New York Stock Exchange

Project Report

Subject:
Business Analytics with R - BUAN 6356.003

Group Members:

- Yukung Cha
- Ismael Isak
- Snehal Khatpe
- Tashfia Mamun
- Naomi Snider

Table of Contents

Abstract

The New York Stock Exchange is an American stock exchange in the Financial District of Lower Manhattan in New York City. It is the world's largest stock exchange by market capitalization of its listed companies at US$30.1 trillion as of February 2018.The average daily trading value was approximately US$169 billion in 2013.

We analyzed the NYSE data from 2010 to 2016 to see what Machine Learning model can be implemented to predict the closing prices of a stock based on Numeric Data.

We cleaned our data to remove missing and null values, and we chose a stock's opening prices, highest price, lowest price and volume as our independent variables, to predict dependent variable "The closing price of stock"

We compared 5 ML models with the following results

- **Ridge Regression:** High MSE and a negative R square led us to conclude Ridge regression is not the best option for the data.
- **Lasso Regression:** We got decent MSE and R square, but for Lasso Regressions, R squared is not always the best way to check a model's fit, as there are not many covariates
- **Regression Tree:** Decent MSE and R square, so a decent choice for predicting this dataset.
- **Neural Network:** High MSE and negative R square indicates its highly unsuitable for this dataset.
- **Linear Regression:** Low MSE and R square close to 1 indicates that its great for predicting our dataset. Analyzing residuals lead us to conclude that residuals are scattered around center, meaning that it's a good fit for.

We conclude that for predicting stock's closing prices, the most accurate model is Linear Regression.

Proposed Work

The overall goal of this project is to better understand the New York Stock Exchange data from year 2010 to 2016 and data from 501 companies.

Broadly our objectives are:

- Predict the best predictors for predicting closing price of a stock
- Compare Machine Learning models to find out which is the best one for predicting closing price.

Summarizing and preparation of data

Data Summary

For the project, four data sets are provided: prices, split_prices, securities, and fundamentals.

- **Prices:**
  Prices is raw as-is daily prices which does not account for stock splits. Records time stamp, open, close, low, high, volume
- **Security:**
  Description of each company with division on sectors
- **Split Prices:**
  Split prices is the same as the prices data, but has been adjusted for stock splits
- **Fundamentals:**
- Metrics from SEC 10K filings

Descriptive Statistics

1. Closing Price for each industry by year

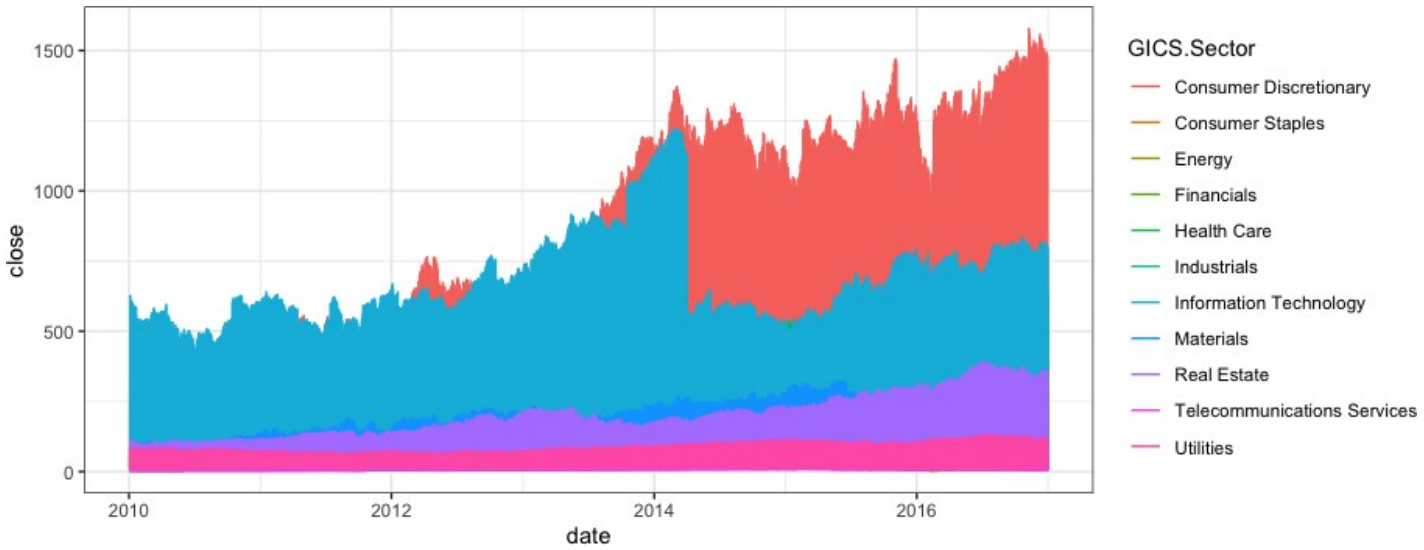


Figure 1 Closing price by industry per year

Observations:

- Biggest growth in consumer discretionary and health sector
- Growth slowed down for most industries in 2015

