DATFUEL

Stock Price Trends Prediction Report

Group Members:

Shourya Paranjape
Zhe Wang
Ayushri Bhargava
NagaVineela Mellachervu
Yuxiang Gao

Table of Contents

I. Introduction

- Business Problem:

Datfuel is an asset management company that runs many mutual funds. The company makes profit by charging administration fee from the customers if they decide to invest in our mutual funds and also from a small part of the revenue that comes from the investment of our customers if they make a profit in the stock market. We as the technical team of the Datfuel company provides the information about the trend of stock prices to the managers of our fund management team in order to increase the net asset value of mutual funds.

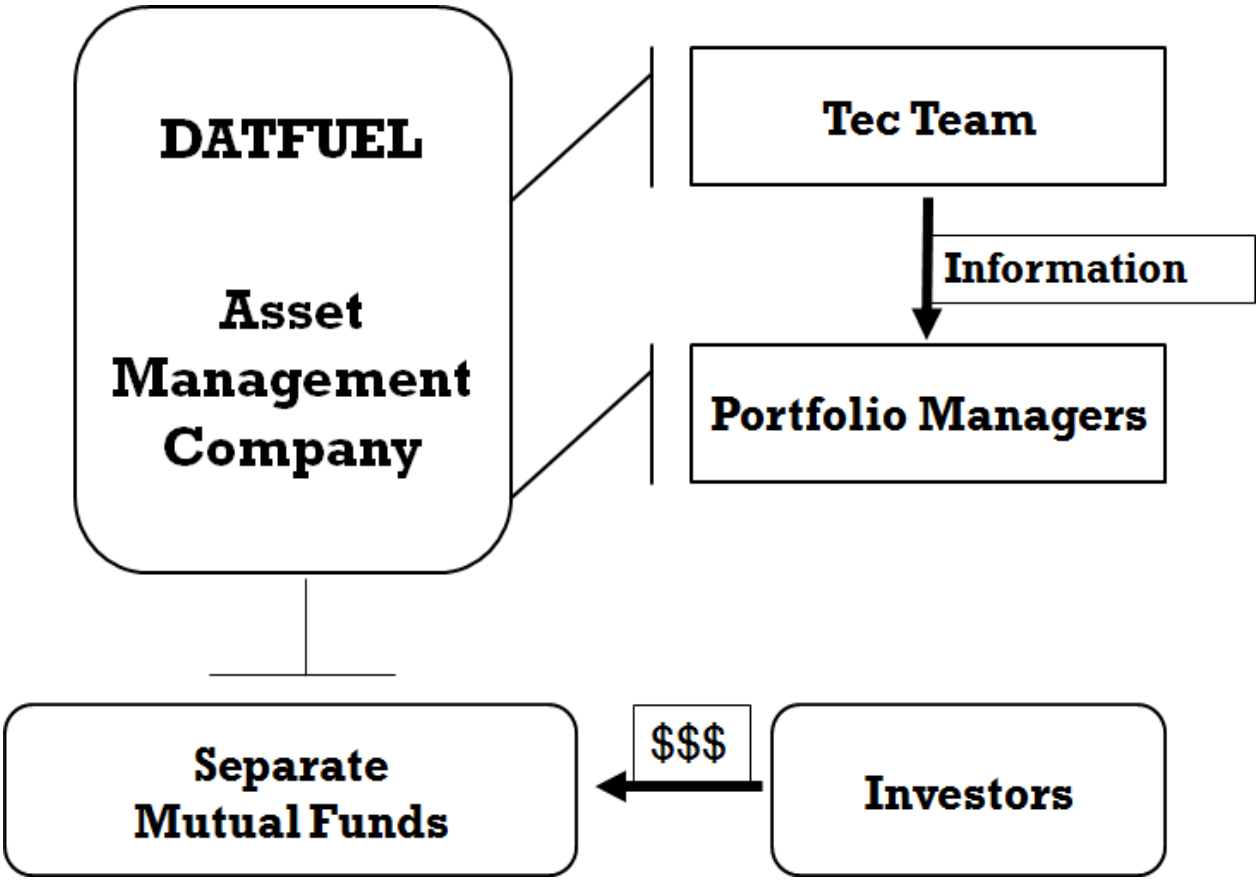

Figure 1

Our company's customers are the investors who have money for investment but do not want to invest in stocks by themselves. Mostly because they don't have time to pay attention to stock markets all the time. They choose to invest in mutual funds. Mutual funds are made up of

money collected from different investors for the purpose of investing in securities. There are several mutual funds operated by portfolio managers in our company.

Better performance of mutual fund means more investors in the market. Therefore, the job of our tech team is to predict the trend of stock prices using Support vector machine classifier, decision tree classifier, and random forest classifier and give advice to our mutual fund portfolio managers. Knowing the trend of stock price change of the next day, portfolio managers can make a better choice of trading to increase profits of the mutual funds and to make more investors interested in our mutual funds, helping in the increase of company's revenue.

