Predicting Stock Prices Using LSTM, Gradient Boosting and Combined Hybrid Models with Sentiment Analysis

Group Members: Archer O'Connell, Ethan D'Costa, Jay Mallard

Abstract

The goal of the paper is to analyze how integrating sentiment analysis can impact the accuracy of stock price prediction on different types of machine learning models. In the study we compare three different models, Long Short-Term Memory (LSTM), Gradient Boosting (XGBoost) and a hybrid model combining the two. Our approach is to train our models on historical stock prices alongside features extracted from the sentiment analysis of financial articles.

I. Introduction

The valuation of a company's stock changes every second of every day and is influenced by innumerous factors such as the overall trend of the market that day, the company's performance and the public's shared opinion. This is why predicting stock prices is a particularly challenging task involving incredibly complicated computations. The models used today by traders and Hedge Funds are often primarily solely based on market data but recent publications have demonstrated the inclusion of non-empirical data, such as sentiment analysis of online publications, can improve prediction accuracy of models.

To test this we use natural language processing to extract sentiment scores from financial articles about a particular company, which are then used in combination with the company's historical stock prices to train a predictive model. In order to gain a better understanding on exactly how the sentiment data is correlated to the stock price we train not just one but three different models: a LSTM model, a XGBoost model and a hybrid model which combines them.

II. Related Work

There have been numerous previous attempts to study different models' accuracy for predicting stock prices. One study titled "Twitter Mood Predicts the Stock Market" uses Twitter posts to analyze the collective mood of the website to test if the public's mood can predict changes in the Dow Jones Industrial Average which resulted in an accuracy score of up to 87.6% (Bollen et al. 2011). Two researchers from Stony Brook University studied the connection between stock market changes and sentiment scores of online news articles and blogs by buying stocks with positive sentiment scores and selling those with negative to see if those trades perform better than traditional strategies (Zhang et al. 2010). A study by Mohapatra et al. creates a hybrid model that creates an initial stock price prediction using LSTM followed by gradient boosting to correct residual errors in predictions increasing accuracy (Mohapatra et al. 2023). Researchers from the University of Washington also created a hybrid LSTM and gradient boosting model to create more efficient and cost effective scheduling of OR rooms in hospitals (Chen et al. 2018). Our project aims to build on these studies, specifically Mohapatra et al., by creating a hybrid LSTM and XGBoost model that relies on sentiment analysis data.

III. Problem Statement and Methods

Can we use *just* sentiment values from financial articles about a stock ticker to predict what its price will be in a given time range?

Methods

1. Data Collection
2. Data Formatting
3. Model training/prediction

Data Collection

Using the AlphaVantage API, given a ticker, retrieve

a) Its price history, dating back to its first listing
b) However many financial articles we can retrieve that mention the ticker, paginating through the API. Usually around 2-3 year time range, with results wildly varying between ticker entity size. (AAPL has many more articles than most penny stocks)

Data Formatting

Retrieve five features from each article:

1. The amount of tickers the article mentions
2. The relevance of our specific ticker
3. The sentiment score for our specific ticker
4. The sentiment score for the overall article
5. The average sentiment score for the publication

To account for timing mismatches (the stock market is closed on weekends, and not every ticker is mentioned in an article every day) we implemented rollovers, where the sentiment for a stock would roll over for every day there wasn't a new article, mutated by a temperature of 0.95. Stock prices for a weekend are Friday prices.

Model Training/Prediction

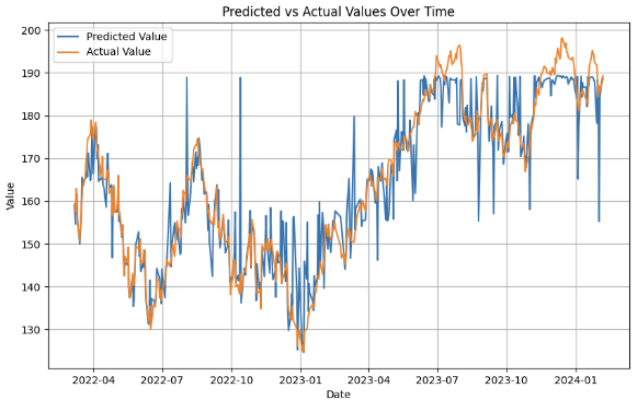
After loading in our sentiment data and price data, we split it into training sets and testing sets. Since the temporal nature is key to our data/problem, we settled on an 80/20 ratio, where the first 80% of days in a row were used as training data, then the final 20% sequentially were used as testing data.

