Course Project: Predicting the S&P500

Ori Rinaman

Bellevue University

**Executive Summary**

Recently signing up for a 401k, I became interested in learning more about the stock market. One of the lead indicators of the United States economy is the S&P500 index. The S&P500 index is an index that tracks the 500 largest companies in the United States. For my course project I proposed to attempt to build a model that would predict of the index would rise or fall in the next month based on previous months data.

The data I used to build the model included the DOW Jones Industrial Average, NASDAQ, Consumer Price Index (CPI), Gross Domestic Product (GDP), and the unemployment rate. All data was retrieved monthly. When it came to cleaning and aggregating the data, I started with creating my target variable called “Status” which is a 1 if the index fell in the next month and a 0 if the index rose in the next month. GDP and NASDAQ were also changed into percent change from previous month instead of the actual value of the indices.

Once the data was cleaned and aggregated to the form I wanted, I began to model using a decision tree algorithm and a random forest algorithm. I split the data into two data sets for modeling. One data set was the training data set that consisted of 75% of the data, and the other was the testing data set that contained the other 25% of the data which will be used to test the accuracy of the model. I created my decision tree model first and was able to get an accuracy of 64% against the testing data. The random forest model I created was only able to produce and accuracy of 61% against the testing data. The decision tree outperforming the random forest model surprised me do to the fact that a random forest is an ensemble model made up of many decision trees. My conclusion to the modeling is that simple sometimes is better. The results were not what I expected but I learned what I can improve upon on my next data science project.

**Introduction/Background of the Problem**

The United States economy has for almost a century been the front runner in world markets. A good indicator of the health of an economy is within the stock market. The stock market is the aggregation of buyers and sellers of stocks, which represent ownership claims on businesses. In the U.S. Economy there are three major indexes that track the markets. The three most widely followed indexes in the U.S. are the S&P500, Dow Jones Industrial Average, and the Nasdaq Composite. Understanding these indexes can give an understanding of the current state of the market. Now what if there was a way to predict how the market will perform in the future?

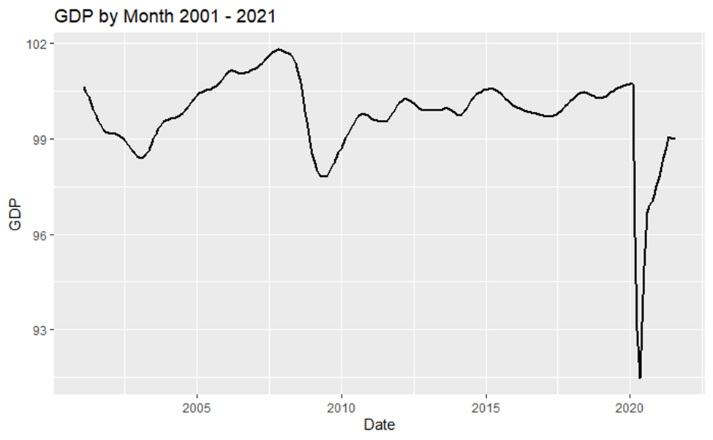
For years many people have tried and failed to beat the markets. People have dumped their whole life savings into the markets hoping that one day that investment will turn into a fortune. The S&P500 as previously mentioned is one of the major indexes that tracks the performance of the 500 largest companies in the United States. Companies of the index are selected by committee. When considering the eligibility of a new addition to the index, the committee assesses the companies merit using eight criteria: Market Capitalization, Liquidity, Domicile, Public Float, Global Industry Classification Standard and Representation of the industries in the United States, Financial Viability, Length of time publicly traded, and stock exchange currently on. As of December 2020, more than $4.6 Trillion was invested in assets ties to the performance of the index. The S&P500 index is a free-float weighted/capitalization-weighted index and the 10 largest companies in the index account for 26.4% of the market capitalization of the index. Currently the 10 largest companies on the S&P500 index are Apple Inc, Microsoft, Alphabet Inc (A and C)., Amazon.com, Facebook, Tesla, Berkshire Hathaway, JPMorgan Chase & Co., and Johnson & Johnson.