- Overview:

Our purpose is to predict the direction of stock prices whether it will increase or decrease the following day. We have used historical daily stock price data for 5 years of 30 companies under Dow Jones index. We have extracted this data from Google Finance, Yahoo Finance, and the Bloomberg Terminal. This model will benefit the fund managers in making informed decisions on when to buy or sell stocks. We have used various models such as Support Vector Machine Classifier, Decision Tree Classifier and Random Forest Classifier for our predictions.

II. Data Preparation

- Describe data and data sources

Our data is from Google Finance, Yahoo Finance, and Bloomberg Terminal. We collected 5-years historical daily stock price from December 1st 2012 to December 3rd 2017 of 30 companies which are a part of Dow-Jones Index. All the data we used was well structured and we dropped the null values while making a histogram to see how the data was distributed, which helped us to create Data Quality report for different companies.

- Data Quality Report

We split the data by 3-days, 5-days and 20-days as three levels of time periods and plot histogram for each time period and each company. Here we use the stock price of American Express (AXP) as an example.

	count	mean	std	min	25%	50%	75%	max
momentum	1170.0	2.462185	8.354208	-24.786514	-2.744774	3.473256	9.034040	20.108127
ROC	1170.0	4.182884	12.010116	-33.340243	-3.547722	4.689818	13.107725	34.007834
volatility	1170.0	0.115745	0.025197	0.062242	0.098230	0.116706	0.131303	0.175221

Figure 2 (90-days)

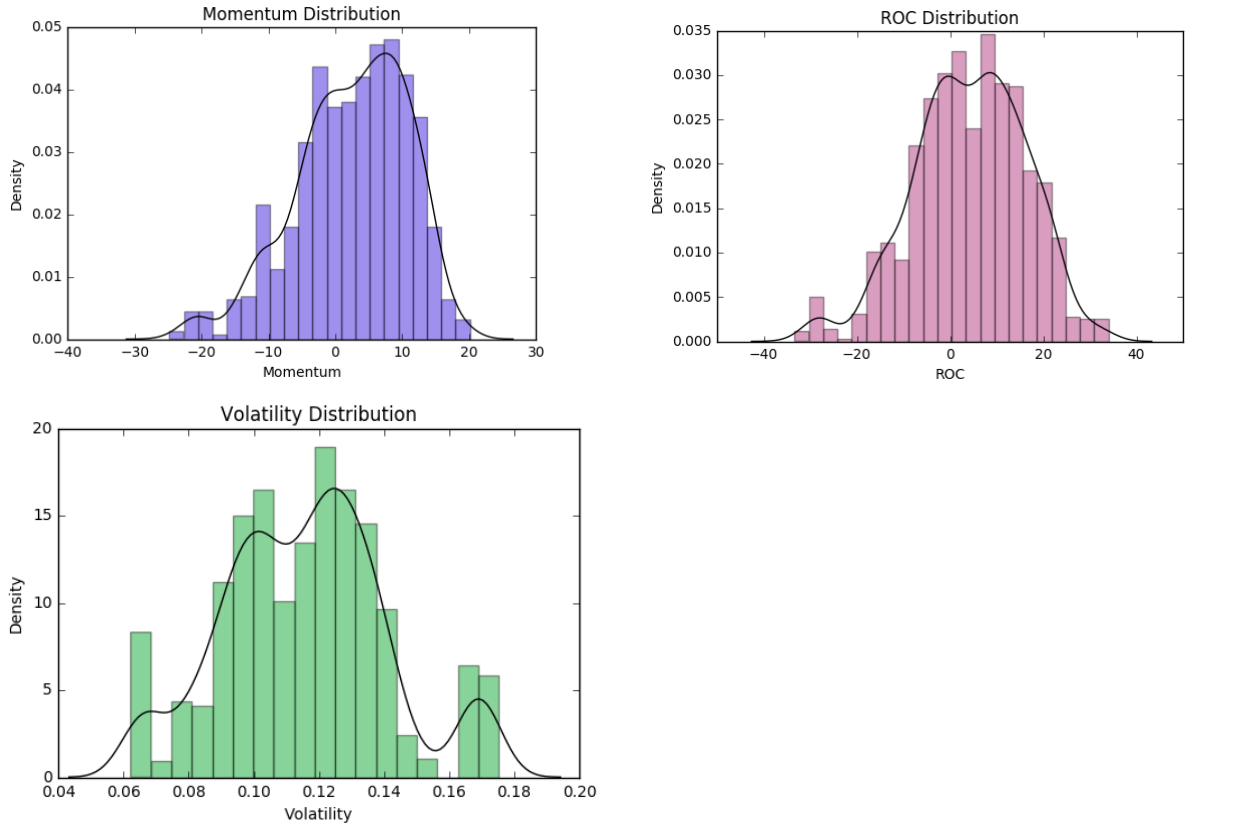

Figure 3 (90-days)

