NSPP: News-Based-Stock-Price-Prediction

Mahdi Soleymani, Nasimeh Heydaribeni

Data Science Bootcamp, Erdos Institute

Summer 2024

Motivation and Overview

- Vast amounts of unstructured data (news) are continuously generated
- Harnessing this data can provide valuable insights that traditional financial models might miss

Overview of the project:

➔ Collect news data with stock related queries

➔ Preprocess the news data using a pre-trained LLM to extract their features

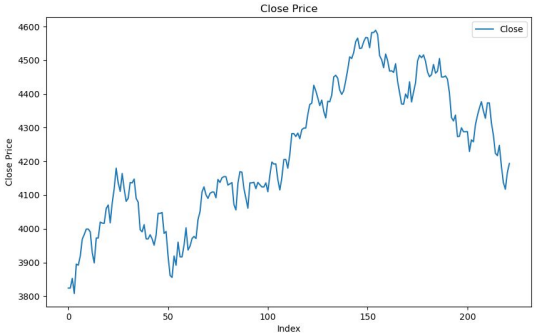
- ◆ Last hidden layer features
- ◆ Semantic information

➔ Use the extracted features in the conventional time series models as exogenous predictors

- ◆ ARIMA with exogenous variables
- ◆ Random walk with news-based drift

➔ Compare the performance with baseline time series models

Data Collection and Preparation



- We collected two types of data:
  a. S&P 500 stock price data for the year 2023
  b. NYtimes news with query S&P 500 over the year 2023

``The cost cutting continues at Twitter. The S&P 500 had a bad 2022, and it's not clear if 2023 will be better. The Labor Department will report jobs numbers for December.''

- We used Hugging Face API of the pre-trained LLM model DistilBERT
  a. **Feature Extraction**: Extracted the last hidden layer activations for all of the abstracts of the news data: a vector of length 768 for each abstract
  b. **Sentiment Analysis**: Classified the abstracts into positive and negative sentiments

  → The above example is classified as negative with score 0.998521
- And of course did some data cleaning to fill in the gaps
- Data corresponding to last two months are assigned to validation and test

Baselines and Proposed Method

- Naive forecasting (Naive)
- Random walk with drift (RW_D)
- Exponential smoothing (EXP_S)
- Rolling average (RA)
- ARIMA w/o exogenous variables (ARIMA)

Proposed Method:

- ARIMA with exogenous variables (ARIMA_EX)
- Random walk with news-based drift (RW_ND)

$$y_0 = \beta_0, \quad y_t = y_{t-1} + \beta_1 * d_{t-l} + \epsilon_t$$

Cross validation with test size = 1 (1 day forecasting) with number of splits being the length of validation set

Metrics:

- MASE (mean absolute scaled error)
  - How well the model is forecasting compared to a naive forecasting model averaged over the training set
- MAPE (mean absolute percentage error)
  - Mean of absolute error divided by the actual value

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=l}^{n}\frac{(y_i - y_{i-1})}{d_{i-l}}$$

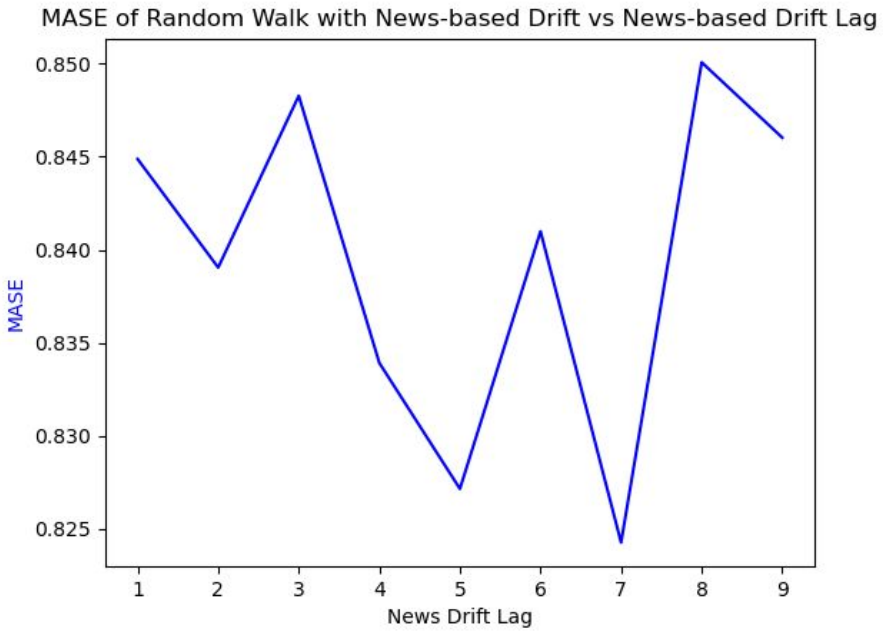
$$\hat{\beta}_0 = y_0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=l}^{n}(y_i - y_{i-1} - \hat{\beta}_1 * d_{i-l})^2$$

