Using Machine Learning Models to Predict Stock Price Trends Based on External Events

Sanskriti Seernani and Stephanie Ryan

Abstract

This paper was aimed at applying machine learning models such as logistic regression and the Naive Bayes Classifier to data samples from external events that often affect stock price in order to determine future prices of company stocks. The stock that was used for experimentation was the Apple Inc. (AAPL) stock. The external events factored in were Apple's quarterly reports, and the news headlines on a said day. Logistic regression and Naive Bayes were applied to both data sets. The quarterly report data sample was small and so, cross validation was applied to it. It was found that logistic regression gave higher *f*-Measure and accuracy values in the case of the quarterly report data where as the Naive Bayes Classifier ave higher *f*-Measure and accuracy values when applied to the daily news headlines. It was then concluded that logistic regression is a better fit for data similar to the revenues pulled from the quarterly reports, while Naive Bayes was more effective for large data samples like the news headlines data set.

Contents

1 Background

Machine learning is a branch of artificial intelligence that comprises of methods for analyzing data via the automation of learning models. It relies on the fact that systems learn from data, identify patterns and work independent of human intervention. Logistic Regression and the Naive Bayes Classifier are two such machine learning models. Logistic regression uses a log function to model a binary dependent variable to estimate the parameters needed while the naive Bayes classifier employs Bayes' theorem coupled with strong independence assumptions between features.
Recently, there has been a significant rise in the number of individuals investing in the stock market. Given how unpredictable the market is, it is essential to build a tool that will help investors maximize their profits. In order to do this, logistic regression was applied to two of the many external factors that influence stock price- 1. Quarterly revenue percentages and 2. Daily news headlines. This data was also used when employing the naive Bayes classifier as a means to compare the results of logistic regression with. This information was then used to predict future stock price values.

2 Related Work

Researchers in Serbia studied the effect of factors such as company size, return on equity, return on assets, earning per stock , book value, and some others, on the company's stock prices on the Belgrade stock exchange. They found that all the factors combined had a unique influence on the stock prices.

Researchers in Nigeria used regression analysis to predict stock prices. Their approach was focused on data mining. The obtained their data from the daily list of stock prices, officially published by the Nigerian stock exchange. They extracted values from the database and used data mining tools to predict future values by using time series data, which used the moving average method.

3 Methodology

3.1 Data

This project was divided into two parts. The first being the prediction of stock price trends based on quarterly report information. The second is the prediction of stock trends due to new headlines on the given day. Two separate datasets were used for each part.

3.1.1 Quarterly Report Prediction Stock Trends

For this prediction, we needed to gather the close price of stock two days prior to the release of quarterly report. We then calculated the percentage of revenue increase or decrease in comparison to the previous term. The percentage change in revenue was determined by calculating the difference in revenue between the current and previous quarter. These three elements acted as the X values in the models while the Y values was 1 or 0 depending on how the stock close price fluctuated from the opening price. If the stock increased, it was represented by 1 and if it decreased, 0 was the label.

Due to time limitations, we focused on Apple stock (AAPL) with quarterly reports dating back to January 2016. The historical stock price data for AAPL was obtained from Yahoo Finance[5] while the quarterly report information was gathered from Apple Newsroom[6].

As this was a relatively small dataset of 22 rows, we decided to used cross validation on all the elements which utilizes the data better. By using this method instead of the test-train split usually implemented, we build 22 different models enabling us to make predictions on all of our data.

3.1.2 News Prediction Stock Trends

In the prediction of stock price trends based on news articles, we used a combined dataset from kaggle called *Daily News for Stock Market Prediction*[7]. This data consisted of the date, the Dow Jones Industrial Average (DJIA) on that given date and lastly the top 25 news articles from Reddit WorldNews Channel. The period of time of the dataset was 2008-06-08 to 2016-07-01.

The data was cleaned up to remove stop words. After special characters were striped for the sentences, each of the words were checked to see if it is in a contraction list to check for words like *"we've"* to convert it into its expanded for *"we have"*. Lemmatization on the words was conducted to group together the inflected forms of words to be analysed as a single item.

To vectorize the news article information, the Global Vectors for Word Representation (GloVe) was used. GloVe is an unsupervised learning algorithm which obtains vector representation of words[8]. By implementing GloVe, the larger common crawl vectors were identified to create the word embeddings for the project. The code for the vectorization of the news was taken from an open source platform on Github[8].

