# Prediction Model for Long Term Investment on the Stock Market

Victor Enchautegui, Joey Chan, Jiazhen Cui
College of Computing & Informatics, Drexel University, Philadelphia, PA 19104, USA

**Abstract**—.In the era of Cloud Computing and Machine Learning (ML), utilizing big data analysis to predict complex trends has become possible. Predicting stock market prices is one of those complex trends that is highly popular due to its large pool of available public data and significant reward if solved. Unfortunately, the stock market is an extremely complex and volatile system. In this paper, we will utilize Cloud Computing and a learning-based model to analyze 22 years of daily data to predict a company's future stock prices. The proposed solution will utilize various techniques to pre-process the big dataset, stage the data to be used for our learning model, and train and test our model to predict price trends.

**Keywords**—Classification, Stock Market, PySpark.

## 1 INTRODUCTION

The stock market is always changing and reliable data used to predict the market is more significant than ever. For investors, it is important to find trends in stocks, view forecasted future stock prices, and investigate the impact of investments and risks. The investor community is reliant on information (i.e., stock data & analysis), which enables investors to participate fully within the investing community and stock market to grow their portfolio.

Though there are many strategies to forecast the performance of the market, machine learning modeling is the most coveted, but also still in its infancy after years of Research and Development (R&D) and utilization. Using a gradient boosting machine learning model implemented with PySpark, this study aims to predict market outcomes of a selected stock (i.e. APPLE) and identify from the dataset attributes if there is a consistent pattern to predict future stock price.. In other words, we would like to investigate and explore if it is possible to remove that uncertainty from a stock's price based on its attributes and historical market data. Once we assemble our final dataset, we will make it freely accessible to anyone who is interested in our data model.

## 2 Stock Dataset

Apple's historical dataset is from a Python module called yfinance module. The module uses Yahoo Finance, which is a free data source with no registration required. The data extracted comprises all data from a 10-year interval to current date. Each row describes the attributes of the stock: Date, Open, High, Low, Close, Volume, Dividends, Stock Splits. The target variable is Close, which is the closing price of the stock at the end of daily trading (shown in Figure 1). The predictor variables are Open and Volume, which are the opening price of the stock at the start of daily trading and the amount of shares traded daily, respectively (shown in Figure 2 and 3). The dataset has a total of 5579 samples and 8 variables. Table 1 shows the sample variables description in this dataset.

| Variable | Description |
|---|---|
| **Date** | The date of trading transactions. |
| **Open** | The price from the first transaction of a business day. |
| **High** | Highest closing price of a stock over the past 52 weeks. |
| **Low** | Lowest closing price of a stock over the past 52 weeks. |
| **Close** | The price from the last transaction of a business day. |
| **Volume** | Measured in the number of shares traded |
| **Dividends** | A distribution of profits by a corporation to its shareholders. |
| **Stock Splits** | Divide stock to increase the number of shares in a company |

**Table 1:** Sample APPLE Dataset Variables Description.

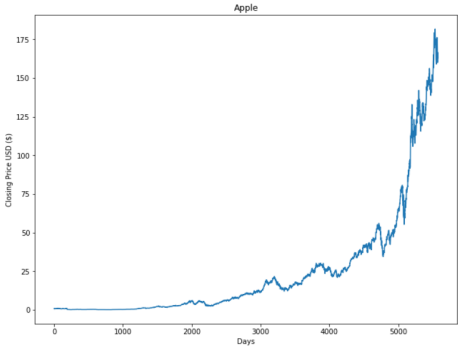


**Figure 1:** APPLE Daily Closing Prices.

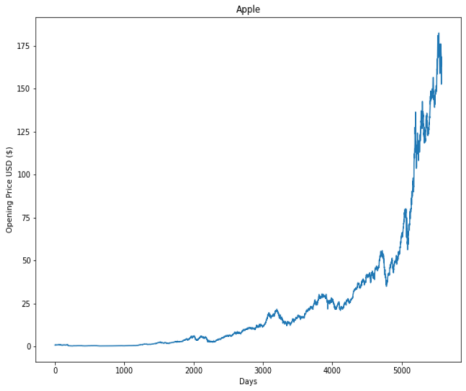


**Figure 2:** APPLE Daily Opening Prices.

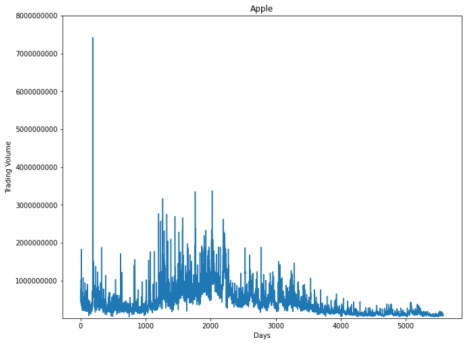


**Figure 3:** APPLE Daily Trading Volume.

## 3 Exploratory Data Analysis

For analysis, the team sought to answer the following exploratory questions:

