Himanshi Gupta, Yoomin Oh, Suditi Bhatt

Professor Bergen

LIGN 167 (Fall 2022)

08 December 2022

LIGN 167: Final Project Write-Up

Link to Code in Our Github Repository:

https://github.com/Yoomin99/LIGN167_Stock.git

Introduction

The stock market is notorious for being volatile and dynamic. There are many factors that affect it ranging from politics and global warming to the financial performance of a company. This makes it tough to predict stock prices accurately. However, this also means that there is a lot of data to analyze and find different patterns. It is useful to have this knowledge about upcoming trends or projected stock prices in order to make investment decisions. In recent years, buzzwords like 'investment portfolios' and 'stock trading' have been thrown around more casually. Companies like Robinhood have further lowered the bar to enter this space for the general public. Thus, nowadays information about stock prices and predictions is relevant to a larger number of people. Online articles about the factors that affect the stock market can help these people make better investment decisions. However, this knowledge is vast and it is hard to fully comprehend the impact of each little factor. This is where machine learning algorithms can help predict stock prices and find general trends in order to aid investment decisions. Our project aims to compare the following machine learning models- linear regression model, logistic regression model, LSTM model, and GRU model, and figure out which model gives the most accurate prediction.

We used datasets from Kaggle - an online data science community platform, for our project. For the logistic regression model, we used a dataset that had keywords from different articles and the stock market fluctuations. For the other three models, we chose datasets containing stock prices for Netflix, Amazon and Google, three well-established and popular companies, from the year 2010 onwards. We trained all our models using the data from 2010 to 2017 and then compared our graphed predictions for the next few years with the actual documented fluctuations in stock prices.

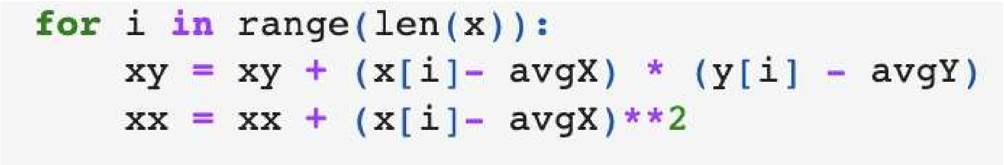
Description of Models

The four models we created and trained to attempt to predict specific future stock prices are the linear regression model, logistic regression model, LSTM model, and GRU model.

The datasets we used contained '*x*' values that represented the time variable (dates) and '*y*' values that represented the stock prices.

Linear Regression Model [1]

First, we calculate the weighted average and sum of all *x* values and then the *y* values. Then we use the formulae shown from our code below to calculate the slope and y-intercept :



```
for i in range(len(x)):
    xy = xy + (x[i]- avgX) * (y[i] - avgY)
    xx = xx + (x[i]- avgX)**2
```

After the above calculations, we trained the data using a subset of the stock prices from the year 2010 to the year 2017. In order to evaluate the model, we compared the predicted stock prices obtained after training with the actual stock prices.

Logistic Regression Model [2]

For our logistic regression model, we implemented a logistic regression line to identify if the positive or negative articles can be used to predict increasing or decreasing trends in stock prices. With our x variable representing the change between the open and adjusted stock prices and our y variable representing the connotation of the articles in the dataset, we trained our model with the LogisticRegression( ) function in order to analyze how the changes in stock prices are associated with article headlines about these companies and their stock values.

LSTM Model[3]

In our implementation of the LSTM model based on the source cited above, we were able to observe how changing the number of dimensions that are used in the model can affect its prediction accuracy. We then trained our model based on the declared number of epochs in order to predict the stock prices of the remaining dates in the dataset for each of the three companies.

GRU Model[4]

In our application of this GRU model based on the source cited above, the structure of this model is similar to the LSTM model in which the model is created based on the value of its set dimensions and the number of layers within it. The trained data is a result of iterations that are equal to the number of epochs and the final prediction of the y variable, which represents the stock price, is a result of the inverse transform of the y predictions based on the trained data from the model.