2. Analysis By industry

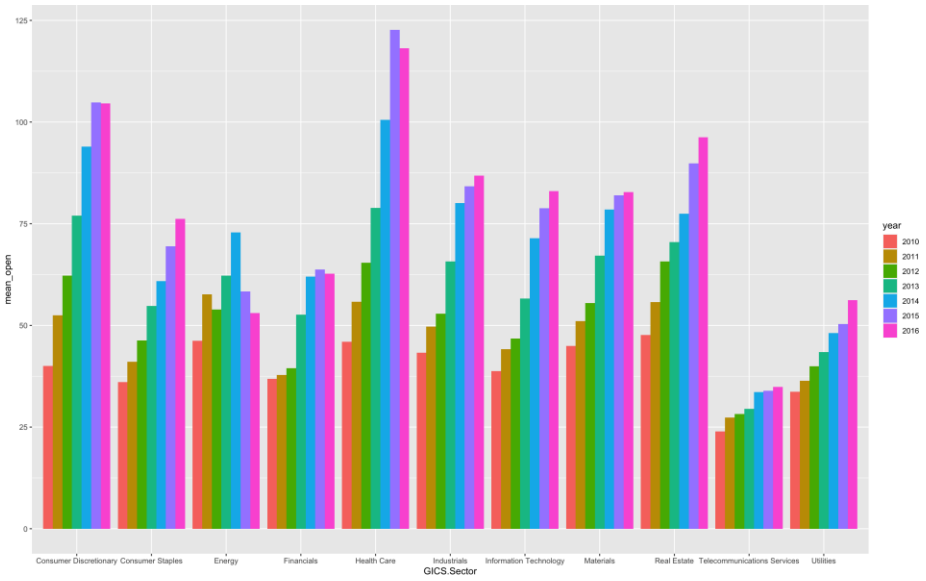


Figure 2 Mean opening prices each year grouped by sector

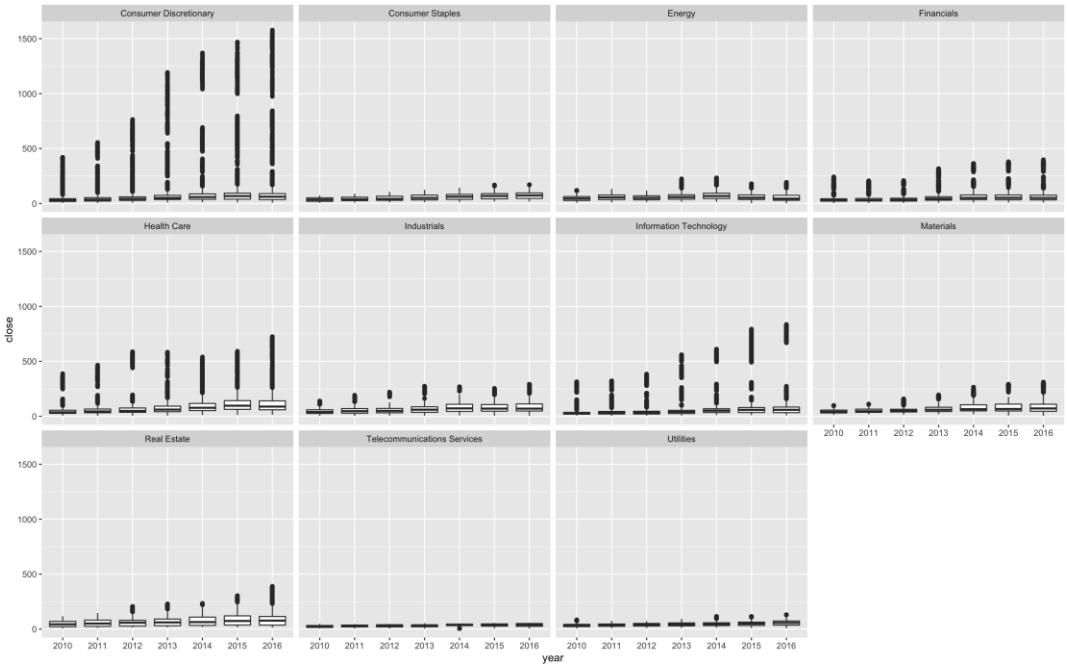


Figure 3 Boxplot of closing prices by industry by year

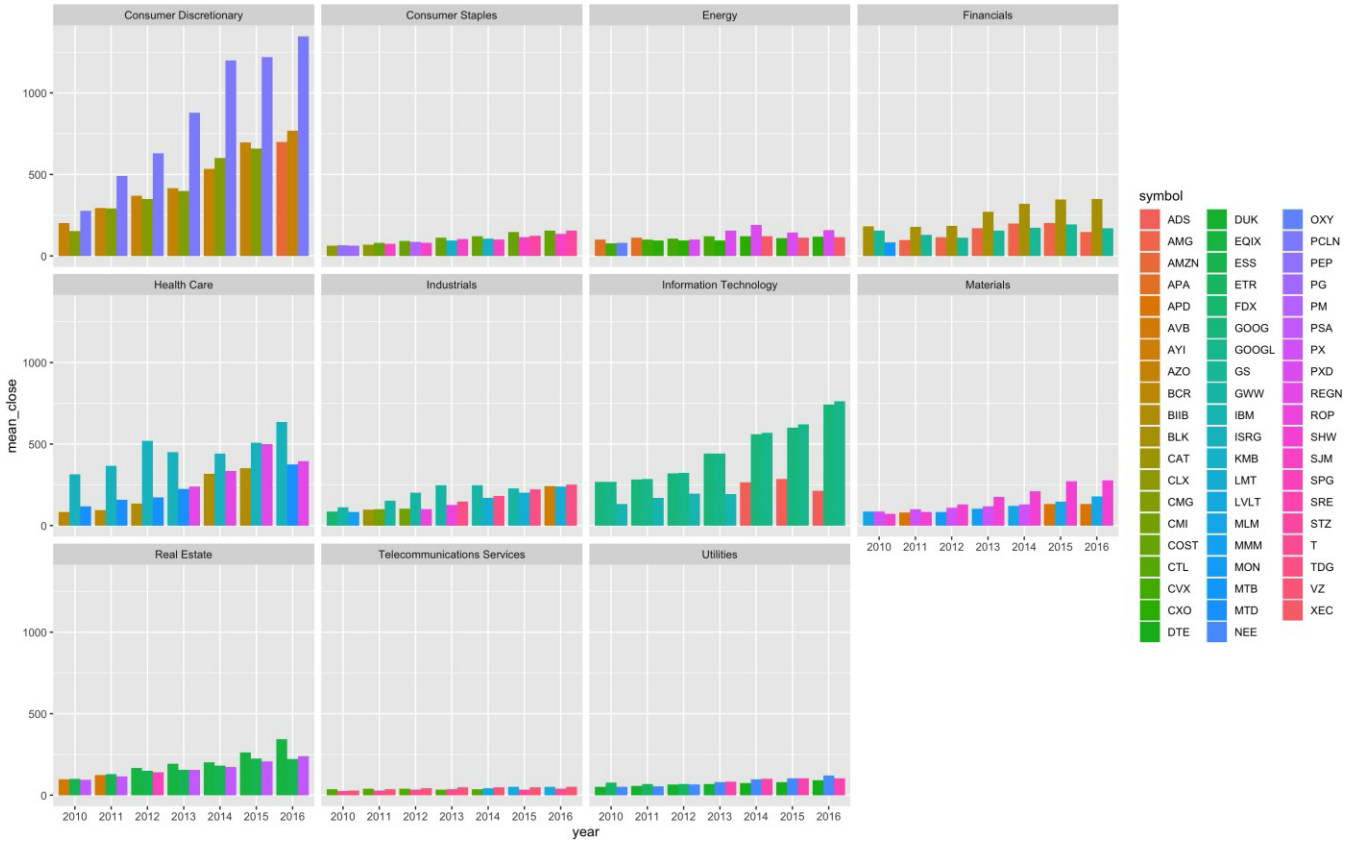


Figure 4 Top 3 companies for each year (based on closing price)

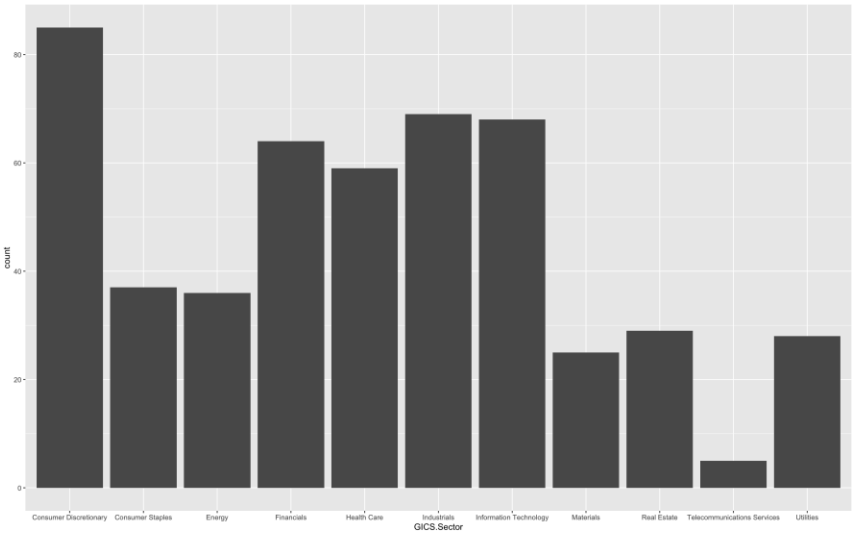


