Interim Report
CS640
MScSE Project
2024 – 2025

Sarbojit Bhattacharjee
23252356

Wu Hao

Stock Market Forecasting Using Sentiment Analysis and Trading Volume Trends

Goals of this project

The main goal of this study is to create a model that uses historical financial data to predict stock prices. When we look at the past prices of stocks, we want to find patterns and trends that will help us guess where the prices will go in the future. It has starting, ending, high, and low prices, trade volume, and daily percentage changes for Apple Inc.'s shares, as well as other past market data. Using this well-organized data, a regression-based model will be made that can fairly guess what stock prices will be in the future.

Implications for the real world are what make the project important. Accurate predictions of stock prices are an important part of financial study that helps buyers, traders, and companies all at the same time. Predicting how a stock's price will move in the future can help buyers make better choices, like whether to buy, sell, or hold on to a stock. For a long time, fundamental and basic study have been the basis for predicting stock prices. However, advances in machine learning have made it possible to create more complex methods. This research looks at these tactics by using regression methods, especially linear regression, to guess what stock prices will be based on how they have changed in the past. This project focuses on predicting stock prices using historical data and regression-based models, emphasizing interpretability over complex deep learning techniques. Unlike prior studies relying on black-box algorithms, this approach prioritizes transparency in feature selection and forecasting. The dataset, sourced from Yahoo Finance, includes Apple Inc.'s historical stock prices.

An important goal of ours is to see how well our prediction model works by comparing it to real changes in stock prices. Several measurements, such as root mean squared error, mean squared error, and mean absolute error, are used to check how reliable the model is. This is because the accuracy of financial predictions is very important. Visualizations of residual distributions, real vs. expected stock prices, and future forecasts may also shed light on how well the model works and where it could be improved.
Key third-party libraries used include pandas, numpy, scikit-learn, matplotlib, and seaborn for data processing, modeling, and visualization. While stock price prediction is widely explored, this project offers enhanced visual insights and structured analysis, distinguishing it from existing models while acknowledging market unpredictability and external influencing factors.
In addition to trying to guess stock prices, this study also wants to look into what machine learning means for the financial markets as a whole. By looking at the pros and cons of prediction models, we can learn a lot about how machine learning can be used for stock trading methods, portfolio management, and figuring out how much risk there is. There is a strong base for this project that emotion analysis, deep learning models, and reinforcement learning for trade methods all build on.

Problems encountered and progress so far

There are a lot of problems that need to be solved when making a model to predict stock prices. These include, but aren't limited to, evaluating performance, choosing a model, and getting data ready. The layout and style of the information were one of the first problems that had to be solved. Because there were so many of them, it was important to think about which traits in the original stock data would help the most with making estimates. It was hard to choose factors that correctly

showed trends in stock prices without adding too much or too little noise. Volume, high, low, and starting price were the factors that were chosen in the end. Test and train data sets are obtained after the classification and analysis from the machine learning approach.

- Problems were also found with the preparation of the data and the handling of missing numbers. There aren't many missing numbers in this project's dataset. In real-world stock market systems, however, gaps in data are common due to breaks, market shutdown, or broken records.
- Managing missing factors well is important for making sure the model is resilient. Data normalization was also thought about in addition to making sure that the model was free of bias caused by differences in feature sizes.

- It was very hard to separate the information into a training set and a test set. In many machine learning situations, data can be split up without any particular reason, but in the stock market, data must be kept in chronological order to stop data leaks.

- Using old train-test division methods that mix up the data would lead to inaccurate performance measures and false situations when future data points are used for training.

- It took some skill to make sure that the model was taught on data from the past and then tested on data from the future.

After looking at several different regression methods, the model selection process finally decided that linear regression was the best and easiest way to do regression. Even though linear regression was a simple method, it was easy to see that it wasn't completely satisfactory when working with non-linear price changes. A lot of different factors, like economic data, trader emotions, and global politics, can affect the price of stocks. Using a simple linear model to account for these factors might be hard, if not impossible.

In its early studies, the linear regression model showed promise in predicting stock prices. However, it completely failed when prices changed quickly and became very volatile. To get a better picture of the differences between what was expected and what actually happened, we found the root mean squared error and the mean absolute error. From the residual analysis, it was clear that the model worked well when prices were stable, but it had trouble catching sudden price changes.

