STOCK MARKET PREDICTOR USING RECURRENT NEUREL NETWORK MACHINE LEARNING MODEL

Omar Khalil, William Knoff, Hesham Elshafey, Ian Curd

University of Rochester Electrical and Computer Engineering Department

ABSTRACT

The stock market is one of the most complex and unpredictable systems in the world. With numerous variables of trends, news, supply and demand, it's incredibly difficult to predict. Using machine learning it may be possible for these trends to be found in a model and used to make your choices. We explore this possibility and put it to the test.

1. INTRODUCTION

We used a dataset that contains information about the S&P 500 stocks from 2020-2022. This dataset set has provided us with both market data (volume, price, etc.) and news data (sentiment, events, etc.). Exploratory data analysis was performed to avoid any inaccuracies that might throw our training model off, such as the concept of stock splits. Following that, we experimented with our recurrent neural network model, such as the layers, to enhance performance. We were then left with conclusions that were drawn based on the results.

2. METHODS

In our project we utilized a Recurrent Neural Network (RNN). This is the optimal way to process ongoing data such as the stock market. A RNN takes trends of previous gradients in the graph in order to predict future gradients. The specific RNN model that we used is a Long Short-Term Memory (LSTM) model. LSTMs fix the exploding and vanishing gradient issue with traditional RNNs. Since our data is a point every day for multiple years this is a large issue for our dataset. LSTMs also deal with gaps in the dataset which is another issue solved for us since the stock market isn't open on weekends. This means that every week we have a gap of two which the LSTM solves.

We have two models for our training that we used, the first of which is using only the target stock in order to train its future performance. The first method we tried was training using four years of data from our old dataset and then predicting the entirety of the fourth year. This caused a compounding error problem which caused large inaccuracies. Our method for determining the success was calculating the MSE and RMSE which were very inaccurate. We then took the closing data from the past 60 days as training in order to predict the next day. We used rolling predictions in order to keep relevancy in the data. Our method for determining success with this approach was to take daily predictions of the stock increasing or decreasing. We then took the percentage of correct and incorrect stocks as our method of success.

The second model for our training was using every stock in the S&P 500 in order to train the target stock's future performance. In this method we had to cut the dataset to make sure all of the stocks were the same size. Since the S&P 500 is an index of the top 500 stocks it will change over time so some of the stocks are shorter than the rest. Once the data preprocessing was done we only did the first method done for the single stock. We chose to not do the second method since there was an overtraining issue and it would've been obsolete. Our method for determining the success was again calculating the MSE and RMSE.

Some of the visualization methods we used included bar charts, heatmaps, and graphs. Bar charts were used for evaluating success, heatmaps were used for finding correlations, and graphs for enhancing performance.

3. EXPERIMENTS

We experimented with different factors, such as the stocks being trained on, the inclusion of sentiment features, and the length of training recall. Our initial model used solely market data and neglected news sentiment; it led to inconsistent RMSE values that ranged from 5 to 30 RMSE when being trained and tested on the same stock. In order to develop a model that encompassed the overall movements of the market, we averaged all stocks in the dataset together to create an unweighted representation of the S&P 500 index, then tested it on individual stocks. However, this led the model to become overly stabilized, predicting near zero change in stock prices with an RMSE of 40.

We switched to a new dataset that incorporated news sentiment data in order to make the model more robust. Unfortunately, none of these features held any significant correlation to stock price percentage change, even when staggered by a week to account for the time it takes for sentiment to make an effect. Using the new dataset, we experimented on using different recall lengths to train the RNN. When predicting the

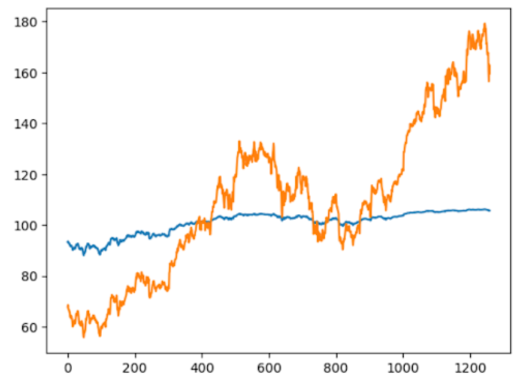

Fig. 1. Model trained on unweighted S&P 500 index predicting price of AAPL

next day's price, the RNN only trains on a certain number of days prior; we tested the range of 20 - 60. At 60 days, the model achieved a reasonably low RMSE of 6; however, it consistently predicted % changes in price between -1 to 1%, which is much lower in volatility than reality. When train-

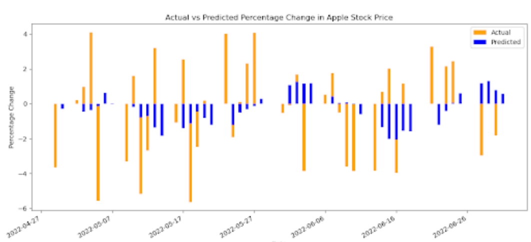

Fig. 2. Model trained on unweighted S&P 500 index predicting price of AAPL

ing on AAPL, a recall horizon of 30 was optimal to match the volatility of the market while maintaining ¡10 RMSE with more correct than incorrect predictions. However, this recall horizon did not apply well when the model was applied to other stocks.

Utilizing a mock portfolio starting at $50 stock, $50 cash that invests $1 when predicting up and sells $1 when predicting down, or model lost money every single time. When switching to a strategy with purchases weighted by magnitude, the model still lost.

4. CONCLUSION

After running multiple experiments using multiple models along with researching the predictability of the stock market, the team concluded that the market is inherently random and unpredictable. We discovered, along with the other obvious reasons, random reasons that caused fluctuations in the stock market prices that are small as appearance of physically attractive CEOs on TV and wrong company names due to investor errors. Our model was able to learn from the previous data relatively well, but the natural randomness of the stock market mandates a methods to track news in real-time and pass them through the model as fast as possible to correct predictions. As a result, our model can serve as a tool for intuition to assist decision making during stock trading, not as a magical way to guarantee profit.

5. INDIVIDUAL MEMBER CONTRIBUTIONS

Hesham: Model Layers , Evaluation, Exploring Predictability of Stock Market
Ian: Data Preprocessing, Created All Stock Input Model
Omar: Visualization, enhancing performance, report
Will: Exploratory Data Analysis and Preprocessing, Mock Portfolio for

6. REFERENCES

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title," Journal, Publisher, Location, pp. 1-10, Date.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, Book Title, Publisher, Location, Date.