12/10/2024

Hiya Mavani

AAPL Stock Prediction with Random Forest

Data 101 Project #2

# Abstract

Stock prediction is a deeply researched topic, and there have been many attempts through technical and fundamental analyses to anticipate the direction of the stock market. In this paper, I will focus on classifying the direction (up or down) of stocks, focusing on the stock price of AAPL (Apple). AAPL is known for its stability in the markets, which makes it a beneficial choice in this study; Other more volatile stocks are known for throwing off algorithms because they are often driven by sentiment. The features in consideration are several different indicators, including oscillators such as RSI, MACD, Williams %R, ADX, and Bollinger Bands. For classification of the stock trend as going up or down, I will employ the ensemble method, Random Forest, which is one of the industry standards for classification in terms of accuracy and prevention of overfitting.

# Introduction

Many individuals aim to invest their money in stocks and take advantage of the seemingly quick, but also risky way of making money. The movement of stock prices are impacted by so many factors such as crowd sentiment, world news, macroeconomic movement, and the intrinsic value of a company’s assets. Stock analysis is divided into two main categories: fundamental and technical analysis. Fundamental analysis focuses on the broader picture; it places an emphasis on a company’s financial records and market capitalization, while technical analysis focuses on creating an almost scientific method of stock prediction that’s based on patterns, trends, and momentum of a stock’s price. Keeping track of all the indicators and signals dictated in the technical analysis of a stock can be complex and time-consuming to do by hand, which is why machine learning techniques can be used to aid this process. Techniques such as Random Forest can essentially capture all these indicators and signals as features and classify the stock price as going up or down. This kind of classification offers investors a buy or sell signal from the technical perspective of stock prediction, which they can verify with real-world sentiments to help make decisions about their investments.

# Features

The original datasets used contain timeseries data with features such as open, high, close, adjusted close, low, and volume. The open is the price that a stock opens at the start of a trading day. The close is the price that a stock ends with at the end of a trading day. The high is the highest trading price for the stock on a given day, and the low is the lowest trading price for the stock on a given day. The adjusted close factors in after-hours adjustments that may be made from shelling out dividends, and the volume is the number of transactions of a stock. The original dataset provides this information for each day ranging from 1980 to 2019. From these features, indicator features were calculated such as the RSI, MACD, ADX, William %R, Simple Moving Average, and Bollinger Bands. Random Forest uses an extension of bagging to reduce overfitting, and each observation is treated independently, which can be inadequate for time-series. A consequence of this is if the dataset contains features such as open, high, close, adjusted close, low, and volume, the model may show a naïvely high accuracy because the Random Forest cheats by having access to values that it wouldn’t have at the time of prediction. This is known as look-ahead bias, and therefore, this classification will only utilize features that rely on historical data prior to the date of the current close price — the indicator features listed previously. These features were also picked because of their popularity in investment strategies and their ability to be implemented and interpreted easily.

## Simple Moving Average & Bollinger Bands

The Simple Moving Average (SMA) is calculated with the formula (n is typically 14):

Where pi indicates the close-price on the i-th day of the n-day period.

This indicates the averaged price movement of a stock and it’s particularly useful as a measure of central tendency. For instance, with the SMA, a Bollinger band can be created to mark out the approximate trading range of a stock’s prices. By taking 2 standard deviations above and below the SMA, a region of price variety is created, and a stock price that travels outside of the region – above the upper boundary of the Bollinger band or below the lower boundary of the Bollinger band – indicates an unusual fluctuation that can signal that it’s due for a correction that brings the price back up or down.

## RSI

The RSI or the relative strength index is an oscillator that is calculated with the formula:

The average gain and loss are calculated over an n-day period, where n is usually 14 days.