Description of Datasets

Dataset of Amazon Stock Prices [5]

The dataset we used for Amazon contains data about the Amazon stock prices during the dates from 2010 to 2020. We used all the stock price data from the dates in 2010 to 2017 as our subset for the training data, which is 2013 training points. We used all the stock price data from the dates in 2018 to 2021 as our subset for the testing data, which is 963 testing points.

Dataset of Google Stock Prices: [6]

The dataset we used for Google contains data about the Google stock prices during the dates from 2010 to 2022. We used all the stock price data from the dates in 2010 to 2017 as our subset for the training data, which is 2013 training points. We used all the stock price data from the dates in 2018 to 2022 as our subset for the testing data, which 1065 testing points.

Dataset of Netflix Stock Prices: [7]

The dataset we used for Netflix contains data about the Netflix stock prices during the dates from 2010 to 2021. We used all the stock price data from the dates in 2010 to 2017 as our subset for the training data, which is 2013 training points. We used all the stock price data from the dates in 2018 to 2021 as our subset for the testing data, which is 951 testing points.

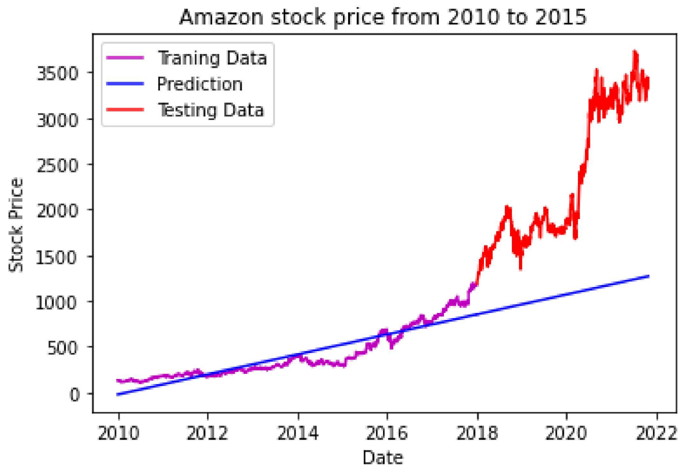
Dataset for Logistic Regression Based on Articles and Stock Price Data: [8][9]

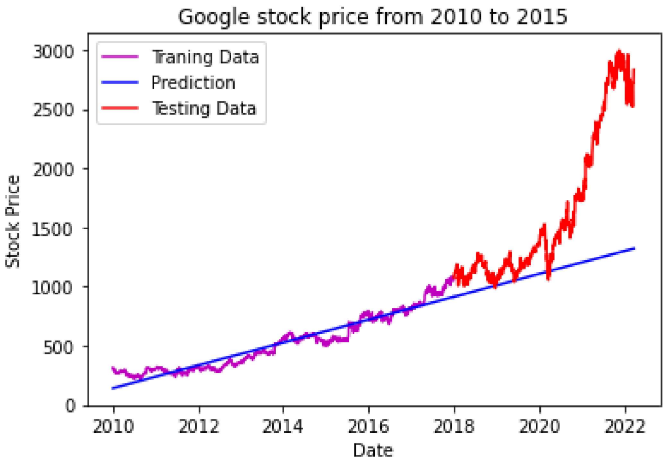
We used this data set that contains data about news data on certain dates along with the corresponding stock prices on those dates. We used the subset of this dataset for the dates from 2010 to 2016.