Figure 5 Histogram of sectors

Observations:

- Biggest growth in consumer discretionary and health sector
- Growth slowed down for most industries in 2015
- Top 3 companies have stayed mostly consistent
- Highest number of companies belong to the Consumer Discretionary industry

3. Analysis on Columns:

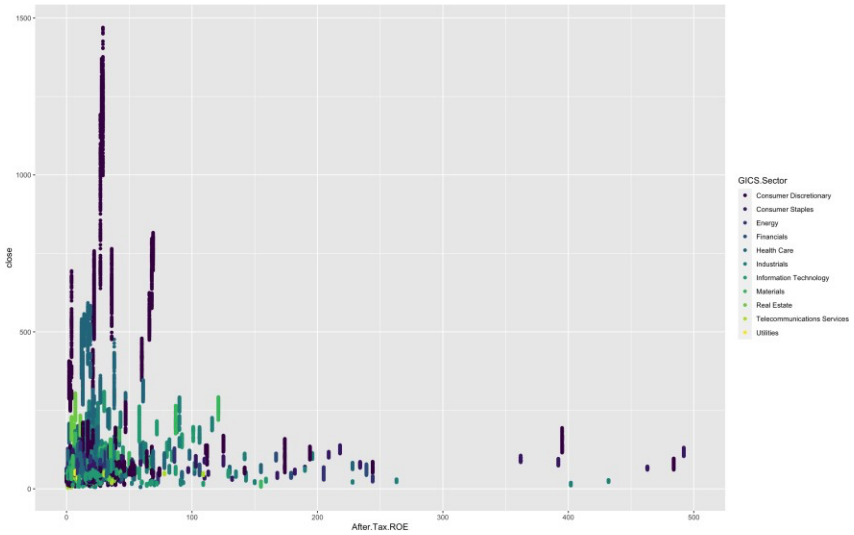


Figure 6 Closing Price vs ROE

Observations:

- No significant linear relationship
- Consumer discretionary sector tend to have high stock price and after tax ROE

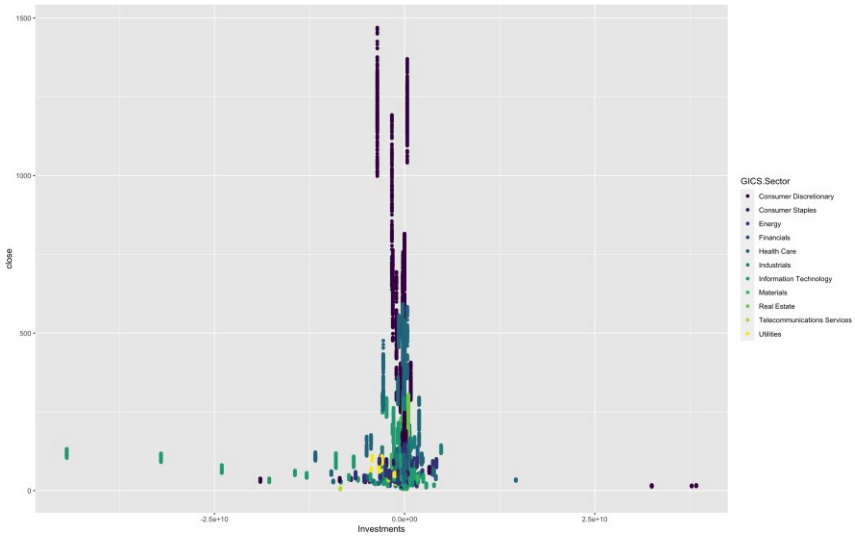


Figure 7 Closing Price vs Investment

Observations:

- No significant linear relationship
- Investment tended to be centered in the middle

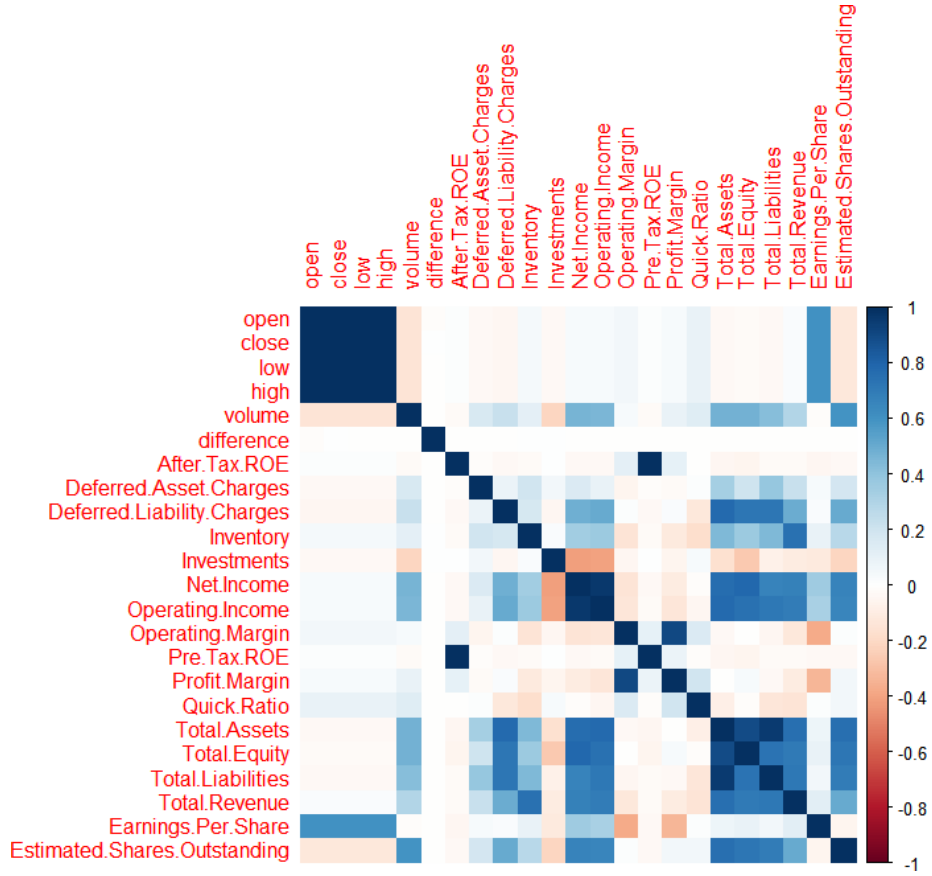


Figure 8 Cor Plot for every column

Observations:

- This Cor plot shows that the closing price has significant relationships with 'open', 'close', 'low', 'high', 'volume', and 'Estimated Shares. Outstanding' columns.

Preparing the Data

Data Cleaning:

- Data was cleaned to remove all missing and NA values.
- Missmap function was called to see if dataset had any gaps.