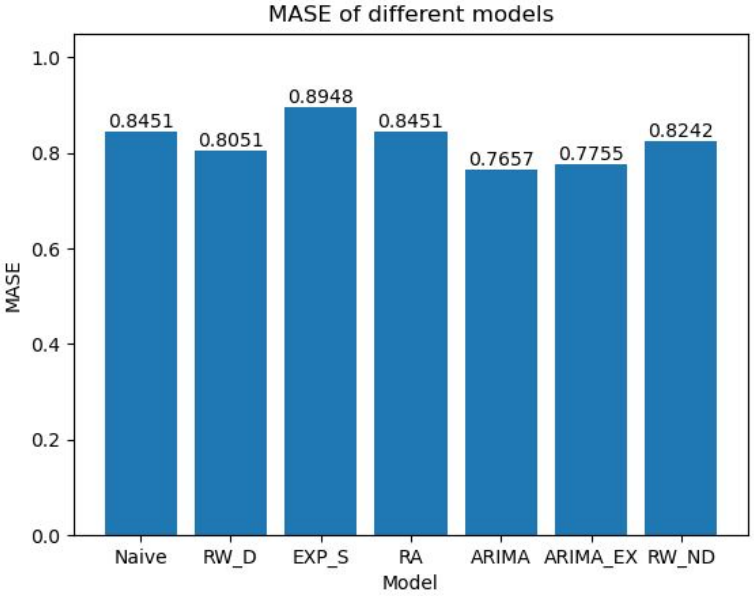
Model selection

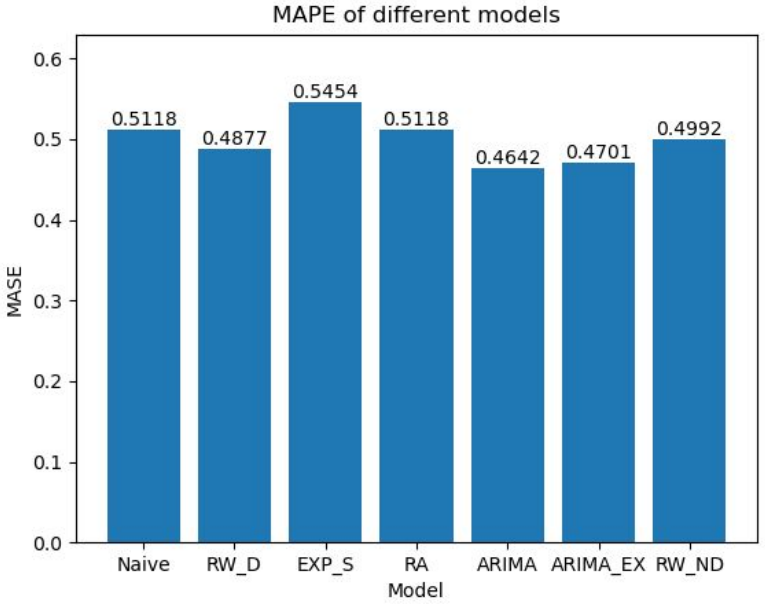
- Random walk with news-based drift
  - News might affect the stock prices with a delay
  - Explored different lags for the news-based drift in cross validation
  - Best lag was 7 with MASE of 0.8242 and MAPE of 0.4992
- We did a grid search for the rest of the models
  - Autoarima gave us worse results
- Cross validation was used for all models

- Arima parameters:
  - Min MASE parameters: (2, 2, 2)



Results





- ARIMA is the winner!
- The news data did not improve the predictions

Conclusion and future work

- The news sentiment has a small correlation with the stock price changes
  - Correlation coefficient: 0.0228
  - Correlation coefficient
  - with lag of 7: -0.1378
  - The correlation is even negative!

- More complex models such as LSTMs might be a better option for such tasks