The the length of a day's news was limited to 200 words, and the length of any headline to 16 words. These limitation values were chosen as they prevent from an excessively long training time while balancing the amount of headlines used and the number of words from each headline.

The padded count of the words in the daily headlines was used as the theta values in the model calculation with the DJIA increase or decrease as the Y values, where 1 denoted the increase in the average and 0 represented a decrease.

3.2 Models

3.2.1 Logistic Regression

Logistic Regression is a statistical classifier that employs the use of a logistic function to model a binary dependent variable. The logistic regression model was applied to the datasets.

The data was first randomized. The data was split into training and testing groups. Both the train and test data were standardized. The *theta* parameters were initialized using random values in the range $[-1, 1]$. Follwing this, a batch gradient descent was performed on the data. This was scheduled to terminate when 1500 iterations had passed or when the absolute value change in loss on the data was less than 2^{-23}. The learning rate, η, was 0.01. Each test sample was then classified using the model and the class labels were chosen based on which class probability was higher. The model used testing data results to compute the *Precision, Recall, F-Measure, Accuracy.*

3.2.2 Naive Bayes Classifier

The naive Bayes classifier is a machine learning model that is based on Bayes theorem i.e. it is a probabilistic model. It applies Bayes' theorem to the data along with strong independence assumptions between features.

The data was first randomized. The data was split into training and testing groups. Both the train and test data were standardized. The training data was divided into two groups, *spam samples* and *non-spam samples*. A normal model was created for each feature for each class. Each testing sample was classified using these models and the class labels were chosen based on which class probability was higher. The model used testing data results to compute the *Precision, Recall, F-Measure, Accuracy.*

3.3 Comparison

The paper examines the data collected from Logistic Regression and Naive Bayes in various ways to pinpoint which model is more appropriate to use for predicting stock price trends due to certain events. Precision and recall rates examine the true-positive, false-positive, and false-negative results to investigate the algorithms accuracy further.
In the classification process of the predicted Y values, a threshold of 0.5 was used. If the the predicted Y was below this threshold, it was considered a 0, i.e. the stock price would go down. On the other hand, if the predicted Y was greater than or equal to the threshold, the Y value was updated to 1, i.e. an increase in the stock price.

Precision calculates the proportion of positive identifications correctly classified with the following formula[3]:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall computes the proportion of true-positives identified correctly using the formula below[3]:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The f-Measure calculates the harmonic mean of precision and recall, which is a measure of accuracy for a given test. It makes use of the following formula[4]:

$$fMeasure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

An f-Measure value of 1.0, the highest possible score, indicates perfect precision and recall; a value near 0, the lowest possible score, signifies poor performance.

We decided that since we are dealing with stock prices and the predictions could be used in decision to either buy or sell the stock, there is no preference in having a higher precision or recall rate. This was due to the observation that both false negative and false positive results could have a significantly negative impact on the traders position in question. Therefore, we are putting more weight behind having a high harmonic balance between precision and recall. Hence, in this scenario, the higher the fMeasure score, the better the model.

4 Experiments and Results

4.1 Data Information

4.1.1 Quarterly Report Prediction Stock Trends

Models	Precision	Recall	*f*-Measure	Accuracy
Logistic Regression	0.627	0.682	0.642	0.682
Naive Bayes Classifier	0.588	0.545	0.563	0.545

Table 1: Quarterly Report Prediction Stock Trends

Applying logistic regression to our data from news articles gave a precision value of 0.627, a recall of 0.682, an *f*-Measure score of 0.642, and an accuracy of 0.682

The Naive Bayes Classifier gave a precision value of 0.588, a recall of 0.545, an *f*-Measure score of 0.563, and an accuracy of 0.545

Due to the limited size of the dataset, cross validation was implemented. It is notable that the precision was less than the recall in the Logistic Regression model which is due to the model producing more false positive results than false negative. One the other hand, the Naive Bayes Classifier produces more false negative results.

As stated before, we are focusing on analyzing the two models from an accuracy and *f*-measure standpoint as, from a traders perspective, they are the most important. The *f*-Measure score and the accuracy value that result from the logistic regression model were higher than the comparative values for Naive Bayes Classifier. The difference between the *f*-Measure scores was 0.079 while the accuracy difference was 0.137.