1. What was the change in price of the stock over time? (Illustrated in Figure 1)
2. What was the daily return of the stock on average?
3. What was the moving average of stocks?
4. How much value do we put at risk by investing in a particular stock? (Illustrated in Figure )
5. How can we attempt to predict future stock behavior? (Illustrated in Figure )

Through statistics and distribution, we may uncover discrepancies or abnormalities in the data, which may require further analysis outside the initial scope of the dataset.

Table 2 shows the average daily return of Apple as approximately -0.0067753% over a period of 10 years with minimum return of approximately -13.64240% and maximum return of ~11.60442%. The standard deviation of the average daily return is approximately ± 2.043099, so the mean is within a range of -0.0067753 ± 2.043099% or -2.0498743 to 2.0363237. Figure 4 shows the distribution of the daily return of Apple over the sample size period.

| summary | Daily Return % |
|---|---|
| count | 5579 |
| mean | -0.00677528780333... |
| stddev | 2.0430985570262115 |
| min | -13.642403644742629 |
| max | 11.604417620254818 |

**Table 2:** Summary of APPLE Average Daily Return.

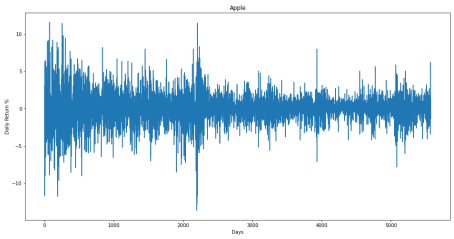


**Figure 4:** APPLE Average Daily Return.

Figures 5, 6, 7, 8 , 9, and 10 show Apple's moving average over various time intervals – 5, 10, 20, 50, 100, and 200 days – to the current date. These figures illustrate that the larger the sample size of data used the better fit the moving average is to the stock's closing price.

