Final Paper Project

Predicting Next Day Closing Stock Price Direction using Machine Learning Methods.

By Achraf Safsafi

Bellevue University, DSC 630, Dr. Brett Werner

Abstract:

Stock price prediction is a priority goal of every investor or trader, so it enables them to reduce risks and increase profits by analyzing past records. This paper focuses on the best independent variables and indicators to predict next-day closing stock price direction using machine learning methods. Different machine learning techniques include Logistic Regression, Gaussian Naive Bayes, Support Vector Classifier (SVC) using RBF kernel, Decision Tree, Random Forest, XGBoost are applied to foresee the stock price of Apple. The models performance would be compared using two metrics, Accuracy, and AUC (Area under the receiver operating characteristic curve). Experiment results suggest that the Random Forest, and XGBoost classifiers are more fitting for Predicting the Next Day Closing Stock Price Trend.

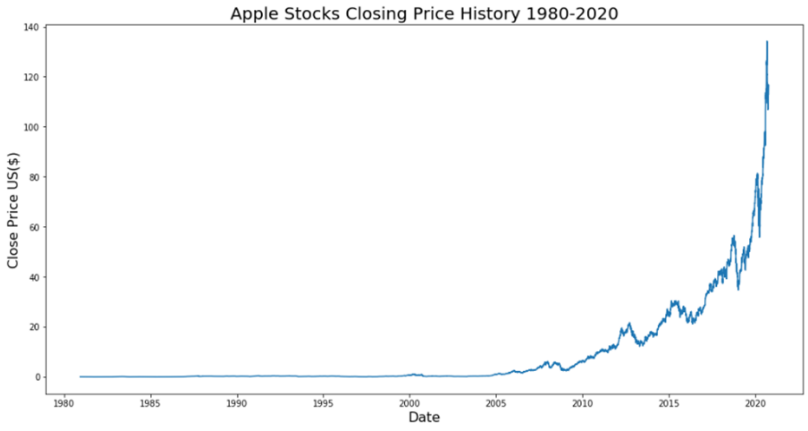
Introduction:

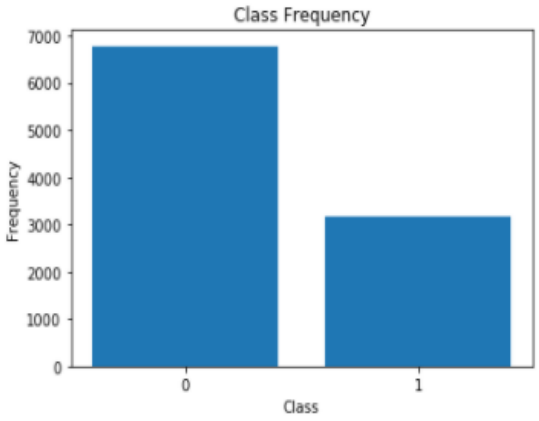
Day trading is a quick way to obtain gains faster. However, it needs having a lot of experience and accessible information to make a decision to buy the right stock that will likely go up. the objective of this project is to develop a prediction model for price trend prediction. The model focuses on a one-day prediction of the closing price trend of the Apple stock. Logistic Regression, Gaussian Naive Bayes, Support Vector Classifier (SVC) using RBF kernel, Decision Tree, Random Forest, and XGBoost are chosen to be applied. The past prices of the Apple stock are taken from the Yahoo Finance website since its debut (1980-12-12) until the present. Besides to Open, High, Low, and Close features, some Buy&Sell signals are using as independent variables.

Data sets:

In this project, the data are extracted from yahoo finance. They are collected from 1980 to up to date.

Open, high, low, and close price are used as predictors. Open and close features represent the opening and closing price at which the stock is traded on a particular day. High and low represent the maximum, and minimum price of the share for a day. Volume is the number of shares traded on a specific day. In addition, Moving Average Convergence Divergence (MACD), Williams %R (WR), Relative Strength Index (RSI) are selected as additional predictors, and 2 days simple moving average on open price (open_2_sma).





The applied classifiers are used to predict a binary outcome, buy, or sell stock. The target variable is taking two values '0' or '1'. The values are classified as class '1' if the closing price of the next day is higher than the closing price of the actual day, else, the values are classified as class '0'. A threshold value of 0.5 is selected. Above 0.5, values are classified into class 1. Else, the values are classified into class 0. If tomorrow's closing price is higher than today's closing price, we can buy the stock, else we can sell it.