For my course project I proposed to attempt to create a predictive model to predict if the S&P500 index will rise or fall monthly. By having the ability to know what the index will do in the future month there is opportunity to increase value in investments and more over attempt to beat the markets.

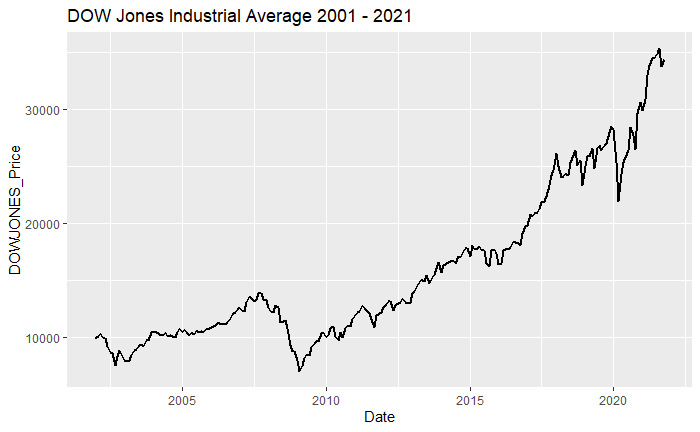
**Methods**

This project started with many hours of research trying to gather ideas for what data could be used to develop a predictive model for the S&P500 index. Knowing that the S&P500 index was solely based off the 500 largest companies in the U.S. I knew that the data I needed had to indicators of what economists use to try and understand the current state of the economy. The indicators that I found for my model were Gross Domestic Product (GDP), Dow Jones Industrial Average (DOW), National Association of Securities Dealers Automated Quotations (NASDAQ), Unemployment Rate, and the Consumer Price Index (CPI). I collected data for the indicators going back to January of 2001.

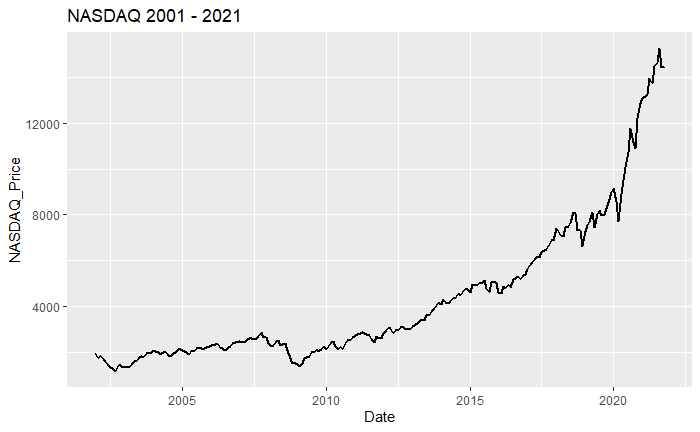
**Gross Domestic Product (GDP)** – monetary measure of the market value of all the final goods and services produced



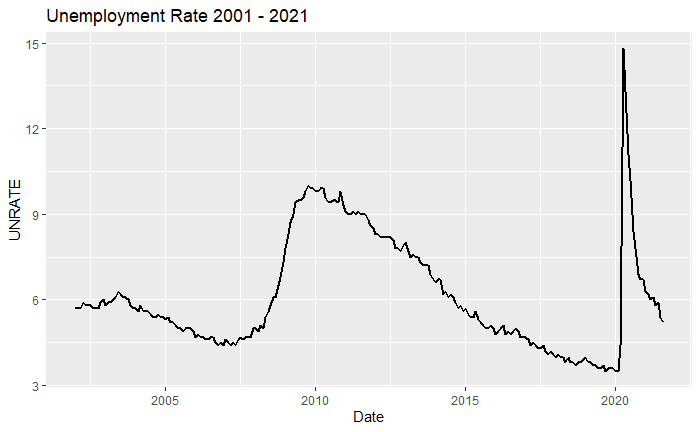
**Dow Jones Industrial Average (DOW)** – a price weighted measurement stock market index of 30 prominent companies listed on various stock exchanges in the United States.



