Samaksh Goyal, Hari Sowrirajan, Teja Veeramacheneni

Motivation and Overview

Problems
- 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.
- Use news article sentiment to predict FER.

Datasets and Feature Engineering

- 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 (CNY), 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

- Created a LDA model to extract the 5 most meaningful topics from the 40 years of news articles.
- Created a Random Forest model to predict FER for each country using the LDA model features as inputs.

Model Infrastructures

Latent Dirichlet Allocation (LDA)
- Hyperparameters: $\alpha$ (Dirichlet Prior for topic distribution), $\beta$ (Dirichlet Prior for word distribution), and $K$ (number of topics).
- We used: $\alpha = 0.2$, $\beta = 0.2$, $K = 5, 10, 20, 25, 50$.

Models for Training/Prediction

Linear Regression
- $\hat{y} = \beta_0 + \sum \hat{\beta}_j x_j$ (Linear Regression)
- $\hat{y} = \beta_0 + \sum \hat{\beta}_j x_j$ (Linear Regression)

Non-Linear SVR
- $\hat{y} = \beta_0 + \sum \hat{\beta}_j x_j$ (Non-Linear SVR)
- $\hat{y} = \beta_0 + \sum \hat{\beta}_j x_j$ (Non-Linear SVR)

Random Forest Regression
- Generated 100 Decision trees with following splitting criterion (Residual Sum of Squares):
- $\sum (y_i - y_{pred, tree})^2$ (Random Forest Regression)

Performance of Models

- 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

- 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