We built three models:

1. an LSTM (Long Short-Term Memory) model, using a Mean Squared Error loss function
2. a Gradient Boosting model to create a strong predictive
3. a Hybrid model, which first trains the temporal LSTM model and then feeds its residuals to a Gradient Boosting model.

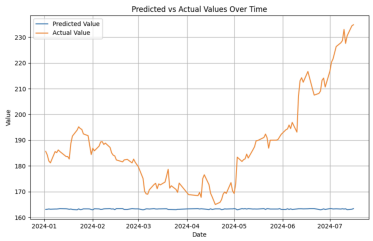
IV. Experiments and Results

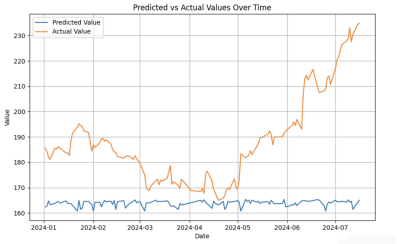
The development process was initially highly promising -- after our first few roadblocks in getting the LSTM model up and running, our base hyperparameters of 10000 epochs, a batch size of 3, and learning rate of 0.001, saw our training loss shoot from the tens of thousands to level off in the low four hundreds range. After tuning the batch size and learning rate, we found that our model performed excellently on our training data (seen on the left) when given the hyperparameters

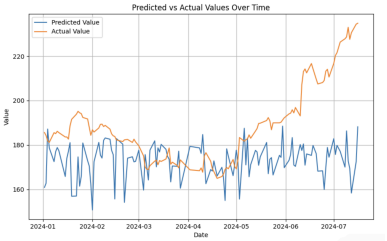


+ $n_{(epoch\ size)} = 10000$
+ $a_{(learning\ rate)} = 0.001$
+ $s_{(batch\ size)} = 20$

Unfortunately, this stellar correlation didn't continue over to our test data. The LSTM model, Gradient Boosting Model, and combination model all obtained similarly disappointing results. There were some moments of false hope -- for example, while working on the Gradient Boosting model, as we increased *a*, it seemed like we were making progress on our testing data (shown below), but upon closer inspection, it was clear that there was no correlation, the amplitude of our incorrect results was just increasing.