Data Preparation:

Data Cleaning:

A few missing values are found. Their percentage is too negligible. Thus, observations that have missing values are dropped.

Feature scaling:

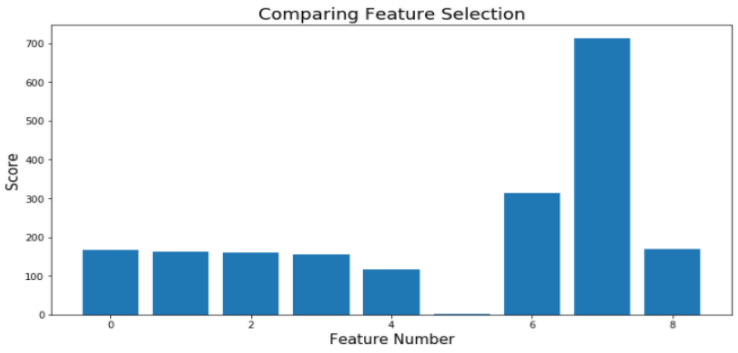
As the scale of some features are different, the data are made on the same scale, between the range of 0 and 1.

Data splitting:

Some data need to be kept for testing the model after the training. the test data must be different from the training set that is used for model training. Thus, the data are split into two different datasets. 67% for training data and 33% for testing data.

Feature Selection:

Feature 0 (open)	: 166.610988
Feature 1 (high)	: 162.526005
Feature 2 (low)	: 161.085385
Feature 3 (close)	: 156.165251
Feature 4 (volume)	: 115.551111
Feature 5 (macd)	: 1.235863
Feature 6 (rsi_12)	: 314.550010
Feature 7 (wr_12)	: 712.764245



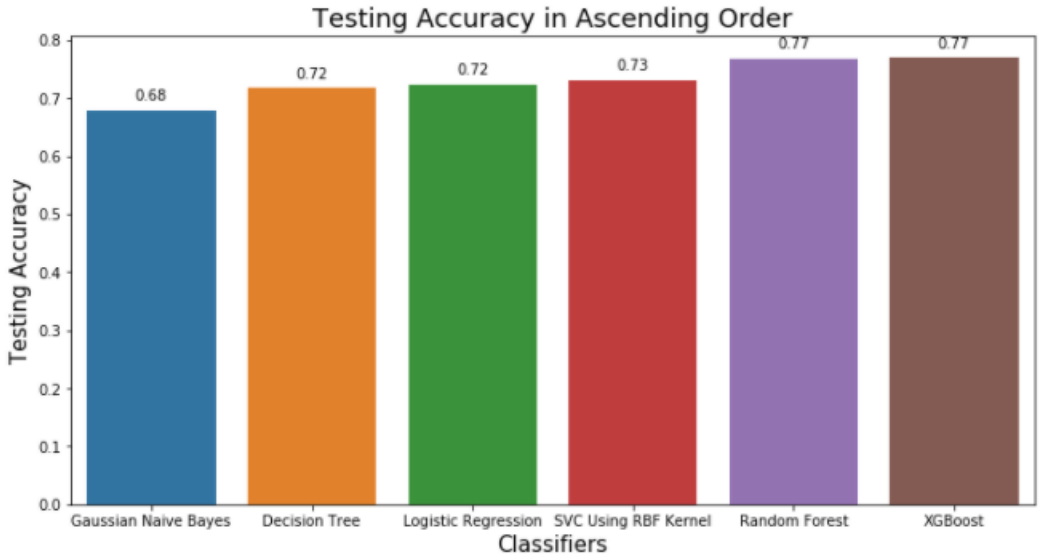
To reduce overfitting, and improves accuracy, we need to identify and select a subset of input features that are most important to the target variable. to do this, Feature Selection using

ANOVA F-value Method will be applied. Usually, this test is used when input data are numerical, and the target variable is categorical. so the result shows that macd indicator has a very low score comparing to others, this means that the indicator is not important and would drop it from the data.