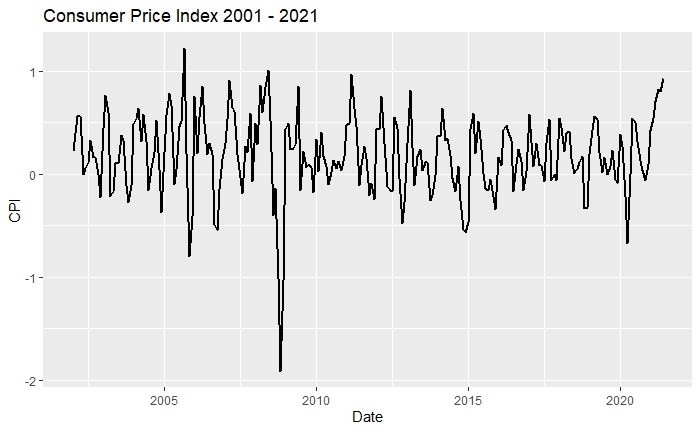
**National Association of Dealers Automated Quotations (NASDAQ)-** the worlds electronic quote system that provides price quotes and supports trading for over-the-counter stocks. Heavily weighted towards companies in the information/technology sector.



**Unemployment Rate**- measures the share of workers in the labor force who do not currently have a job but are actively looking for work. People who have not looked for work in the past four weeks are not included in this measure.



**Consumer Price Index (CPI)**- the price of a weighted average market basket of consumer goods and services purchased by households.



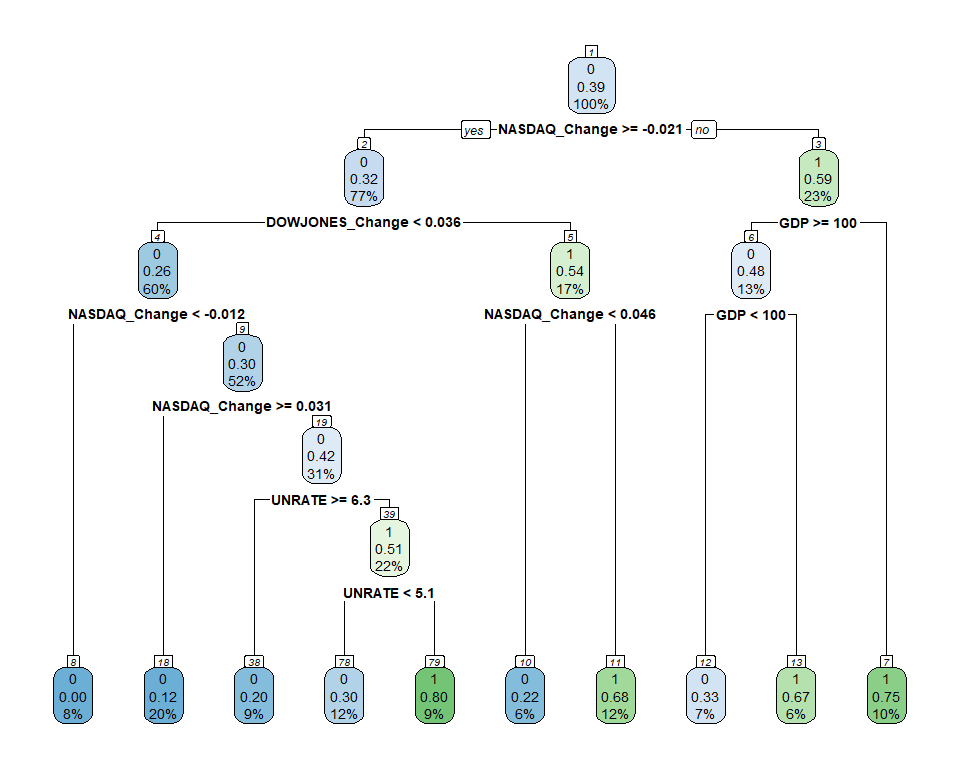
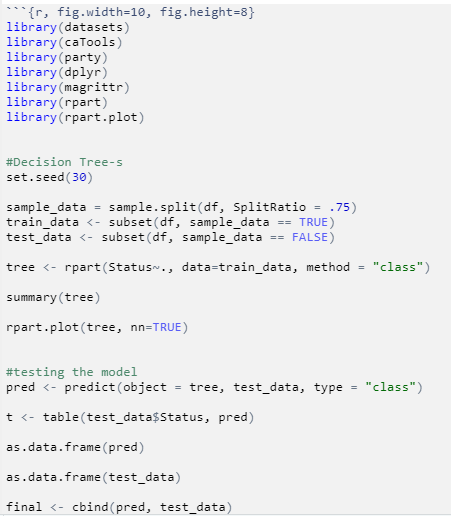
After getting data for all the indicators it was time to prepare the data for modeling. To clean and prepare my data I used the R programing language. R is a programming language and software environment for statistical computing and graphics. Using R I was able to join all data tables containing all of the indicator variables for my model. Next, I created a new variable called “Status” which would indicate a 0 if the S&P500 index rose from the previous month or a 1 if the index fell from the previous month. Then to get the data to line up to predict the next month I shifted the Status variable up one month. At this point my variables were all in sync with the date and my target variable Status would be shown as the next month’s value. After the data finished being prepared for modeling it was time to decide which technique would be used to build the predictive model.