Model Preparation:

- Perform Feature Selection on data, then test multiple models to find the lowest Mean Square Error (MSE)

Feature Selection:

- To determine what features/variables to use in the model, forward and backward stepwise regressions were performed.
- Similar to the data visualization that was shown closing price had the strongest relationship with opening price, highest and lowest price, and stock volume, which is why these features were selected to predict closing price

Techniques Used/Models Tested:

Each of the following models were tested to find the best MSE

- Ridge Regression
- Lasso Regression
- Regression Tree
- Neural Network
- Linear Regression

Evaluating Algorithms

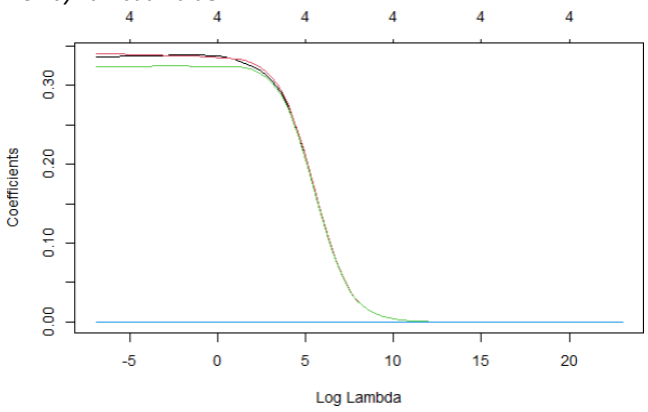
To determine which model functioned best for the stock exchange data we evaluated the following algorithms to notewhich one had favorable value for MSE and R square.

Ridge Regression

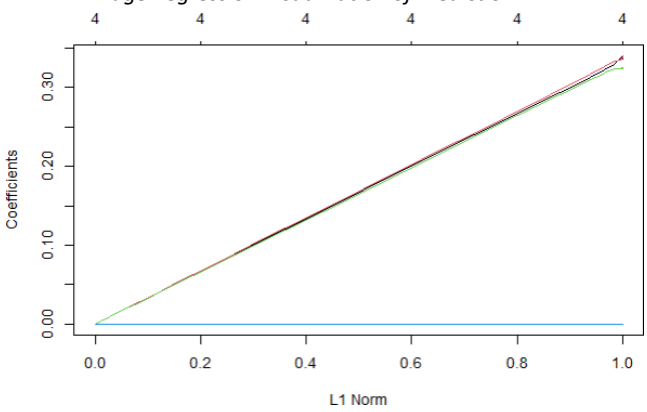
Ridge Regression is a penalized regression method, meaning that it regularizes the regression coefficients and as a result some of the coefficients are forced to shrink.

Lambda is used to calculate the penalty applied to the model. Best Lambda was calculated by running the ridge regression and taking the lambda that produced the lowest MSE.

MSE by Lambda Value



Ridge Regression Visualization of Prediction



To predict Closing stock value, we have taken numerical variables (Open, Low, High, Volume) as predictors. The following results were yielded using these variables:

- MSE: 366908.4
- R squared: -51.20822

Overall, the model was not a good fit for the data. With a negative R squared, the model is less accurate than the constant mean of the data.

Lasso Regression

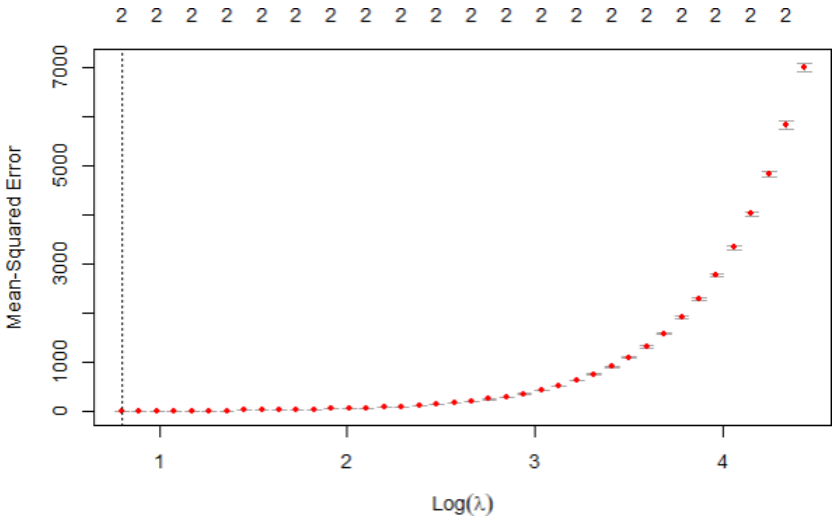
Like Ridge Regression, Lasso Regressions are a penalized regression method. Variables included were (Open, Low,High, Volume). The best lambda was 2.23, which produced an MSE of 7.05.

To predict Closing stock value, we have taken numerical variables (Open, Low, High, Volume) as predictors. The following results were yielded using these variables:

- MSE: 7.053108
- R squared: 0.9989964

Although the R value is high, the MSE is still not ideal. Additionally, R squared is not always the best indicator of a model's accuracy or fit, as there are not very many covariates and R squared tends to favor models with many covariates.