The concept behind the RSI is that every stock has a trading range, and when a stock price reaches the bottom of its range, there will be a support force that pushes the stock price back. Contrarily, when the stock price reaches the top of its range, there’s a resistance force that pulls the stock price back down. The resistance level and support level of a stock’s trading range are measured by an RSI index of either 70 and 30 or 80 and 20 respectively (70/30 for this report). If a stock price touches the 70/80 mark, it indicates that the stock is overbought, and market forces will naturally bring the stock back down (trend reversal). If a stock price touches the 30/20 mark, it indicates that the stock is oversold, and market forces will naturally bring the stock back up. In other words, the RSI measures the strength of a stock price’s momentum.

A graph of a graph

Description automatically generated with medium confidence

Using the AAPL dataset, here is a graph of the RSI compared to the stock close prices for a subset of about 100 days of timeseries data (The closing prices have been scaled down to observe the patterns better). The RSI in March reaches the lower strength index of 30 and you can see the trend reversal to a brief uptrend in the close prices.

Note however, that the RSI can give a false alarm when the existing trend is strong. For example, the RSI reaches the resistance bar (upper bar) in mid-December, but this is following an already strong uptrend. Instead of the price going down after hitting the resistance bar, it continues to go up resulting in a breakout. This emphasizes the importance of verifying multiple signals from different indicators to make a clearer prediction.

## MACD

MACD or the Moving Average Convergence Divergence Oscillator is another popular indicator used in technical analysis of the stock market. Essentially, it measures the momentum (the speed and direction) of price movements. The MACD indicator has two parts: the MACD and the signal.

The MACD value is calculated by subtracting the 12-Period and 26-Period EMA (exponential moving average) of the historical close prices. The former is a short-term moving average, and the latter is a long-term moving average, so the difference of the two provides more insight on the direction of the market. The exponential moving average is like the simple moving average, but it places more weight on the more recent data points in the timeseries data.

MACD = EMA (12) – EMA (26)

A math equations with numbers

Description automatically generated with medium confidenceSource: Investopedia

The signal value is the 9-Period simple-moving average of the **MACD** line, and it’s used as a standard of comparison for the MACD.

Signal = SMA(9)

*9-day period*

The main points of interest with the MACD indicator are crossover points, where the MACD and Signal lines cross over each other, as well as the magnitude of the momentum at those points. Generally, when the MACD line is above the Signal line, this is a buy signal, and when the MACD line is below the Signal line, this is a sell signal.

Each graph below highlights 2 cross-over points. The first point shows that the MACD line surpasses the Signal line indicating a buy signal, and the subsequent price changes are shown in the close price graph, where the buy signal precedes a bullish rise in the close-price, ands where the sell signal from MACD foreshadows a huge price drop in mid-February.

A graph showing the growth of the stock market

Description automatically generated

A graph showing the growth of a stock market

Description automatically generated

## ADX, DMI+, DMI-, DI

Some indicators that are used in conjunction with MACD are the average directional index (ADX), directional movement index plus and directional movement index minus. ADX identifies the existence of a stock trend, however it only provides the magnitude of the strength of a trend and not the direction. The directional movement (positive line DMI+) and directional movement (negative line DMI-) provide information about the direction of the trend. The overall directional index (DI) is calculated by taking the difference between the DMI+ and DMI-. When the DMI+ is above DMI-, the price is moving up. Conversely, when the DMI+ is below the DMI-, the price is moving down.

The ADX can be evaluated with this table.

|  |  |
| --- | --- |
| ADX Value | Trend Strength |
| 0-20 | Absent or Weak Trend |
| 25-50 | Strong Trend |
| 50-75 | Very Strong Trend |
| 75-100 | Extremely Strong Trend |

For example, a positive directional index (DMI+ > DMI-) and an ADX value of 45 indicates a strong uptrend. A negative directional index (DMI+ < DMI-) and an ADX value of 15 indicates a weak downtrend.

In this stock trend classification, the ADX and directional index are represented in one feature by multiplying the ADX by -1 if the directional index is negative (the DMI+ < DMI-); the ADX is kept positive if the directional index is positive (the DMI+ > DMI-).

In the circled area of both graphs, the ADX is above the 0-baseline indicating that the directional index is positive and the ADX value is 30, which indicates a strong upward momentum. This is also reflected in the close price where there’s a strong upward movement in the price from December to Mid-February.