The first attempt that was made to create the predictive model was a decision tree. A decision tree is a type of supervised machine learning where the data is continuously split according to a certain parameter. It is a flow like decision making tool where each node represents a test on a feature, each leaf node represents a class label, and branches represent conjunctions of features that lead to those class labels. The reasoning for choosing a decision tree for the first attempt is because it is a simple algorithm that can be easily explained and still can yield good results.

The second attempt at modeling used a random forest model. A random forest model is also a supervised machine learning algorithm that is constructed from multiple decision tree algorithms. Random forest is known as an ensemble model which is a machine learning technique that combines several base models to predict the optimal predictive model.

**Results**

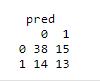
After the data was aggregated and formatted, I began my attempt at modeling with the decision tree algorithm. I began by separating the data into two datasets. The first data set was my testing set which consisted of 75% of the data. This data will be used to train my decision tree model. The second data set was the test data set which will be used to test the accuracy of the model. Once again, I used R to create the decision tree model and below is the resulting decision tree I created.



R code for Decision Tree

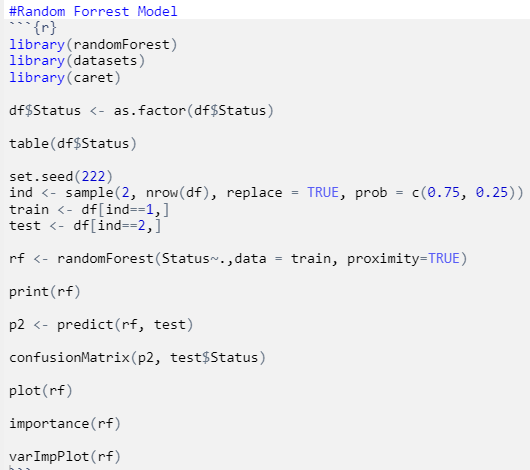
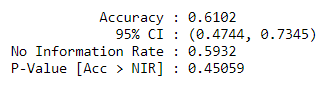
Decision Tree

The decision tree above was able to produce and accuracy of 64% against the testing data. Below is the confusion matrix of the results running the model against my testing data set.