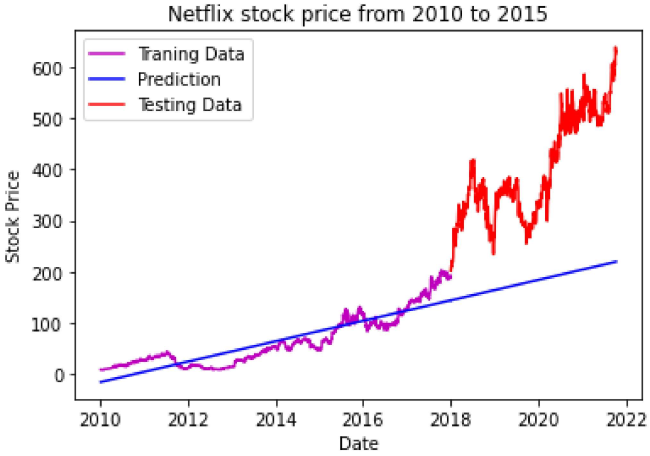
Results and Analysis

We obtained our results by training each model based on a subset of our dataset for each of three companies, which are Amazon, Google, and Netflix. For training our linear regression, LSTM, and GRU models, we used the subset of the stock price data from 2010 to 2017 for Amazon, Google, and Netflix. After each of these three models were trained, we tested our models by predicting the stock prices for the remaining dates that were not used for training in the dataset for each company and then comparing the results to what the actual stock prices that were in the dataset ended up being on those dates. We then graphed these test results for each of three companies and included three plotted lines that represent the training data, testing data, and the prediction data with the x-axis as the dates and the y-axis as the stock price of the respective company on each data for the testing subset. We have attached the graphs of the results from these three models below in this write-up. While the linear regression model was not very accurate in predicting the stock prices on the subset of dates used for testing, the LSTM model was more accurate and the GRU model was the most accurate as the plotted line for the predicted stock prices was very similar to the plotted line for the actual prices on the future dates. These graphs helped us easily visualize the prediction results of each model and analyze how close the predicted stock prices were in comparison to the actual future stock prices that were recorded on those dates. Based on these results and graphs, we have concluded that the GRU model is the most accurate in predicting the future stock prices of Amazon, Google, and Netflix. For the logistic regression model that we implemented for this topic of stock price prediction, we applied this model in a different method than the other three. We decided to see how a logistic regression model can be used to predict more general increasing or decreasing trends in stock prices for these three companies based on online articles in order to incorporate a more linguistic approach to solving our problem. By using a dataset about articles related to companies and their prices in the stock market, we trained a logistic regression model to utilize the wording in these articles to predict if future stock prices will increase or decrease. Based on the results from our model, the correctness of linking the positive or negative connotations of the article headlines with the changes in stock prices is about 0.465 for Amazon, 0.4524 for Google, and 0.4921 for Netflix. These results indicate that the logistic regression model may be able to predict some kind of general trend for these companies' stock prices based on the article headlines, but it may not always be the most accurate since the results of correctness for the correlation is less than 0.5000. After testing and analyzing the results of all four models, the GRU model seems to be the most accurate in predicting the future stock prices of these three companies since its prediction results were the most similar to the actual prices recorded as the points of comparison for our testing data.

