Marlene Aviles

630: Predictive Analytics

Final Course Project

**Executive Summary**

The concept of a stock market and the components that shape up this business can be intimidating to any individual. As awareness of the concept of stock markets is increased, the terms and flow of the market become more straightforward to understand. A stock market is made up of several different exchanges, such as the New York Stock Exchange, the S&P 500 index, and the Nasdaq. The exchanges where the stocks are listed are composed of buyers and sellers together, and it can be viewed as a market for the shares of those stocks (O'Shea, 2020). Standard and Poor's 500 Index was selected for this predictive analytics project, given that it is made up of the 500 largest U.S. publicly traded companies. The S&P 500, for short, contains the most popular and profitable tech companies we have all heard of such as Apple, Amazon, Google, and Microsoft. Tech companies were a focus for this project given that they currently make up approximately 18% of the S&P 500 market cap. This makes them desirable companies to consider in selecting and investing in for investors. During this predictive analytics project, data from these four particular tech companies were analyzed, to assist institutional investors in deciding what companies to invest in within the forthcoming years. This analysis was conducted to identify prior trends from the market and predict upcoming trends. This analysis was made up of two different predictive models, known as the Long Short-Term Memory (LSTM) model, and an Auto-Regressive Integrated Moving Average (ARIMA) model. The utilization of these models helped in predicting the future stock value of the most appealing tech company out of the four selected, which ended up being Amazon. Amazon was the focus of the analysis after initial data exploration was conducted, given that it had the highest closing amounts in the stock market from 05/2019 to 05/2020. The overall results of this analysis reflected that the use of the LSTM model was the most accurate model when predicting future stock market closing prices. This particular model also had the lowest margin of error out of the two models and should be preferred to assist institutional investors.

**Abstract**

Standard & Poor's 500 Index has become a popular market-capitalization-weighted index based on the 500 largest U.S. publicly traded companies. The S&P 500, for short, has over time become an indication used as one the most appropriate measure of large-cap U.S. equities (Kenton, 2019). The S&P 500 index and the most recent historical data from the years 2019 to 2020 will be applied to create a predictive model that will assist institutional investors in deciding what companies to invest in within the forthcoming years based on prior trends from the market. The companies included in the S&P 500 index can additionally be part of other leading indexes such as the New York Stock Exchange (NYSE) or the Nasdaq Composite Index. The S&P 500 plays a significant role in measuring how the overall U.S. stock market and economy are doing. Given that tech companies control about 18% of the S&P 500 market cap, this predictive analytics project will focus on the most popular and profitable tech companies of the index only, which include Apple, Amazon, Google, and Microsoft.

**Introduction & Background of the Problem:**

The term stock market refers to the collection of markets and exchanges where transactions such as buying, selling, and issuance of shares of publicly held companies take place (Chen, 2020). The stock market and terms such as the S&P 500, the Dow Jones Industrial Average (DJIA), and the Nasdaq Composite Index are well-known terms in this domain. Many people have heard these terms, but only a handful of individuals know the ins and outs of how the stock market works. Hence this general dilemma, this analysis will look at the historical stock market data from 2019 to 2020 of the most profitable tech companies to attempt and identify which tech company out of the four selected should an institutional investor invest in. The companies that partake in the S&P 500 index without question are part of the 500 largest U.S. publicly traded companies, without in-depth analysis, the institutional investor may become overwhelmed with which tech company to focus on and invest in. This analysis will help concentrate on the most successful tech company, to assist the institutional investor or investors on investing on this one company, based on the patterns and trends found during the year 2019 - 2020.

