NewsWorthy

Executive Summary - Data Science Project - Erdős Institute, Summer 2024

Research Question: Can we predict the stock price movements of a company using the sentiment scores of financial news headlines?

Team: Jem Aizen Guhit | Tim Alland | Nawaz Sultani | Ogonnaya Michael Romanus | Kenneth Anderson | Sarasi Jayasekara

GiTHub: https://github.com/jguhit/Erdos-DS-2024-newsworthy

Approach:

- Explore news sentiment data to find features that correlate with stock price movement.
- Develop and validate predictive models that use data from financial news headlines, alongside features from stock data, to predict the stock price movement of S&P 500 companies (focusing on 15 of them).
- Run a simulation of an investment portfolio directed by the best model, and evaluate how it performs. Compare the model's yield to the performance of a simple "Buy & Hold" strategy.

Data Gathering & Processing:

- Using Stock News API and Yahoo Finance, 5 years worth of news & stock data was collected, corresponding to the timespan: 2019/03/15 - 2024/03/15.
- Sentiment Analysis Scores on headlines were obtained using Vader & FinVader. Ultimately, the FinVader scores were used as features in the predictive models.

Initial Decisions:

Train-Test-Split: From the data set that spans 5 years,

- The last year of data (March, 2023 - March, 2024) was kept aside as the testing set.
- Among the first 4 years, the last year (March, 2022 - March 2023) was broken further into four 3 months increments as validation sets.

Models & Validation:

- Baseline Models: ARIMA, Buy and Hold
- Other Models: Logistic Regression, Gradient Boosted Trees, XGBoost, LSTM
- The portfolio simulation was run based on each model during each validation period. The average growth yielded by each model was compared to each other, and out of the models considered, Gradient Boosted Trees resulted in the highest average growth during the validation period.

Final Results & Conclusions:

- When the portfolio simulation was run based on the predictions of Gradient Boosted Tree, it resulted in a 2% average growth in the portfolio. This was higher than the 1% average growth yielded by the Buy & Hold strategy.
- This indicates that making Buy v Sell decisions based on sentiment data can yield greater accuracy than simply buying and holding.

Possible Further Developments:

- Future studies should examine the interplay between financial news and social media commentary on stock movement.
- Simulations and additional feature engineering that assess the recency and length of news cycles (7 days, monthly, quarterly, yearly, etc) on stock movement may improve model performance.
- Models that investigate bidirectional relationships between news and stock prices over time may inform future model deployment (i.e. news influencing stock prices and stock prices influencing news).