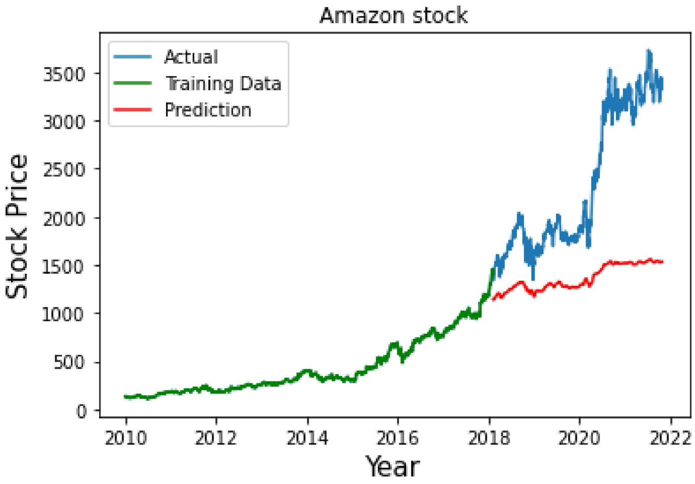
Graphs for Our Linear Regression Model

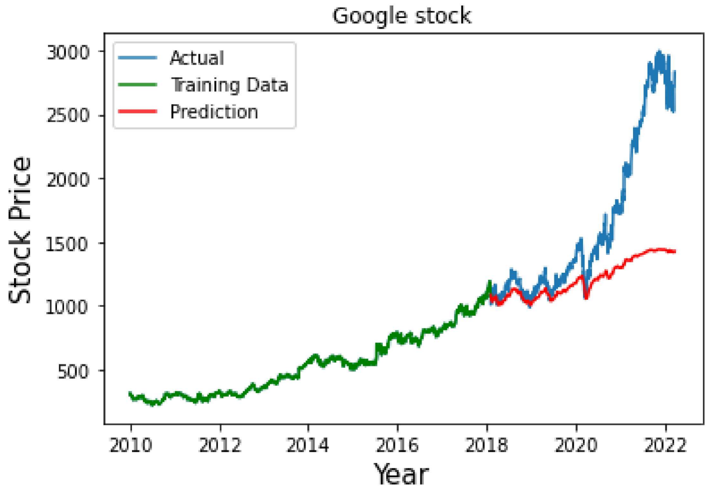


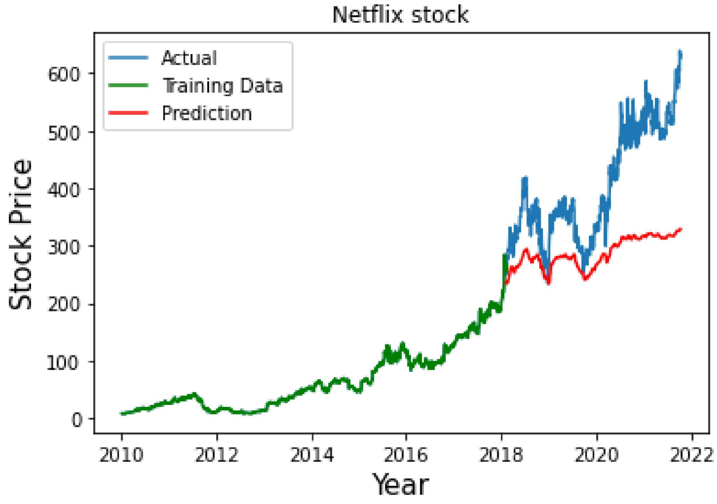




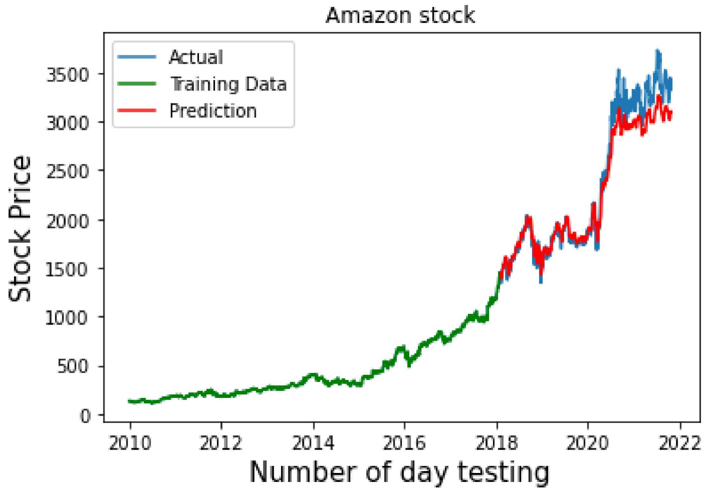
Graphs for Our LSTM Model

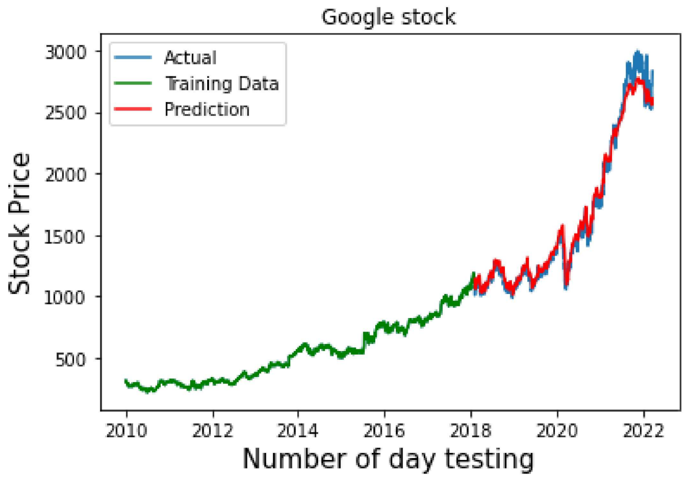


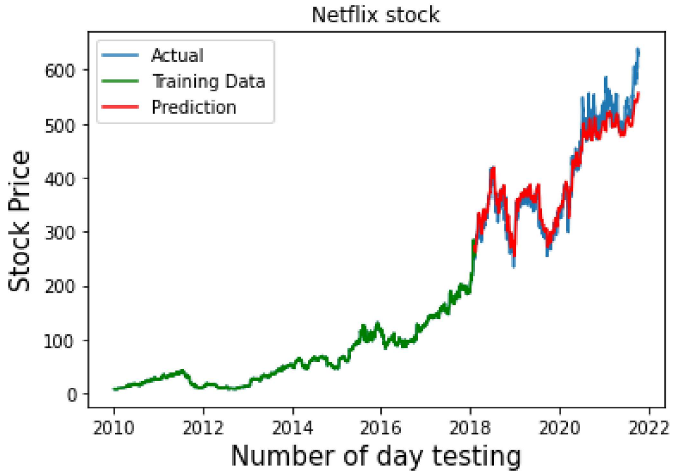




Graphs for Our GRU Model







Conclusion

From our evaluations of these four different models in our project, we are able to learn more about which model is efficient for which kind of specific purposes. While the linear model can be useful in predicting outcomes in a linear relationship between two factors, it is not the best tool to use for stock price prediction since the stock prices of companies, especially the ones in the technology industry that our datasets represented, had a more of an exponential growth in the past twenty years. Based on the results of the logistic regression model, this kind of model is more appropriate for more binary kinds of conditions. While we implemented and adapted a logistic regression model to determine the relationship between articles about these companies' stocks and the general trend in the change of stock prices, this model was not ideal for our original purpose of predicting more specific trends in the stock prices of the three companies we focused on in our experiments. As our graphs of the LSTM and GRU models that we researched and applied to our goal illustrate, these two models are more suitable for data that is fluctuating more over longer periods of time. Based on our testing and results in our experiment, we have learned that the GRU model is the most accurate in predicting future stock prices based on a data set of the previous stock prices of the corresponding company. Through this final project, we were able to further explore and apply models that we learned both in this LIGN 167 class and during our extra research in order to achieve our initial purpose to solve the problem of evaluating which kind of models are the most accurate in predicting the future trends and stock prices of multiple companies.