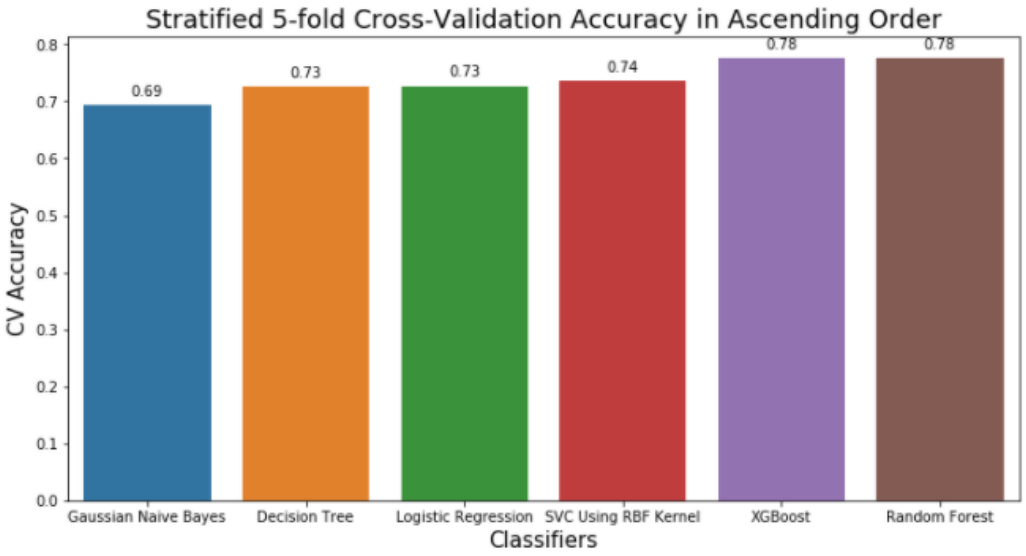
Algorithms:

Common classification machine learning algorithms are chosen for this project include Logistic Regression, Gaussian Naive Bayes, Support Vector Classifier (SVC) Using RBF Kernel, Decision Tree, Random Forest, and XGBoost.

Results /Analysis:

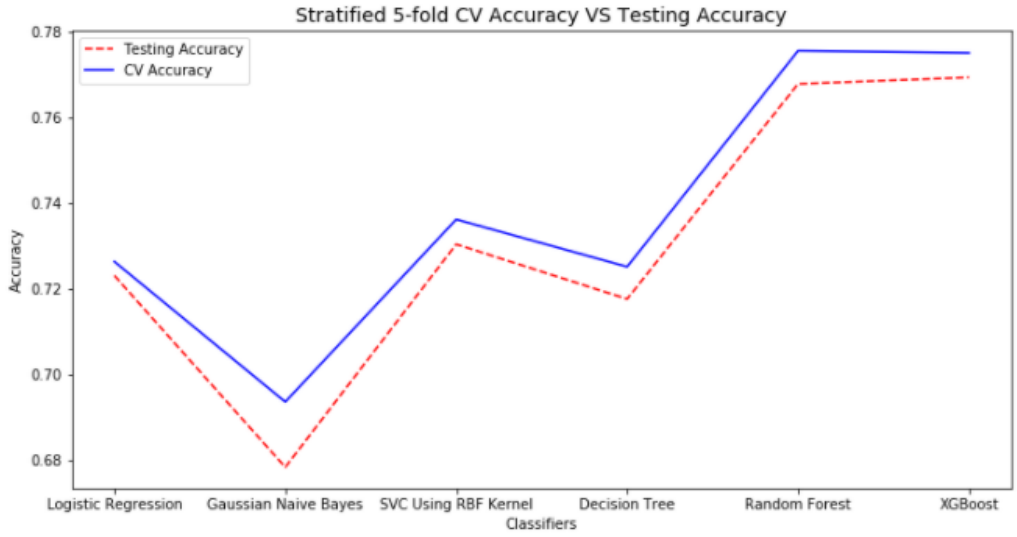


After training the models, their testing accuracies are compared. The result shows that Gaussian Naive Bayes has the lowest testing accuracy 0.68. The testing accuracy of Decision Tree, Logistic Regression, and SVC using RBF kernel are 0.72, 0.72, 0.73, respectively. Random Forest and XGBoost have a higher testing accuracy of 0.77.