A graph showing the number of days and months

Description automatically generated



A graph showing the growth of the stock market

Description automatically generated



## Williams %R

The William Percent R indicator shows how the current price relates to the highest high over the last 14 periods. It’s calculated with the following formula:

The highest high is the high over the last 14 days and lowest low is the low over the last 14 days. The traditional representation of Williams %R has a range of 0 to -100, where -20 is the overbought line and -80 is the oversold line. The R library used in this report uses an alternative representation where the values of Williams %R are positive, but the graph has a reversed y-axis where 0 is at the top and 100 is at the bottom.

Any value below 20 indicates that the corresponding price is overbought, and crossing the y = 20 line indicates a sell signal. Any value above 80 indicates that the corresponding price is oversold, and crossing the y = 80 line indicates a buy signal.

The Williams Percent R indicator can be seen as a non-smooth version of the fast stochastic oscillator that is known for identifying overbought and oversold conditions in the market. However, potential limitations include its fast responsiveness resulting in false signals, which is why it’s used with the MACD and RSI indicators to filter out false alarms.

A graph with red and blue lines

Description automatically generated

A graph showing the growth of the stock market

Description automatically generated

The highlighted points on the Williams %R graph and Close Price graph show that around late February, the Williams %R indicates a sell signal and oversold conditions as the line is above the 20-line and crosses the 20-line. Correspondingly, the close price graph shows a drop in the price that continues until late March. It’s also evident that the Williams %R is sensitive to market changes due to the volatility of the graph, and there are also many false signals present.

# Materials

1. The dataset used was a Kaggle Stock Dataset with a multitude of stocks and ETFs that contained the features: date, open, high, low, close, adjusted close, and volume.
   1. The AAPL dataset initially contained 9,909 observations and 7 columns.
   2. After data wrangling and clean up, this was reduced to 9,876 observations.
2. The Integrated Development Environment was R Studio and the code for the Random Forest Classification was written in R.
3. The 2 classes considered in classification in the direction column:
   1. 1 for an uptrend (the stock price will go up)
   2. 0 for a downtrend (the stock price will go down)
4. The final features/predictors used for the final classification dataset:
   1. SMA/Bollinger Bands
   2. RSI
   3. MACD
   4. MACD Signal
   5. ADX & DI
   6. William %R
5. Indicator Calculation Library TTR
6. Investopedia Technical Analysis Blog

# Methods

### Data Plotting

Data plotting was used to analyze the AAPL data in relation to all the features to examine how volatile the AAPL stock prices were and how relevantly the features captured patterns in the stock. As provided in the *Features* section, the different indicator features were plotted along side the close price data to examine how they worked. This was done primarily using line graphs.

A correlation matrix was conducted to see correlations between the features. In this case, correlations were expected because of how some of the indicators are calculated, but Random Forest is skilled at handling multicollinearity.

Class imbalances were also examined at the start to ensure that the model wouldn’t predict using the most frequent class. The proportionality of the two classes will be provided in the *Results* section.

### Random Forest Classification

1. **What is Random Forest?**

Random Forests is a machine-learning technique that can be used for regression or classification, and it is an ensemble method that uses multiple decision trees to come to a predictive conclusion. For regression, the prediction of each tree is averaged to come up with the final classification. Relevantly, for classification, this conclusion is reached through a majority-wins rule, where if most trees classify the data point as a 1, the final classification is a 1. This technique also addresses many deficiencies of standard decision trees by ensuring that each tree is independent of the others, which helps to reduce overfitting of the data.