MSE by Lambda Value

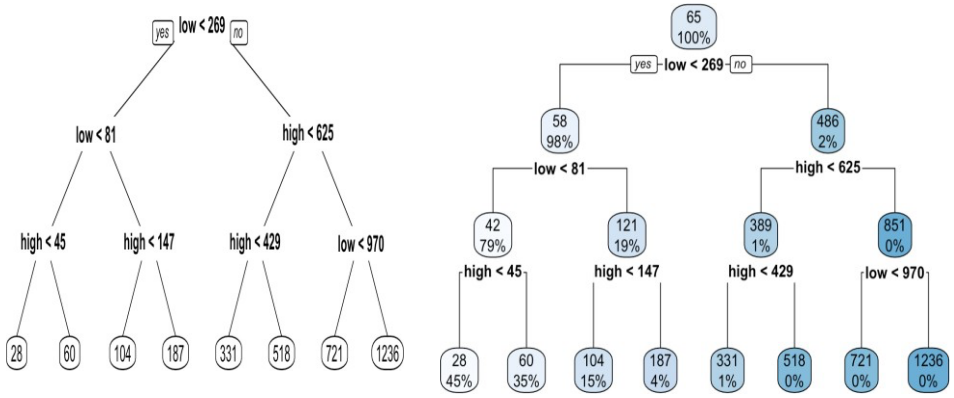


Decision Tree

A decision tree is a type of algorithm in machine learning that uses decisions as the features to represent the result in the form of a tree-like structure.

We have used Regression Algorithm; Regression trees are used when the dependent variable is continuous. In a regression tree, the values of the terminal nodes represent the mean values of the dependent variable in each subset of the sample. So, the predictions for the new observations are made using the mean values.

As part of the first step build a full tree and the variables used to construct the tree are high and low.

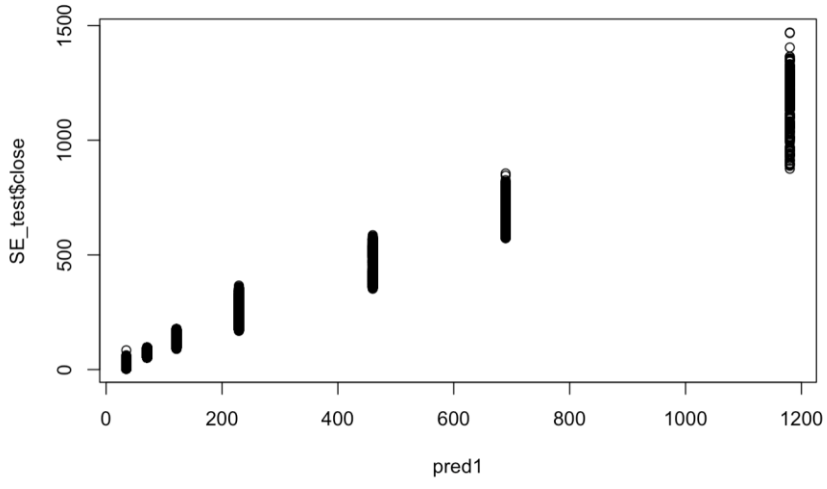


To build a regression tree we would require 'rpart' library. This library implements recursive partitioning and is very easy to use.

Modelled data with training dataset and we got:

- MSE: 122.8483
- R squared: 0.956

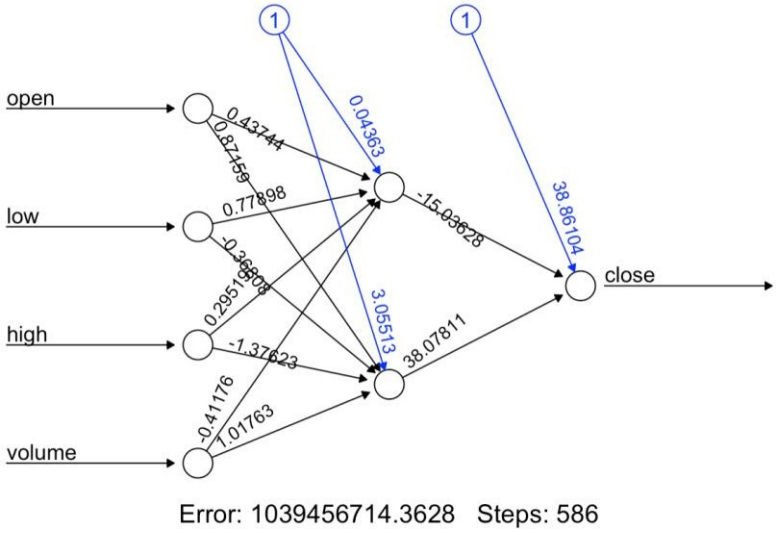
Plotting r**eal vs predicted values, we see that results are quite accurate, also indicative of our high R square value.**



Neural Network

Neural networks are powerful modeling tools that can detect complex non-linear relationships between inputs & outputs.

To predict Closing stock value, we have taken numerical variables(Open, Low, High, Volume) as our predictors. We have taken 2 hidden layers.

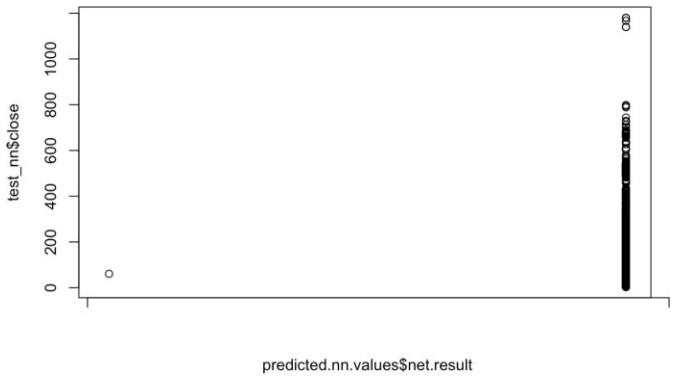


From our Neural Network, we found that:

- MSE: 2205.179
- R squared: -0.06

Negative R square implies model's predictions are worse than a constant function that always predicts the mean of the data.