Visualization was a key part of judging how well the plan worked. When looking at the first set of graphs, there were times when the real and expected stock prices were pretty close to each other, and times when they were very different. The residual figure showed that the mistakes were not spread out evenly, so the model had to be changed. We added more graphics, like scatter plots and residual histograms, that showed the difference between real and expected values so that we could get a better idea of the model's limitations.

Next steps

Several critical actions must be undertaken to improve the model and provide more accurate predictions about stock prices. The assessment of increasingly more advanced device gaining knowledge of methodologies is a number one vicinity requiring development. despite the fact that linear regression gives an extraordinary start line, it can be fine to discover other methods, including choice timber, random forests, and guide vector machines, to enhance the accuracy of destiny predictions. Deep mastering models, specially LSTMs and RNNs, may additionally facilitate the analysis of temporal correlations amongst fluctuations in inventory costs.

Incorporating improved traits into the records is a following indispensable step. The cutting-edge new release estimates stock prices via the usage of historic records. however, it is possibly elevated if it blanketed outside factors along with market sentiment, macroeconomic signs, and international financial hobby. A valuable approach to apprehend dealer behavior and its effect on inventory expenses is to investigate public sentiment on financial records and social media trends.

it is also essential to decorate the version's software by way of refining the methodologies utilized in statistical practise. it's far very conceivable to beautify the education dataset by using feature engineering techniques, integrating random facts, and more efficiently managing lacking values. To realize more complex market conduct, one may also use volatility metrics, transferring averages, and time series analysis(Zaman,2023).

Incorporating extra elements into the version evaluation method may also provide a greater complete assessment of fulfillment. One may additionally get extra perception into the efficacy of the model by using comparing it the usage of different metrics, consisting of as R-squared ratings, absolute percentage blunders, root imply squared mistakes, and imply absolute mistakes. To enhance the accuracy of predictions, we would use improved window validation and further pass-validation strategies mainly tailor-made for time collection forecasting.

Allocating more time for forecasting is another crucial aspect of future planning. The model now has the capacity to anticipate just a few days ahead; its accuracy would significantly improve if it could extend its predictive range while maintaining correctness. It may be beneficial to examine integrated solutions that combine machine learning with general economic models to achieve an optimal balance between issue comprehension and data-driven insights.

The last phase is to contemplate how to implement the notion in the actual world. Developing a platform that provides real-time inventory price projections will likely enhance the utility for investors and purchasers. To demonstrate its use, the version might be integrated into a financial analysis tool or offered as a web application.

General comments

1. In financial forecasts, the steps of collecting data, preparing it, modeling it, and evaluating it have shown what prediction models can and cannot do. Stock prices change because of many complicated and often unexpected factors. This means that more advanced modeling methods are needed than linear regression, which is a simple and easy-to-understand method.

2. Due to the natural instability of financial markets, there is no single model that can correctly predict stock prices. One of the most important lessons is this. But statistical learning techniques and past data can be used to build models that can make correct guesses and help people make decisions. The information from this study could be used to look into more complex methods, like deep learning, reinforcement learning for trading tactics, and multi-factor models that use many market factors(Mehta,2021).
3. When it's done, this project will serve as a spark for more study into financial forecasts. To make stock price prediction tools that are more accurate and useful, model improvements, the addition of new features, and testing with different methods can all be used. A lot of interesting new themes are coming up where finance and machine learning meet, like algorithmic trade, risk management, portfolio optimization, and predicting stock prices.

Timeline:

Task	Nov-24	Dec-24	Jan-25	Feb-25	Mar-25	Apr-25	May-25	Jun-25
Project Planning & Proposal								
Data Collection & Preprocessing								
Exploratory Data Analysis (EDA)								
Feature Engineering & Model Selection								
Model Training & Evaluation								
Prediction & Visualization								
Report Writing & Documentation								
Final Review & Submission								

References

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in International Business and Finance*, *64*, 101881.

Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, 115019.

Ko, C. R., & Chang, H. T. (2021). LSTM-based sentiment analysis for stock price forecast. *PeerJ Computer Science*, *7*, e408.

Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, *7*, e476.

Nemes, L., & Kiss, A. (2021). Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, *5*(3), 375-394.

Wu, S., Liu, Y., Zou, Z., & Weng, T. H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*, *34*(1), 44-62.

Zaman, N., Ghazanfar, M. A., Anwar, M., Lee, S. W., Qazi, N., Karimi, A., & Javed, A. (2023). Stock market prediction based on machine learning and social sentiment analysis. *Authorea Preprints*.