**Methods**

Technical Approach:

The analysis will focus on creating a predictive model for the most successful tech company using the most recent data from the S&P 500 stock market index to find any significant patterns and trends. Predictive analytics will be used for this stock market analysis as it can generate substantial improvements in efficiency, decision-making, and return of investment for the institutional investor(s) (Abbott, 2014). The technical approach taken for this project will consist of following the cross-industry process for data mining (CRISP-DM) steps to make sure the adequate data and problem are analyzed. The CRISP-DM steps will include the following:

• Business Understanding: Establish a problem

• Data Understanding: Explore the data and identify issues or missing values in the data

• Data Preparation: Apply the proper data cleaning tools and create new variables/dataset

• Modeling: Build a predictive model(s)using machine learning techniques

• Evaluation: Assess the models applies and select the best one

• Deployment: Select the best model and deploy

Data Source:

The data utilized for this project consists of historical stock data collected from the Yahoo Finance Website, which provides access to data from the Exchange to developers and engineers for free. This analysis focuses on four of the top tech companies from May 2019 to May 2020, as this can help reflect the forthcoming trends and patterns of these companies. A second dataset was extracted from the Yahoo Finance website regarding the most profitable company, Amazon, from May 2012 to May 2020 to create a more representative predictive model.

Analysis:

Business Understanding:

Institutional investors may become overwhelmed with the number of companies available. This analysis will help to minimize the number of companies that the institutional investor or investors should focus on investing based on the patterns and trends found during the years, to only one. Predictive modeling will be applied to the most appropriate tech company data for this problem. The patterns and trends will help identify if these tech companies will be profitable or not in the forthcoming years.

Data Understanding:

The dataset collected from Yahoo Finance on the S&P 500 index is very extensive given that it collects data for many years, given this scenario the most recent data from 05/2019 to 05/2020 were filtered out to create a representative subset. This data will be utilized to identity the most profitable company to apply predictive modeling and predict future closing prices for the most appropriate tech company selected. A total of seven variables in the dataset will be analyzed for the four tech companies.

Variables in the dataset:

Date - in format: yy-mm-dd

High - Highest stock price reached in the day

Low - Lowest stock price reached in the day

Open - Stock price when the stock market opened (in USD)

Close - Stock price when the stock market closed (in USD)

Volume - Number of shares traded

Adj Close - Adjusted closing stock price

company\_name - Name of the tech company

A screenshot of a cell phone

Description automatically generated

Data Preparation:

The dataset was prepared by first filtering out the tech companies for the analysis, and the appropriate date ranges were entered for proper extraction. The company\_name column was modified to rename the tech companies to their full company name. This assisted in creating a subset of the four tech companies using the Yahoo Finance website. The summary statistics and visualizations of each company were then analyzed to identify the unique values that it contained, along with any missing values and outliers. A second dataset was created with Amazon historical data from May 2012 to May 2020 for predictive modeling. For the second dataset with Amazon data, dropping the "Close" column was required for predictive modeling. The data was then split into a test and training set using an 80:20 ratio. The data were then standardized and transformed using the MinMaxScaler to make sure the stock data ranged from 0 to 1. When using this method, missing values are treated as missing values and are disregarding in the fit. Below is a screenshot of the second dataset, which contains stock data from Amazon only.

A screenshot of a cell phone

Description automatically generated

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) was conducted to identify the most profitable tech company from the four selected ones. EDA managed to look at factors such as correlation, the adjusted closing stock prices, the daily traded volume, and the moving averages (MA) for 10, 20, and 50 days.

![A screenshot of a cell phone

Description automatically generated]()

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

A picture containing text, map

Description automatically generated

A close up of a map

Description automatically generated

Modeling:

A total of two predictive models were applied to the most profitable tech company in this case Amazon, for deeper analysis with the use of packages such as scikit-learn and pyramid.

• Long Short-Term Memory (LSTM) model was selected given that it creates an artificial recurrent neural network (RNN) and attempts to model time or sequence-dependent behavior.

• Auto-Regressive Integrated Moving Average (ARIMA) model was selected given that it uses time-series data to better understand a data set and predict future trends.