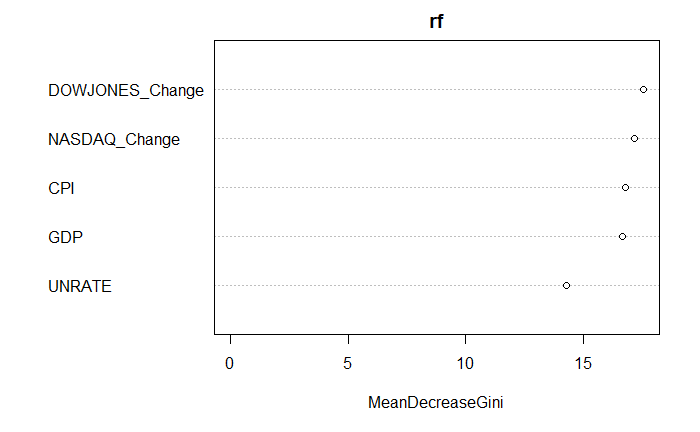
Confusion Matrix

Trying to improve upon my results from the decision tree I next went and attempted to model using a random forest technique. Similar to the decision tree model, I used a training data set that consisted of 75% of the data and a testing data set that consisted of the other 25% of the data. I originally assumed that a random forest model would produce better results than the decision tree due to the fact that a random forest model is made up of multiple decision trees. To my surprised I was wrong and was only able to get an accuracy of 61% using the random forest model. Below is the R code I used to create the random forest model as well as the results of putting the model against the training data and the variable importance plot.



Output of model vs. predicted data set

R code for Random Forest Model

****

Variable Importance Plot

The results were not as desirable as I would have expected, but the results are decent considering the difficulty of trying to predict anything related to the stock market due to its unexpected variability. Hypothetically if I had to choose a model, I would go with the decision tree. It produces an accuracy of 64% and decision trees are easy to understand and deploy.

**Discussion/Conclusion**

Going into this final course project I was aware that the task of trying to model or predict anything related with the stock market was going to be extremely challenging and difficult, but I wanted to try it for myself. I will admit during the first few weeks of working on the project I was extremely confident that I could accomplish building a decent model to make an accurate prediction. In the later weeks when model development began, I ran my first few models with little to no success. The accuracy of the initial models was extremely low. With more research into better indicators of the S&P500 index I was able to get the five final variables for the modeling. With these variables the new models created were able to yield an accuracy of about 64%. The results were not bad, but also not great.

Some challenges that I faced when attempting this project were very unexpected. First, I found it very difficult to find data going back so many years. If I did find data, it was almost impossible to download it without paying a fee. Luckily, I was able to combine some free data from various websites to get my final data set. Another issue I ran into was trying to create a variable to predict. I had switched my initial target for prediction multiple times until I came up with trying to predict whether the S&P500 index would rise or fall in the next month. This took time to really understand what I wanted out of my model as well as how I was going to create the target variable for my supervised learning techniques. Lastly, I wish I would have had more time to investigate and learn more about time series analysis. With the little knowledge I have on the subject I feel that including this into my investigation and modeling could have produced a better model with higher accuracy. Looking forward into future works with this project I would recommend to myself to slow down and really understand the problem I am trying to solve. I found myself so excited to begin modeling that I would catch myself missing crucial steps in the data science process. All in all, this was a fun and interesting project. I am glad I decided to challenge myself with a difficult problem even though I failed to meet my goal of developing a model with a high enough accuracy for use. This was a great learning experience and I look forward to attempting more projects in the future.

**Acknowledgements**

I would like to start by thanking professor Alsaleem for a fun and enjoyable semester. I learned many new skills that will help me along with my journey of becoming a data scientist. I also would like to send thanks to all my classmates who peer reviewed my various milestones throughout the course. Their feedback was extremely valuable to assisting me in completing my project. I wish everyone luck in their future endeavors and hopefully I will see some of you in class next semester!

**References**

“Decision Trees in R: Examples & Code in R for Regression & Classification.” *DataCamp Community*, https://www.datacamp.com/community/tutorials/decision-trees-R.

Finnstats. “Random Forest in R: R-Bloggers.” *R*, 13 Apr. 2021, https://www.r-bloggers.com/2021/04/random-forest-in-r/.

Johnson, Daniel. “R Random Forest Tutorial with Example.” *Guru99*, 7 Oct. 2021, https://www.guru99.com/r-random-forest-tutorial.html.

Team, The Investopedia. “What Is the History of the S&P 500?” *Investopedia*, Investopedia, 27 Oct. 2021, https://www.investopedia.com/ask/answers/041015/what-history-sp-500.asp.