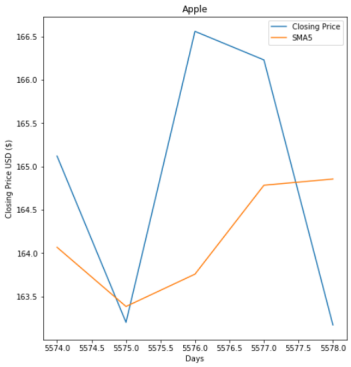


**Figure 5:** APPLE Closing Price vs SMA5.

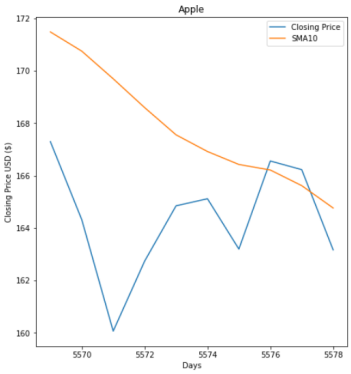


**Figure 6:** APPLE Closing Price vs SMA10.

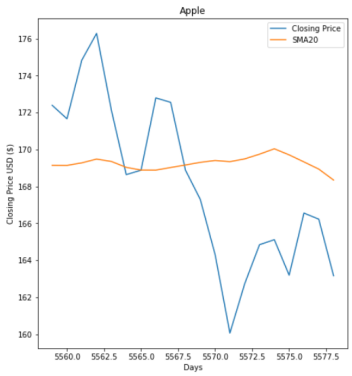


**Figure 7:** APPLE Closing Price vs SMA20.

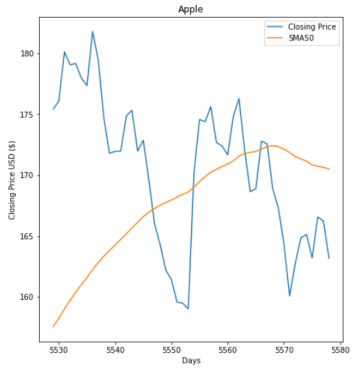


**Figure 8:** APPLE Closing Price vs SMA50.

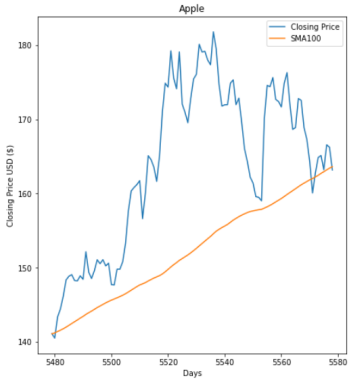


**Figure 9:** APPLE Closing Price vs SMA100.

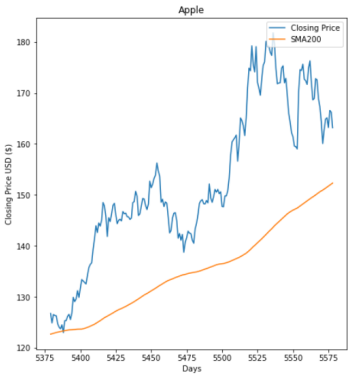


**Figure 10:** APPLE Closing Price vs SMA200.

## 4 METHODOLOGY

### 4.1 Data Preprocessing

In order to predict the closing price of Apple shares, the following preprocessing operations are required to transform raw data into an understandable input for our machine learning model: (1) Data Cleaning: 'High', 'Low', 'Dividends', and 'Stock Splits' were dropped from the dataset as these attributes were not valued as strong features within the analysis. For the Apple stock, there was not any missing data from our source. Null fields are more frequent in stock price data prior to 2012. (2) String Indexing and Encoding Categorical Data: The stock dataset is all numerical. The analysis is based on predicting one stock ticker; there are no stock tickers or any other string values within the dataset. (3) Vector Assembling: a vector of 'Open' and 'Volume' features was created. (4) Dataset Splitting: the data was split to 80-20 train-test. These preprocessing steps were set as "stages" within our script in order to be utilized within a Pipeline.

### 4.2 Machine Learning Modeling

The training set was trained on a binary classification model, gradient-boosted tree (GBT). After training, the model is evaluated on the test set. The metric that is used is root-mean-square deviation (RMSE). Predictions made from the GBT model were added into the current dataset for Apple Stock to create a new dataset used to visualize and compare Apple closing price and GBT model predictions over various intervals of time (i.e., 5, 10, 20, 50, 100, and 200 days). In this project, the team built an end-to-end machine learning pipeline using PySpark.

## 5 RESULTS AND DISCUSSION

The results of the predictive modeling somewhat sufficiently affirm the predicted closing price to actual closing price based on the attributes and historical Apple stock data. RMSE is used as the main scoring metric for this project. It determines how skillful a classifier is at predicting quantitative data. In this case, RMSE is relative to the data, so Apple prices are within the range of $150 - $160 based on recent data illustrated in Figures 11 - 16.

The RMSE value yield from the test data is 4.085066659756554, which is high considering RSME is usually between 0 and 1. However, when RMSE is done on the train data, it yields 3.5996544460802746, which is similar to test data RMSE. This indicates that the dataset is valid and that the data is not over fitted or under fitted as:

RMSE of test > RMSE of train => OVER FITTING of the data.
RMSE of test < RMSE of train => UNDER FITTING of the data.

Therefore, our machine learning modeling is acceptable. From this, we can say that the risk is low using this model to invest in any particular stock. When comparing and contrasting Figures 5 - 10 to Figures 11 - 16, our model better predicts stock behavior and closing prices than moving averages over various time intervals.

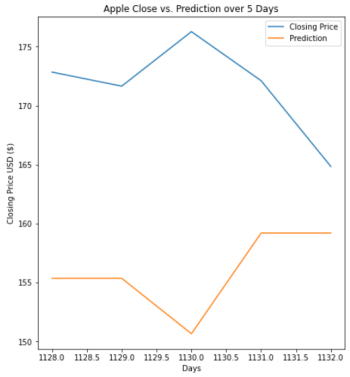


**Figure 11:** APPLE Closing Price vs Prediction over 5 Days.

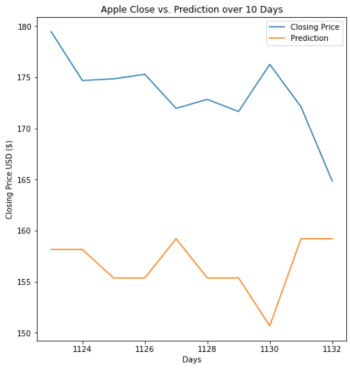


**Figure 12:** APPLE Closing Price vs Prediction over 10 Days.

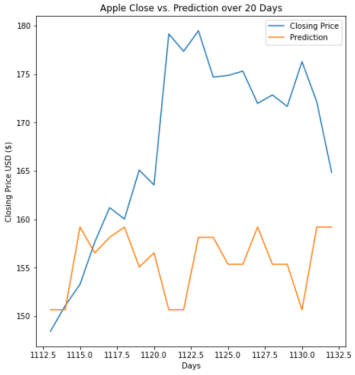


**Figure 13:** APPLE Closing Price vs Prediction over 20 Days.