4.1.2 News Article Prediction Stock Trends

Models	Precision	Recall	*f*-Measure	Accuracy
Logistic Regression	0.496	0.326	0.393	0.465
Naive Bayes Classifier	0.540	0.558	0.549	0.511

Table 2: News Article Prediction Stock Trends

Applying logistic regression to our data from news articles gave a precision value of 0.496, a recall of 0.326, an *f*-Measure score of 0.393, and an accuracy of 0.465.

The Naive Bayes Classifier gave a precision value of 0.540, a recall of 0.558, an *f*-Measure score of 0.549, and an accuracy of 0.511

The data sample here was significantly large, especially in comparison to the quarterly report data. It is interesting to note that the precision was greater than the recall in the Logistic Regression model, implying that there were more false negative results than false positive. However, the opposite can be said for the results produced by the Naive Bayes Classifier. This is inconsistent with the Quarterly report prediction which could be due to the sample size being drastically different.

By again putting a focus on the *f*-Measure score and accuracy of the models, we can see that, contrary to the results from the analysis of the quarterly revenues, the Naive Bayes Classifier resulted in larger values for the *f*-Measure score and the accuracy than the logistic regression model.

5 Conclusion

This paper explored the possibilities of applying machine learning models, namely the logistic regression model and the Naive Bayes Classifier, to data samples of external factors such as quarterly revenues and daily news headlines to predict their effects on stock prices in the future. The stock used for experimentation was the Apple (AAPL) stock. Percentage quarterly revenue differences were pulled from Apple's quarterly reports. The daily news headlines were pulled from Kaggle's *Daily News for Stock Market Prediction.* The two machine learning models in question were applied to both data sets. The quarterly revenue data sample was significantly small and in order to get better results, cross validation had to be applied to that data. The *f*-Measure and accuracy values that were obtained from applying logistic regression were higher than those obtained from applying the Naive Bayes Classifier. This implies that logistic regression would be the better model to use for data similar to quarterly revenues.
On the other hand, when working with a very large data sample that comprised of numerous theta values such as daily news headlines, the Naive Bayes Classifier yielded higher *f*-Measure and accuracy values than logistic regression, making it the better suited model for similar data.

6 Future Work

For future extensions of this project, the researcher can experiment with a larger and more diverse data set. Data such as different stocks and indexes over a longer time period. The researcher can also explore other external factors like COVID focused news articles, tax season, dividend release etc. Additional machine learning models can also be employed to provide for comparison such as the least absolute shrinkage and selection operator (LASSO) model.

References

[1] Miloevic Avdalovic, S., Milenkovi, I. (2017). IMPACT OF COMPANY PERFORMANCES ON THE STOCK PRICE: AN EMPIRICAL ANALYSIS ON SELECT COMPANIES IN SERBIA. Economics Of Agriculture, 64(2), 561-570. doi:10.5937/ekoPolj1702561M

[2] Olaniyi, Abdulsalam S., Adewole G, Jimoh. (2011). Stock Trend Prediction Using Regression Analysis A Data Mining Approach. ARPN Journal of Systems and Software. 1. 154-157.

[3] Google. (n.d.). Classification: Precision and Recall. Machine Learning Crash Course. Retrieved June 3, 2021, from https://developers.google.com/machine-learning/crash-course/classification/precision-and -recall.

[4] Brownlee, J. (2020, January 3). How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Machine Learning Mastery. Retrieved June 3, 2021, from Analysis of Facial Processing Technology https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-class ification/.

[5] Yahoo! (2021, June 3). Apple Inc. (AAPL) Stock Historical Prices amp; Data. Yahoo! Finance. https://finance.yahoo.com/quote/AAPL/history?p=AAPL.

[6] Apple. (n.d.). Newsroom - Search. Apple Newsroom. From https://www.apple.com/newsroom/search?q=quarterly%2Breports.

[7] Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. from https://www.kaggle.com/aaron7sun/stocknews.

[8] Pennington, J. (n.d.). GloVe: Global Vectors for Word Representation. from https://nlp.stanford.edu/projects/glove/

[9] Currie, D. (2017, May 2). Predicting the Stock Market with the News and Deep Learning. Medium. from https://medium.com/@Currie32/predicting-the-stock-market-with-the-news-and-deep-learning-7fc8f5f639bc.