Plotting the real and predicted value we can see that the model cannot accurately predict results, and all the predicted values are same for different predictors, making it an unsuitable model for us.



Linear Regression

Linear Regression is used to predict the value of a variable based on the value of other variables. Since our model should predict the stock price based on other features of the data, we decided to try utilizing Linear Regression.

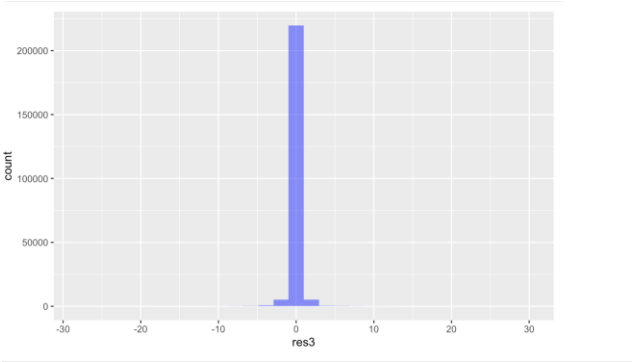
Selecting Variables

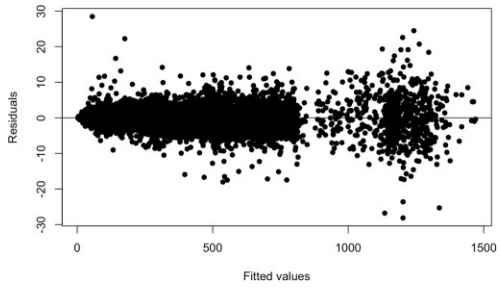
For the Linear Regression model, we decided to predict ‘close’ price with our numeric variables ‘open', 'low', 'high', 'volume', 'difference',

As a result, we got:

- MSE: 0.2199
- R squared: 0.99989

This result showed us that this model estimates the values very well and it fits the data very well. The residual plot showed that the values are randomly scattered.





Comment

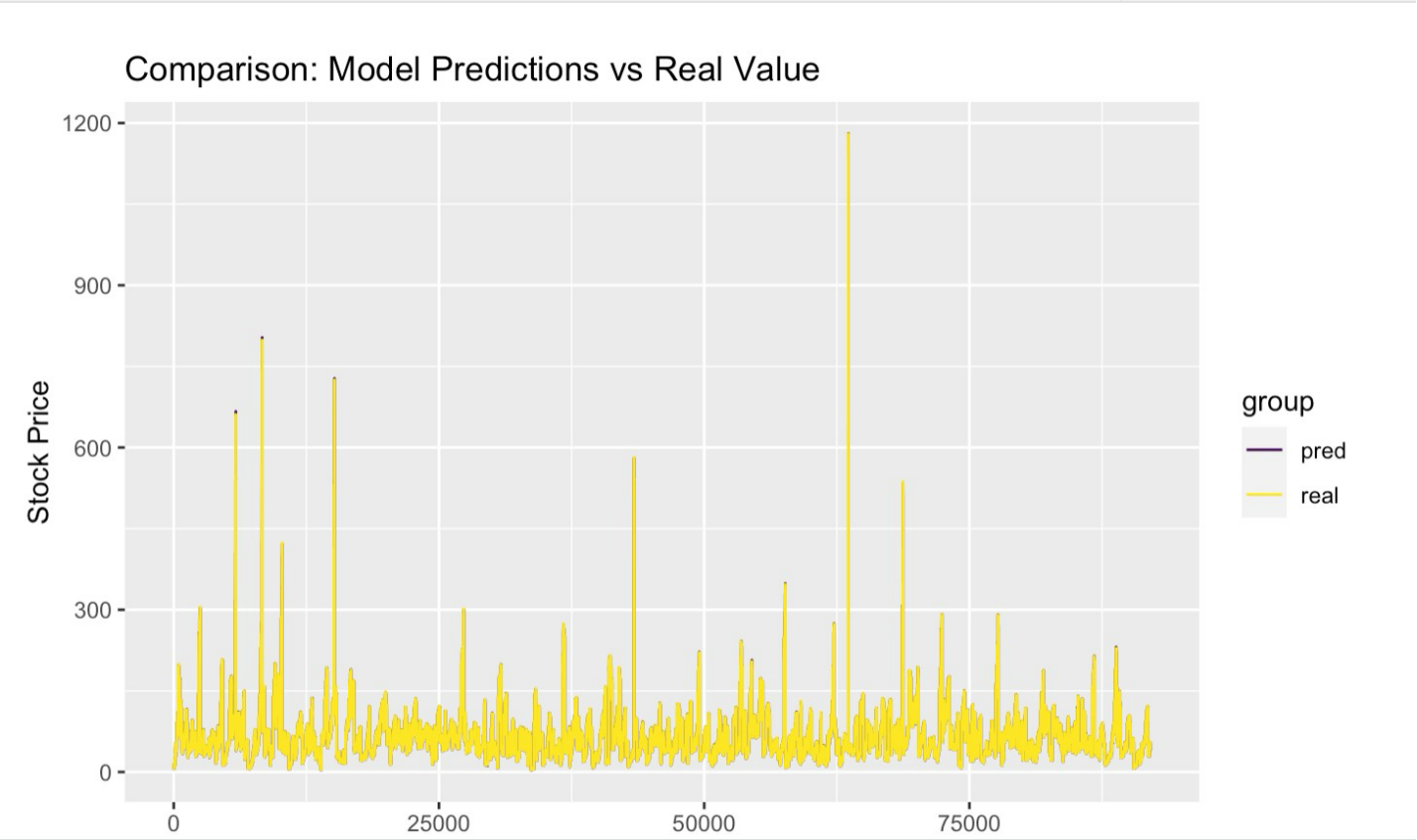
Low MSE and high R squared value does not always mean the model is the best fit. The scattered residual plot shows that its an accurate model.