Source Citations

[1] Mahmood, A. (2019, April 20). *Linear regression with pytorch*. Medium. Retrieved December 8, 2022, from https://towardsdatascience.com/linear-regression-with-pytorch-eb6dedead817/

[2] Swaminathan, S. (2019, January 18). *Logistic regression - detailed overview*. Medium. Retrieved December 8, 2022, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[3] Saldanha, R. (2020, June 3). *Stock price prediction with pytorch*. Medium. Retrieved December 8, 2022, from https://medium.com/swlh/stock-price-prediction-with-pytorch-37f52ae84632

[4] Saldanha, R. (2020, June 3). *Stock price prediction with pytorch*. Medium. Retrieved December 8, 2022, from https://medium.com/swlh/stock-price-prediction-with-pytorch-37f52ae84632

[5] Ravinther, K. (2021, October 27). *Amazon Stock Price (all time)*. Kaggle. Retrieved December 8, 2022, from https://www.kaggle.com/datasets/kannan1314/amazon-stock-price-all-time

[6] Verma, A. (2022, March 25). *Google Stock Data*. Kaggle.

Retrieved December 8, 2022,

from https://www.kaggle.com/datasets/varpit94/google-stock-data

[7] Abhi. (2021, October 12). *Netflix stock price (all time)*. Kaggle. Retrieved December 8, 2022, from https://www.kaggle.com/datasets/akpmpr/updated-netflix-stock-price-all-time

[8] Aaron7sun. (2019, November 13). *Daily News for Stock Market Prediction*. Kaggle. Retrieved December 8, 2022, from https://www.kaggle.com/datasets/aaron7sun/stocknews?resource=download

[9] *TFIDF + scikit-learn SVM*¶. TFIDF + scikit-learn SVM - Podium 2020 documentation. (n.d.). Retrieved December 8, 2022, from https://takelab.fer.hr/podium/examples/tfidf_example.html