News-Based Stock Price Prediction

Mahdi Soleymani, Nasimeh Heydaribeni
Data Science Bootcamp, Erdos Institute
Summer 2024

Introduction

Predicting stock prices is a challenging endeavor due to the dynamic nature of financial markets. Enhancing existing time series prediction techniques is both difficult and crucial, particularly in identifying the variables that have an influence on the market changes. One potential variable is news articles. This project aims to predict the S&P 500 stock price using time series analysis, augmented by sentiment analysis of news from the New York Times.

Data Collection and Preparation

We collected S&P 500 closing price data and the NY Times news articles containing the S&P 500 keyword for the year 2023. We used the NY Times API to obtain the articles. In order to extract features form the articles, we used the Hugging Face API, and employed a pre-trained LLM model (DistilBERT). After cleaning and transforming the data to make it compatible with Hugging Face pretrained models, we utilized the MinIO object storage system. This system enabled us to upload data to a virtual server running on a local machine and then download it for use in Hugging Face queries. For each article abstract, we obtained a feature vector of length 768 from the model's last hidden layer activations. Additionally, we performed sentiment analysis, classifying the abstracts into positive and negative sentiments. The data was cleaned, and the last two months were reserved for validation and testing.

Stakeholders

Hedge Funds and Private Equity Firms: These results could interest financial firms, such as hedge funds, that aim to predict stock price movements. Additionally, private equity firms could leverage this information to explore internet text for new investment opportunities, particularly in the rapidly growing startup environment.

Key performance indicator

We used MASE (Mean Absolute Scaled Error) and MAPE (Mean Absolute Percentage Error) metrics to measure the performance of all models.

Methodology

We compared several baseline models:

- Naive forecasting (Naive)
- Random walk with drift (RW-D)
- Exponential smoothing (EXP-S)
- Rolling average (RA)
- ARIMA without exogenous variables (ARIMA)

Our proposed methods included ARIMA with exogenous variables (ARIMA-EX) and Random Walk with news-based drift (RW-ND). We used cross-validation with a test size of one day and optimized models using MASE and MAPE. For ARIMA model selection, we performed a grid search to determine the optimal values, resulting in $p = 2, d = 2, q = 2$.

Table 1: Model Accuracy Comparison

Model	**MASE**	**MAPE**
Naive	0.8451	0.5118
Random Walk with Drift (RW-D)	0.8051	0.4877
Exponential Smoothing (EXP-S)	0.8948	0.5454
Rolling Average (RA)	0.8451	0.5118
ARIMA	0.7657	0.4642
ARIMA with Exogenous Variables (ARIMA-EX)	0.7755	0.4700
Random Walk with News-based Drift (RW-ND)	0.8242	0.4992

Results

The Random Walk with news-based drift (RW-ND) showed a MASE of 0.8242 and a MAPE of 0.4992 with a lag of 7 days for news impact, which was worse than the performance of the Random Walk with drift (RW-D). Furthermore, ARIMA model with exogenous variables (ARIMA-EX) did not yield better predictions compared to its non-exogenous counterpart (ARIMA). ARIMA showed the best performance overall.

Conclusion and Future Work

The sentiment analysis of news showed a minimal correlation with stock price changes, with correlation coefficients of 0.0228 (and -0.1378 for a 7-day lag). These results suggest that while news sentiment has some influence, it is not a strong predictor of stock prices. Future work could explore more complex models such as LSTMs to better capture the relationship between news sentiment and stock price movements.