Results

Overall, the linear regression was the best model to use for the New York Stock Exchange dataset as it had the lowest MSE and yielded the closest results to actual values.

Model	MSE	R Squared
Linear Regression	0.2199	0.99989
Neural Network	2205.179	-0.06
Decision Tree	122.84	0.956
Lasso Regression	7.053108	0.9989964
Ridge Regression	366908.4	-51.21

Comparing Real and predicted values with Linear Regression in this graph, we see the values are very close



Best Performing Stock:

To create a baseline comparison between all the stocks we have available we calculate the lifetime return percentage. Which is the last price of the stock subtracted from the original price divided by the last price. This means that regardless of day-to-day fluctuations in price we will have a stock that gives us the best return for our money over the long term.

For example, if a stock is originally bought for $19/share and we purchase 10 of those stocks that means we have spent $190. If the price of the stock the day, we decide to sell it is $60/share then that means we have made $600 - $190 = $410.

If instead we used that same $190 to buy a stock that went from $5 to $20, we would have made $570. Even though the final stock price is lower than the other stock.

That's because we are buying stock by the share and not by the dollar amount, so the bigger the difference between the stock prices at the beginning or end of our investment the bigger our profit.

Here is a table of our best performing stocks along with the first opening price, and final closing price with the percentage next to each.

Table 1: Best Performing Stocks

Stock	Opening.Price	Closing.Price	Lifetime.Return
NFLX	7.93	123.80	0.935933536055246
REGN	24.24	367.09	0.933967146301639
ULTA	19.23	254.94	0.924570487765196
URI	9.92	105.58	0.906042812918302
ALK	8.71	88.73	0.901893387741686
AVGO	18.30	176.77	0.896475654319723
AAL	4.84	46.69	0.89633754329273
STZ	16.02	153.31	0.895505836481715
CHTR	35.00	287.92	0.878438460615101
ORLY	39.24	278.41	0.859056781594673

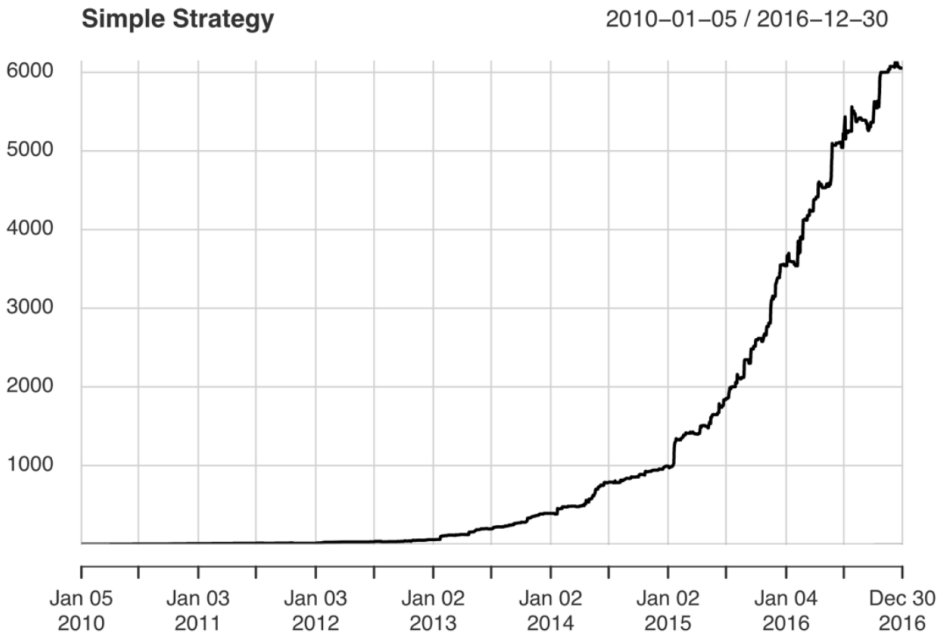
Trading Strategy:

Next, we can also perform a method of trading strategy where we only buy stocks based on the previous days change inprice.

We start with the first day and observe the closing price of our chosen stock, in this case NFLX, and if the opening price for the following day is higher than the previous closing price by 5% we buy the stock. By the end of the day, we then sellstock, where we then lock in whatever profits we made for that day and move on to the next.

Since we already know that we have chosen the best performing stock we can expect many days where our prices will be higher, and we will be trading often. This strategy is simplified but like computer trading algorithms that take many factors into account when executing a trade. When prices begin to slide lower some computers are designed to automatically begin selling shares, and when stock prices go higher, they begin buying hoping to gain a profit. Which arethen reinvested into the program to do it all again.

Here is the cumulative return with our simple strategy of buying and selling at the end of the day.



This would then compare to the cumulative returns for no strategy and buying selling everyday regardless of price change.

Discussion:
This shows us that even a simple strategy can improve the returns on our investment and how computer algorithms can trade so quickly and bring massive returns to their owners. This does come with a risk though and that includes a situation where the underlying stock is doing fine but because a program met a threshold and sold a large volume of stocks that could lead other programs to do the same and cause a rippling effect in the market. So carefully building the algorithms is just as important as monitoring them because computers sometimes lack the ability to understand context.