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

Samaksh Goyal, Hari Sowrirajan, Teja Veeramacheneni

Motivation and Overview

Problem

- Foreign Exchange Rates (FER) are highly indicative of global economic health and economic ties between countries.
- Current predictive models utilize granular, purely economic metrics: GDP, Trade Balance, etc.
- Does not consider how sentiment analysis from newspapers affects FER.

Our Solution

- Use news article sentiment to predict FER.
- Collect and Cluster news articles from past 40 years.
- Used proportion of articles per year falling in each cluster as an indication of sentiment to predict the next year's exchange rate.

Datasets and Feature Engineering

News Articles

- Utilized NYT Article Search API to collect metadata from 2000 articles per year (most relevant) from 1981-2016. Specific Article Categories: Your Money, Job Market, Business, World, Business Day, Technology.
- Used metadata from API to harvest 72,000 full-text articles.

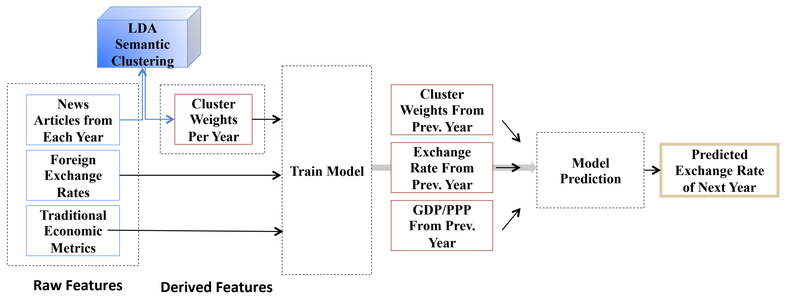
Foreign Exchange Rates

- Yearly Exchange Rates, relative to the USD, collected for the following countries from 1981-2016: China (CHN), India (INR), Great Britain (GBP), Canada (CAD), Japan (JPY), Switzerland (SWS).

Traditional Economic Indicators

- GDP, PPP per capita for each country as a contrast

Feature Modeling



Model Infrastructures

Latent Dirichlet Allocation (LDA)

- Hyperparameters: α (Dirichlet Prior for topic distribution), β (Dirichlet Prior for word distribution), and K (number of topics).
- We used: α = 0.2, β = 0.2, K = 5, 10, 20, 25, 50.

Clustering Results for K = 5

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

Models for Training/Prediction

Linear Regression

$x \in R^m \rightarrow m = \#\text{input features}$

$\theta^T x \rightarrow$ next year's rate

$\downarrow$

Baseline

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

Ridge (L2)

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

$\lambda = 0.01$

Non-Linear SVR

$w^T x + b \rightarrow$ next year's rate

$$\text{minimize: } \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i + \xi_i^* \text{ with } C = 1$$

$$y_i - wx_i - b \le \varepsilon + \xi_i$$
$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

$\varepsilon = 0.1$, Kernel: Radial Basis

Random Forest Regression

Generated 100 Decision tree with following splitting criterion (Residual Sum of Squares):

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

$y^{(L)}$: mean y-value (FER) for left node

$y^{(R)}$: mean y-value (FER) for right node

Performance of Models

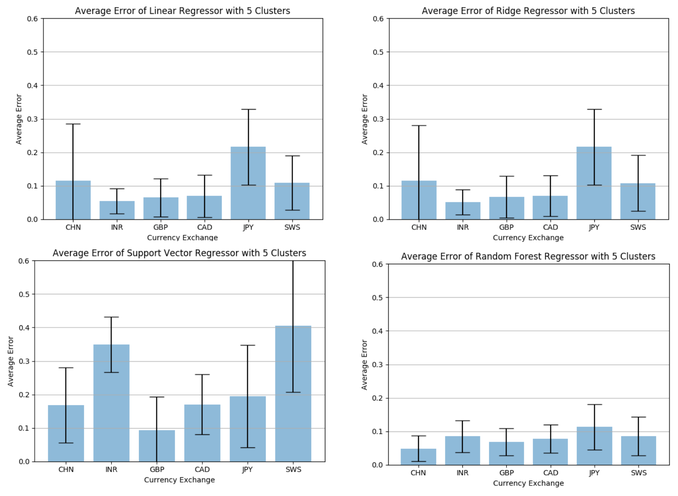


Figure 1: Average Loss for K=5 clusters with each Regressor

- Train: Yearly FERs from 1981-2005 | Test: 2005-2015
- Error: (predicted FER – actual FER) / actual FER
- When K = 5, the Random Forest Regressor had the lowest average test error for 5 of the currency exchange rates. SVR performed extremely poorly, while the Linear Regressions performed in between.

Feature Evaluation

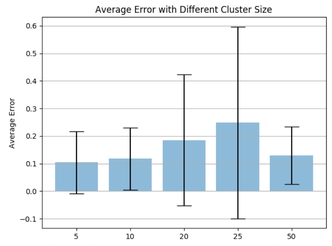


Figure 2: Comparison of Accuracy For Different Cluster Sizes

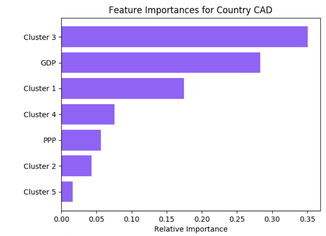


Figure 3: Relative Feature Important In Random Forest Regressor

- Bias and Variance lowest when K = 5.
- Most important feature for CAD was a cluster weight (War/Foreign Policy), indicating that news articles were important in the random forest.

Discussion

- Random Forest Regressor had the lowest bias and variance.
  - Bias: Likely was able to capture non-linear trends
  - Variance: Good tuning of min sample split and leaf size parameters
- K = 5 had superior performance, however when K > 5, clustering yielded highly specific topics.
- Likely performance drop when K > 5 due to overfitting from limited # of training years.
- Relative Feature Importance Metric provides evidence that there is a link between news clusters and FER, though more work is needed to confirm.

Future Work

- Investigate what other news datasets could be used to definitively confirm a correlation with FER: other newspapers, categories/sections, quantification of relevance.
- Refine LDA model to extract clusters with the most meaning and investigate economic rationale behind the weighting of certain clusters.
- Build on the model with more granular economic variables: trade balance, interest rates, etc.

References

New York Times API for Articles. https://developer.nytimes.com
World Bank Database for Exchange Rates: https://data.worldbank.org/indicator/pa.nus.fcrf