**Results**

Evaluation:

Long Short-Term Memory (LSTM)

The Long Short-Term Memory models were created for Amazon to predict the future closing stock values of the company for the rest of the year 2020. This model was created by using Keras simplest sequential model, which contained the following layers.

model = Sequential()

model.add(LSTM(50, return\_sequences=True, input\_shape= (x\_train.shape[1], 1)))

model.add(LSTM(50, return\_sequences= False))

model.add(Dense(25))

model.add(Dense(1))

This particular model was evaluated by using the Root Mean Square Error (RMSE), given that this metric is the standard deviation of the prediction errors or residuals. Residuals are a measure of how far the regression line data points are, and the RMSE measures how spread out these particular residuals are. The LSTM model gave a significantly low RMSE of 6.39 and the accuracy was noted via a time series visual comparing the trained values, valid values, and predictions from the model. A data frame reflecting the valid and predicted values was also created.

A close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Auto-Regressive Integrated Moving Average (ARIMA)

The ARIMA model was created for Amazon to predict the future closing stock values for the company for the remainder of the year 2020. This model used training and validation to forecast predictions. The RMSE was extremely high in this model, which executed RMSE results of 531. This large RMSE value shows that this model was not accurate in predicting the Amazon closing stock prices. This large inaccuracy was also noted via a time series visual given that the predicted values were not close to the valid values of the model.

A close up of a map

Description automatically generated

Deployment: None of the two models were deployed, given that further analysis and modifications need to be prepared for deployment. The parameters for both models need to be further adjusted for the best results.

**Discussion/Conclusion**

The following analysis of predicting the closing stock prices for the most profitable tech company, Amazon, was a great challenge given that predicting stocks is a very complex task. The best predictive model out of the two created was beyond doubt the Long Short-Term Memory (LSTM) model. This model was able to give a very low RMSE, which only means that the predicted values from the model created are very close to the actual valid numbers. This type of results gives a confident notion that this model can be further adjusted, via its parameters, to recognize that this company is worth investing in and predicting that it will not crash. Throughout the time that the data was extracted, the world was dealing with a global pandemic of COVID-19. During the early months of 2020, many companies were directly affected, causing their stocks to crash or greatly decrease as many individuals in the United States were left out of work, and investment was considerably affected. Amazon's stock data and this project have managed to show investors that this company was not enormously affected during this world crisis, and its stocks continued to succeed. This type of pandemic could have not been predicted by anyone, but the results from this analysis should help investors in choosing to invest in this company, given that their stocks managed to greatly rise after a minimal decrease during this pandemic while many other companies did not.

**Acknowledgments**

I want to thank all of the writers and data enthusiasts that have made this foreign subject more bearable with the vast amount of resources they have managed to provide the world with. The resources created by these individuals have managed to offer excellent explanations of how the S&P 500 index operates. Lastly, I want to thank my spouse for being patient with me while working on this project and contributing to some incredible brainstorming sessions on this subject manner.

**References**

Abbott., D. (2014). Applied Predictive Analytics. John Wiley & Sons, Inc.

Chen., J. (2020). Stock Market. Retrieved from https://www.investopedia.com/terms/s/stockmarket.asp

Kenton., W. (2019). S&P 500 Index – Standard & Poor's 500 Index. Retrieved from

https://www.investopedia.com/terms/s/sp500.asp

O'Shea., A. (2020). Stock Market Basics: What Beginner Investors Should Know. Retrieved from https://www.nerdwallet.com/article/investing/stock-market-basics-everything-beginner-investors-know

S&P 500 Stock Data. (2020). Retrieved from https://www.kaggle.com/alexsumt/sp-500-stock-data

S&P 500 stock data: Historical stock data for all current S&P 500 companies. (2018). Retrieved from https://www.kaggle.com/camnugent/sandp500#AAPL\_data.csv

Sozzi., B. (2020). 5 powerful tech companies now make up 18% of the stock market — here's why this could be a bad thing. Retrieved from https://finance.yahoo.com/news/5-powerful-tech-companies-now-make-up-18-of-the-stock-market-heres-why-this-could-be-a-bad-thing-145710881.html