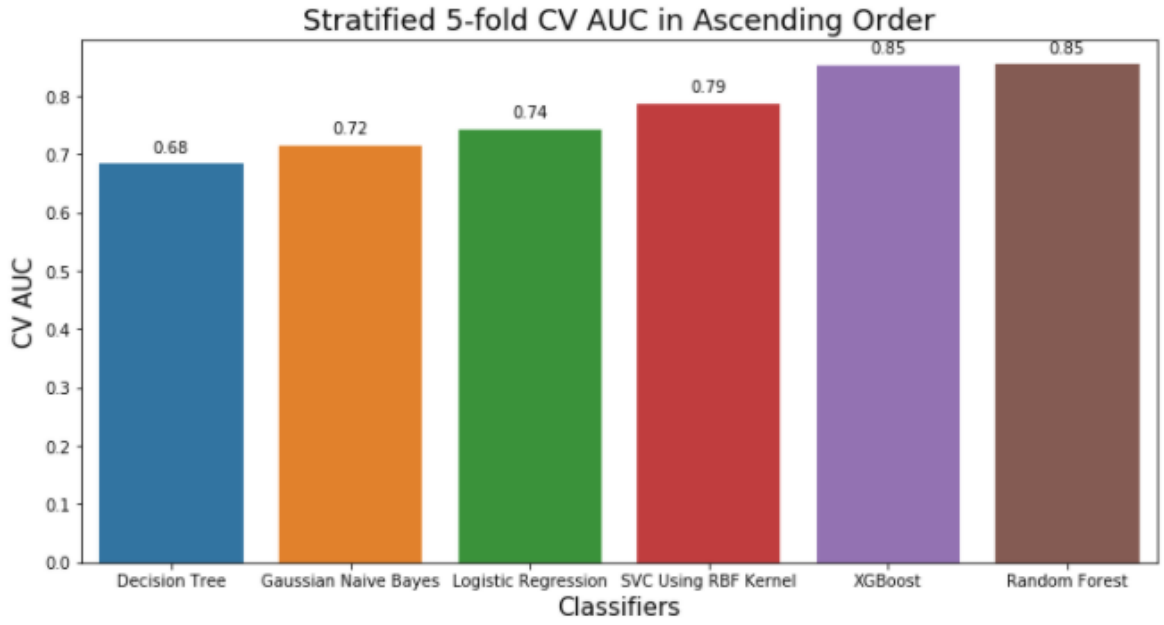


Besides testing accuracy, their cross-validation accuracies are compared. As we have a binary classification problem, and to make sure that the target variable is approximately identically distributed in each fold, stratified 5-fold cross-validation is used for this task.

The experiment results indicate that Gaussian Naive Bayes has the lowest cross-validation accuracy of 0.69. The cross-validation accuracy of Decision Tree, Logistic Regression, and SVC using RBF kernel are 0.73, 0.73, 0.74, respectively. Random Forest and XGBoost have a higher cross-validation accuracy of 0.78.



Comparing the testing accuracy and cross-validation accuracy, we find that the validation accuracy is slightly higher than the testing one. The difference is around 1 %.



Using the cross-validation AUC metric, the results show that Decision Tree has the lowest cross-validation AUC score of 0.68. AUC score of Gaussian Naive Bayes, Logistic Regression, and SVM using RBF kernel are 0.72, 0.74, 0.79, respectively. Random Forest and XGBoost have a higher cross-validation accuracy of 0.85. This means that both classifiers show good discrimination.

Conclusion:

The experiment results indicate that Gaussian Naive Bayes has the lowest cross-validation accuracy of 0.69. Random Forest and XGBoost have a higher cross-validation accuracy of 0.78. The results also show that the Decision Tree has the lowest cross-validation AUC score of 0.68. Random Forest and XGBoost have a higher cross-validation accuracy of 0.85. According to those results, the Random Forest, and XGBoost classifiers are more fitting for predicting the next day closing stock price trend.

References:

Stock Market Prediction Using Machine Learning Algorithms. (2019, July). Retrieved from https://www.ijrte.org/wp-content/uploads/papers/v8i2S4/B10520782S419.pdf

PDF) A new approach of stock price trend prediction based on logistic regression model. (2009, August). Retrieved from https://www.researchgate.net/publication/224595103_A_New_Approach_of_Stock_Price_Trend_Prediction_Based_on_Logistic_Regression_Model

Stock market prediction using logistic regression analysis -A pilot study. (2020, July). Retrieved from https://www.academia.edu/43787578/Stock_Market_Prediction_using_Logistic_Regression_Analysis_A_Pilot_Study

Brownlee, J. (2019, September 19). Machine learning mastery. Retrieved from https://machinelearningmastery.com