1. **Bootstrapping & Out-Of-Bag Error Analysis** 
   1. To ensure that each tree is developed independently, Random Forests uses bootstrapping, and the technique itself is an extension of bagging. Each tree resamples the original dataset with replacement to create new samples that are the size of the original dataset. Each of these resampled datasets are used per tree, which promotes independence.
   2. The Out-Of-Bag Error (OOB Error) allows Random Forests to perform its own validation while training the model. Approximately, each resampled dataset will contain 67% of the original data, so about 33% of the original dataset hasn’t been seen by each tree. This allows for that 33% of “out-of-bag” datapoints to be used as a validation set for each tree. The OOB error represents how many of the out-of-bag points were misclassified. This is important because it opens the door for cross-validation of tuning parameters based on the OOB Error.
   3. The OOB Error can also be compared with the error for the testing dataset. If the two errors are relatively close, the model isn’t overfitting on the training dataset, but if the errors aren’t close, this could indicate an imbalance between the training and testing datasets or indicate overfitting.
2. **Tuning Parameters (Number of Trees, Feature Subsetting)**
   1. When training the Random Forest model, one must provide parameters such as the number of trees for the model and how many features to subset for each tree. For the latter point, another factor of independence in Random Forest is that not all the predictors are used for each tree. Instead, for each tree, the root of the predictors is randomly taken from the original p predictors. Each level of the tree will only be able to choose the best predictor from the random sample of root p predictors. This reduces the correlation between decision trees. In this model, about 2 predictors are tried at each split.
3. Variable Importance
   1. The importance of each predictor to the accuracy of the model can be generated from the Random Forest Model. The predictors that contribute to the highest information gain and the greatest reduction in the Gini index (contributing to the most node purity) are the most important predictors.

### Feature Engineering

In this classification, feature engineering was used to reduce the number of features and increase their relevancy to the model.

For example, the original features in the dataset such as the open, low, close, and high prices were all used to calculate the indicators to avoid look-ahead bias and provide more meaning within the features.

Another linear combination was with the ADX and DI. The direction of the momentum’s strength was indicated by the directional index DI. If the DI was negative (uptrend), the ADX was multiplied by -1, and when the direction was in the positive direction (uptrend), the ADX remained positive. This combination allowed the ADX feature to mean more in terms of both direction and magnitude.

# Results & Discussion

### Data Plotting

**Line Plots**

The explanations of the below graphs are in the *Features* Section.

A graph showing the price of a stock market

Description automatically generated

A graph showing the number of months and months

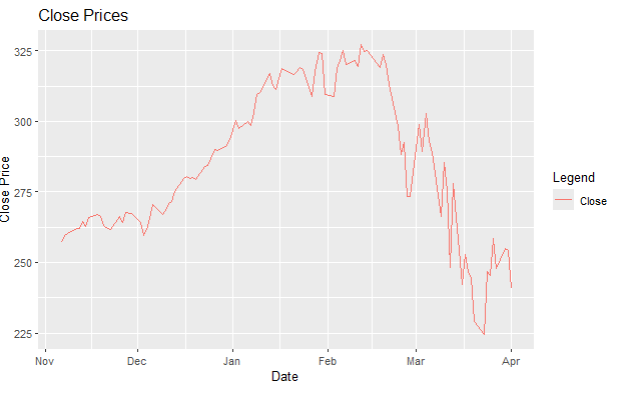
Description automatically generated

A graph with lines and numbers

Description automatically generated with medium confidence

A graph with red line and green line

Description automatically generated



A graph with red and blue lines

Description automatically generated

**Correlation Matrix**

A diagram of a stock feature

Description automatically generated with medium confidence

The correlation matrix shows high inverse correlations between Williams %R and the RSI. This is reasonable because both indicators use the high, low and close values in their calculations. They’re also calculated in a 14-day period. Trend strength is also highly correlated with RSI and inversely correlated with Williams %R.

**Training/Test Dataset Class Proportionality Tables**

A number with numbers on it

Description automatically generated with medium confidence

The class balances for both training and testing set are about the same for the direction of the stock price, so shuffling of the data wasn’t necessary.

## Random Forest Model

**Training Results (OOB Error)**

A close-up of a white background

Description automatically generated

The Out-Of-Bag Error for the in-built validation of the Random Forest was about 30%, which compares well to the error of the model on the test data, which is about 32% (1 – Accuracy). This shows that there’s a low chance of overfitting occurring with the data. The accuracy of the model on the OOB points was about 70%, which indicates that the model is performing about 20% better than random but some refinement is needed. The results also show that 2 variables were tried at each split in the trees in the Random Forest.

