic algorithms and technical analysis." arXiv preprint arXiv:1805.11232 (2018).

[9] Tenti, Paolo. "Forecasting foreign exchange rates using recurrent neural networks." Applied Artificial Intelligence 10.6 (1996): 567-582.

[10] Mueller, Hannes, and Christopher Rauh. "Reading between the lines: Prediction of political violence using newspaper text." American Political Science Review 112.2 (2018): 358-375.

[11] Schumaker, Robert, and Hsinchun Chen. "Textual analysis of stock market prediction using financial news articles." AMCIS 2006 Proceedings (2006): 185.

[12] Krauss, Christopher, Xuan Anh Do, and Nicolas Huck. "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the SP 500." European Journal of Operational Research 259.2 (2017): 689-702.

[13] Abreu, Gonalo, Rui Neves, and Nuno Horta. "Currency exchange prediction using machine learning, genetic algorithms and technical analysis." arXiv preprint arXiv:1805.11232 (2018).

[14] Carlson, John A., and Michael Parkin. "Inflation expectations." Economica 42.166 (1975): 123-138.

[15] Cheung, Yin-Wong, et al. "Exchange rate prediction redux: new models, new data, new currencies." Journal of International Money and Finance 95 (2019): 332-362.

[16] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[17] Saeys, Yvan, Thomas Abeel, and Yves Van de Peer. "Robust feature selection using ensemble feature selection techniques." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2008.

[18] Breiman, Leo, et al. "Classification and regression trees. Wadsworth Brooks." Cole Statistics/Probability Series (1984).

[19] World Bank. "GDP (current US)." World Development Indicators, The World Bank Group, 2019, https://data.worldbank.org/indicator/ny.gdp.mktp.cd. Accessed 10-6-2019

[20] OECD (2019), Exchange rates (indicator). doi: 10.1787/037ed317-en (Accessed on 11 June 2019)

[21] The New York Times. Retrieved from https://developer.nytimes.com/