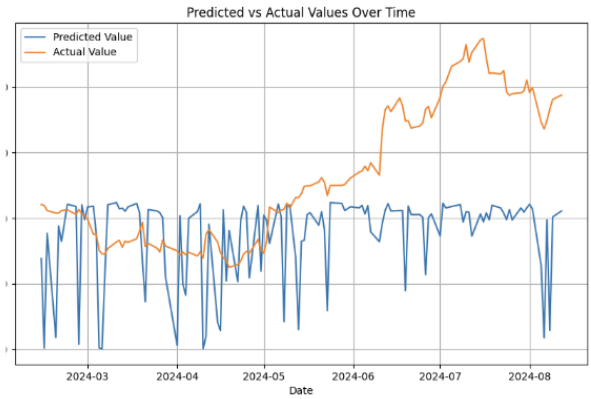






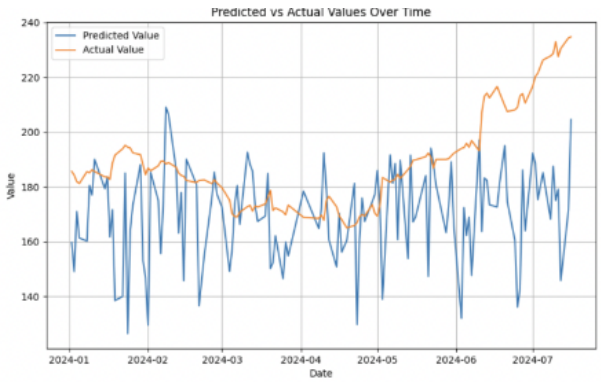
Frequent constant tuning of many different permutations of values of *n, a,* and *s* (domains of $n \in [10000, 100000]$ $a \in [0.0001, 1]$, $s \in [3, 100]$) didn't have much effect, leading us to believe that there is not an effective correlation between *just* financial news sentiment data and stock market prices. As you can see in the figures on the next page, despite the nature of Gradient Boosting models meaning they don't account for the temporal nature of data, meaning it's very unlikely for it to be, alone, a good stock price indicator, the gradient boosting standalone model does just about as well as the other two.

LSTM



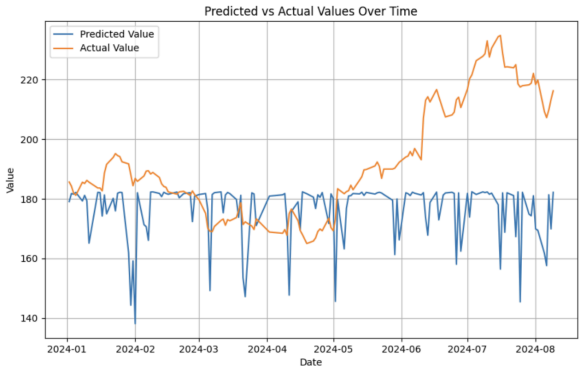
n = 100000
s = 20
a = 0.0001

XGBoost



n = 10000
a = 1

Hybrid



As referenced in the related work section, many models are built using the same hybrid setup as ours, but with stock price also fed into the LSTM as input, with sentiment value taking more of a backseat role. However, our aim for this project was to see if it's possible with *just* sentiment.

n = 10000
s = 20
a = 0.0001

V. Conclusion

Unfortunately, demonstrated by the poor results from all three models run on different metrics, our conclusion is that there is very little correlation between the sentiment scores of discussions about a company and its stock price. Due to lack of time and resources one of the limitations of our models is a limit in diversity and size of data which can be altering the results. Further, since our models heavily rely on the sentiment data, inaccurate or biased sentiment scores would have a severe negative impact on predictions. One way to improve the models would be to use different sources for sentiment scores such as social media posts in order to improve the quality of sentiment data. Another improvement could be to include more features and types of market data such as history of market volatility and including the opening as well as closing stock price.

Team Contributions Statements

Ethan built the LSTM model and UI, Jay worked on the XGBoost model and the body of the report, and Archie worked on collecting/formatting the training data and building the XGBoost model and hybrid model, along with designing the presentation.

GitHub Repository Link: https://github.com/HayBirdJay/StockPriceAI

References

Bollen, Johan, et al. "Twitter Mood Predicts the Stock Market." Journal of Computational Science, vol. 2, no. 1, Mar. 2011, pp. 1–8, https://doi.org/10.1016/j.jocs.2010.12.007.

Chen, Lundberg, and Su-In Lee. "Hybrid Gradient Boosting Trees and Neural Networks for Forecasting Operating Room Data." (2018). https://arxiv.org/pdf/1801.07384.

Mohapatra, Pratyush Ranjan, et al. "Gradient Boosting and LSTM Based Hybrid Ensemble Learning for Two Step Prediction of Stock Market." Journal of Advances in Information Technology, vol. 14, no. 6, 22 Nov. 2023, pp. 1254–1260, https://doi.org/10.12720/jait.14.6.1254-1260.

Zhang, Wenbin, and Steven Skiena. "Trading Strategies to Exploit Blog and News Sentiment." Proceedings of the International AAAI Conference on Web and Social Media, vol. 4, no. 1, 16 May 2010, pp. 375–378, https://doi.org/10.1609/icwsm.v4i1.14075.