**Test Data Results**

A screenshot of a computer

Description automatically generated

The results of the Random Forest on the test data indicate the following metrics:

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.6781 |
| Sensitivity | 0.7434 |
| Specificity | 0.6181 |
| PPV | 0.6418 |
| NPV | 0.7235 |

The positive class was 0 (the stock will go down).

The accuracy of the model on the test data was about 68% which is slightly worse than the 70% accuracy of the model on the OOB data. It performed about 16% better than random, which indicates that a lot of refinement is needed.

The sensitivity of 0.7434 indicates that the model correctly predicted that the stock would go down about 74% of the time.

The specificity of 0.6181 indicates that the model correctly predicted that the stock would go up above about 62% of the time which is significantly worse than its ability to predict that a stock will go down.

The PPV however shows that even though the model was able to correctly predict that the stock will go down more often than it will go up, there was a higher likelihood of false alarms. The model was only right about the stock going down about 64% of the time. Contrarily, the model’s NPV was better: Of the stocks predicted to go up, about 72% of them actually went up.

**Variable Importance**

A number and a number on a white background

Description automatically generated with medium confidence

The variable importance results indicate that the most important variables to the accuracy were Williams %R, and the MACD indicator. The variables that contributed to the most information gain were Williams %R, RSI, MACD, and trend strength (ADX). Essentially, the best predictor was Williams %R with respectively about 54.6% mean decrease in accuracy and 974.3 decrease in the Gini index.

# Conclusion

The Random Forest model performed fairly in terms of accuracy, but it needs more work to be a reliable form of investment advice. It’s ability to predict that the stock will go down was slightly better than its ability to predict that the stock will go up, but it’s precision for predicting the former was worse than the latter. The most significant predictors were the Williams %R and RSI, which indicate that these indicators could be more reliable than others in terms of technical analysis. Overall, the model was only trained with the AAPL stock data, so it’s important to consider tuning for volatile stocks and even seeing how the performance varies between different market sectors such as tech companies vs. healthcare companies. For future improvement, employing a smoothing function and focusing on making more meaningful features would likely improve the model’s performance.

# Literature Cited

areed1192. “Sigma\_coding\_youtube/Python/Python-Data-Science/Machine-Learning/Random-Forest/Resources - Research Papers/Predicting the Direction of Stock Market Prices Using Tree Based.pdf at Master · Areed1192/Sigma\_coding\_youtube.” *GitHub*, 2019, github.com/areed1192/sigma\_coding\_youtube/blob/master/python/python-data-science/machine-learning/random-forest/Resources%20-%20Research%20Papers/Predicting%20the%20Direction%20of%20Stock%20Market%20Prices%20Using%20Tree%20Based.pdf. Accessed 10 Dec. 2024.

Dev, Ajaykumar. “Data Science: Understanding Random Forest Machine Learning Model for Stock Price Prediction with Bajaj Finance Stock Data.” *Medium*, 22 June 2024, medium.com/@nikaljeajay36/data-science-understanding-random-forest-machine-learning-model-for-stock-price-prediction-with-88d1b2f93047.

Engel, Claudia A. *Data Wrangling with R*. *Cengel.github.io*, cengel.github.io/R-data-wrangling/.

*Package “TTR” Type Package Title Technical Trading Rules*. 2023.

Seth, Shobhit. “Technical Analysis Strategies for Beginners.” *Investopedia*, 2019, www.investopedia.com/articles/active-trading/102914/technical-analysis-strategies-beginners.asp.

“Stock Market Dataset.” *Www.kaggle.com*, [www.kaggle.com/datasets/jacksoncrow/stock-market-dataset](http://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset).

\*ChatGPT was used for some R Code with the plotting, usage of TTR functions, and dplyr library.

# Appendix

The R Knit File is attached to the assignment. This file contains all the code and output for this assignment.