Stock Price Movement Prediction For Gamestop

Sean Yeo, Yufei Chen, Sean Sui
SENG 474 Summer 2021, University of Victoria

Summary

We apply machine learning techniques to predict price movements for Gamestop Corporation approximately year-to-date using daily historical data from Yahoo Finance, and sentiment scores of text data from Reddit. Using decision trees and LSTMs, we learn that LSTMs perform better in predicting price movements than decision trees. Additionally, it is found that sentiment data from Reddit does not yield better prediction results for our problem.

Dataset & Features

We downloaded data from 1-01-2021 to 7-25-2021 from Yahoo Finance. Linear interpolation is used on open, high, low, and close values to increase the size of the data set from 140 to 10000 rows.

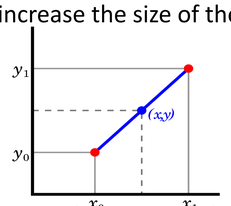

Figure 1: Linear interpolation example

Using these interpolated values, 4 technical indicators are calculated: RSI (Relative Strength Index), ADX (Average Directional Movement Index), SMA (Small Moving Average), LMA (Long Moving Average). Technical indicators are values that can be calculated from open, high, low, close, and volume data to help predict trends in stock prices [1].

Figure 3: RSI and ADX used in combination

Figure 4: Moving Averages Use Case

The Pushshift API is used to scrape posts and comments from specific Subreddits from January 1,2021 to July 25, 2021. We performed sentiment analysis on the reddit data and classified all posts as positive, negative, and neutral. We appended the sentiment as a new column to our dataset.

```
 datetime.datetime(2021, 1, 8, 0, 0): ['Rampant illegal s
p the abuse \n\nFails to deliver $GME shares for second h
  'This is way too reasonable of a thread for r/wallstree
  '$GME price should eventually go up bigly 🚀🚀🚀',
```

Figure 5: Reddit comment snippet

Decision Tree

We set up a decision tree to predict price movements by tuning the maximum depth, minimum sample splits, and minimum sample leaves. We do six time series splits and calculate the AUC (area under curve) score and accuracy for with and without sentiment data.

The model is trained and tested with and without sentiment data. Afterwards, the results are compared

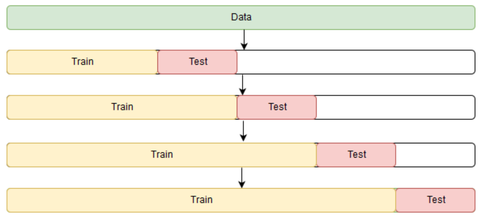

Figure 6: time-series splits

LSTM

LSTM (Long Short-term Memory) neural networks are commonly used to process sequences of data such as speech, video, and handwriting. For this reason, we thought they would be well suited to work with time series data from the stock market.

The first 70% of the enlarged data set is used for training, and the last 30% is used for testing. For testing accuracy, we decided to switch between different activation functions in LSTM layers to see if one would perform better than others. We decided to choose ReLU, sigmoid, and softmax to test this.

The network consists of two LSTM layers, followed by a dropout layer, and a single output node for prediction output [2]. For input, TimeSeriesGenerator objects are created to input sequences of data into the LSTM.

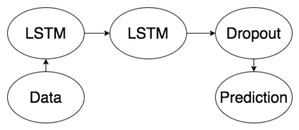

Figure 7: Simplified LSTM workflow

```
<keras.preprocessing.sequence.TimeseriesGenerator
[[1 2]] => [3]
[[2 3]] => [4]
[[3 4]] => [5]
[[4 5]] => [6]
[[5 6]] => [7]
[[6 7]] => [8]
[[7 8]] => [9]
[[8 9]] => [10]
```

Figure 8: TimeSeriesGenerator example

The model is trained with and without sentiment data. Afterwards, the results are compared. Additionally, as another point of comparison, the number of time steps used in the input sequence is varied, and accuracy is compared.

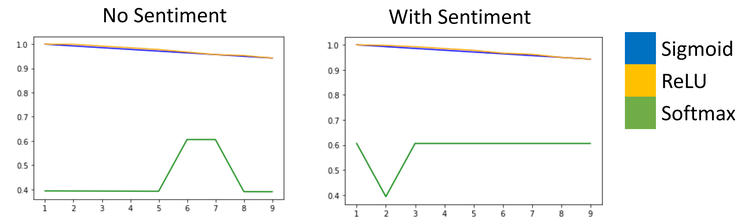

Figure 9: LSTM accuracy as the number of time steps increases

Results

Decision Tree model accuracies:

No Sentiment	AUC	Accuracy
Training	0.67	68.60%
Testing	0.50	49.52%

Sentiment	AUC	Accuracy
Training	0.67	68.60%
Testing	0.50	49.45%

LSTM neural network accuracies:

# timesteps	1	2	3	4	5	6	7
No Sentiment Testing	99.9%	99.2%	98.5%	97.8%	97.1%	96.4%	95.7%
Sentiment Testing	99.9%	99.2%	98.5%	97.8%	97.1%	96.4%	95.7%

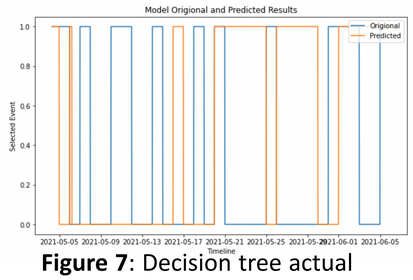

Figure 7: Decision tree actual vs predicted results

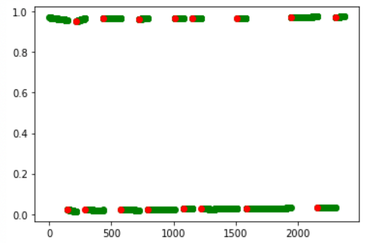

Figure 8: accuracy plot for LSTM

Conclusion

- LSTMs provide more accurate results for price movement prediction on this particular data set.
- The sentiment data makes no measurable difference in the prediction accuracy of the chosen methods.

Future Work

- Use XGBoost versus LSTMs, and using RMSE (root mean-square error) for actual price values as opposed to binary up/down labels.
- Adding Noise to Linearly Interpolated Data and see how both models respond to the addition of noise.
- Gather data from all investing subreddits to achieve better sentiment score that reflect the opinions of a more diversified group of people.

References

[1] J. Chen, “Technical indicator definition,” *Investopedia*, 05-Aug-2021. [Online]. Available: https://www.investopedia.com/terms/t/technicalindicator.asp. [Accessed: 12-Aug-2021].

[2] S. Bhattiprolu, “multivariate time series forecasting using LSTM,” *YouTube*, 08-Dec-2020. [Online]. Available: https://www.youtube.com/watch?v=tepxdcepTbY. [Accessed: 12-Aug-2021].