Figure 2 and Figure 3 are based on a 90-day period and each feature shows an irregular shape and pretty spread out distribution. The width of distribution implies a significant variance that each feature possesses.

	count	mean	std	min	25%	50%	75%	max
momentum	1257.0	0.072525	1.275297	-7.727921	-0.553886	0.122444	0.719543	6.593830
ROC	1257.0	0.115700	1.772533	-12.644769	-0.761336	0.163809	1.008136	11.151803
volatility	1257.0	0.016719	0.012663	0.000404	0.008436	0.014259	0.021398	0.121505

Figure 4 (3-days)

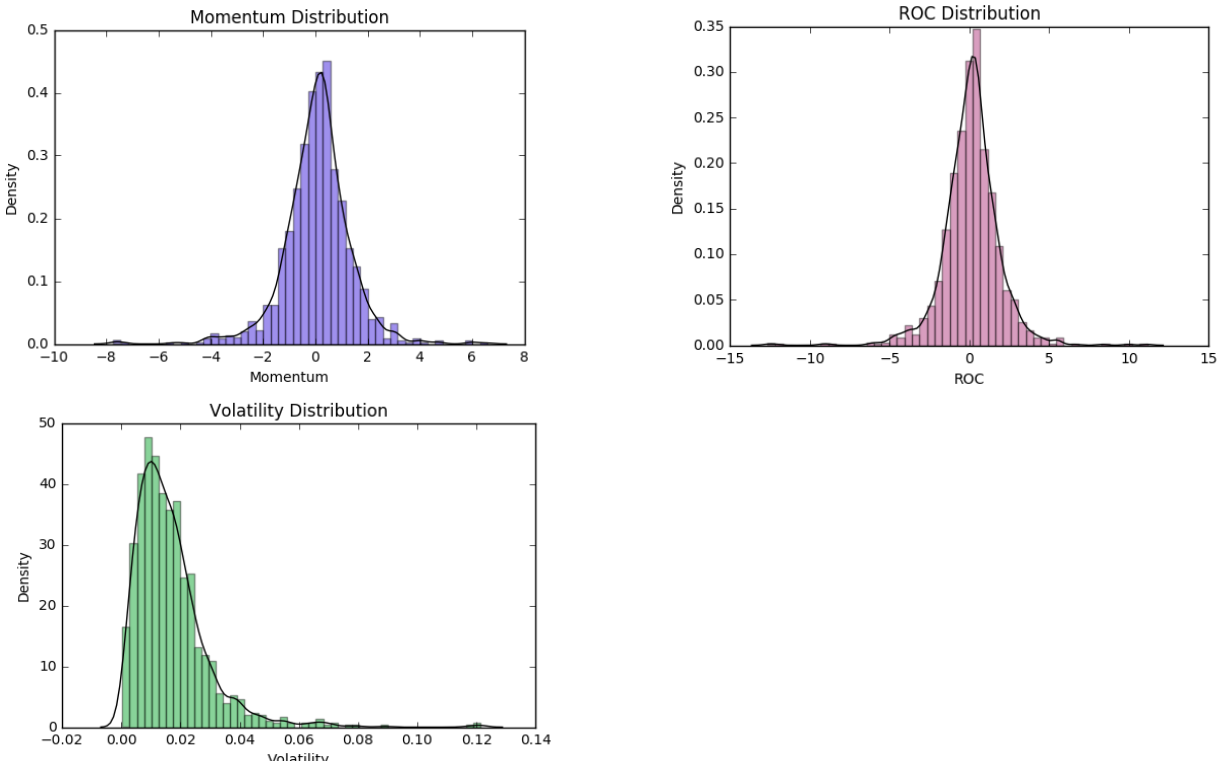

Figure 5 (3-days)

Figure 4 and Figure 5 are based on a 3-days period and each feature shows a sharp bell-like distribution, which also implies a small variance that each feature possesses.

Comparison:

Comparing the graphs based on two different time period, clearly we can see that data split by 3-days have smaller variance than the variance of data split by 20-days, which means that 3-day time period gets us better insights and accuracy for prediction. However, we did try to test on 2-days period with data. The result turned out to be data overfitting. After testing all

different time periods, we pick the 3-day interval that has the best accuracy as our time interval for models and prediction.

- Feature selection

Using the data, we picked 3 features of stocks price in order to predict the trend of stocks change.

- Momentum:

  Price difference for a fixed time interval. It measures the rate of the rise or fall in stock prices.It is a trend following indicator that helps you find the trend without being distracted by the price fluctuations and high volatility and many other indicators are also built based on moving averages. It is calculated by taking the difference between the current considered price and the price of the stock ‘n’ number of days before.

- Price Rate of Change (ROC):

  ROC is also commonly used as a divergence indicator that signals a possible upcoming trend change.Percentage change in price between the current price and the price n periods in the past. It is calculated by using the following formula:

$$\text{ROC} = \frac{Most\ recent\ closing\ price - Closing\ price\ n\ periods\ ago}{Closing\ price\ n\ periods\ ago \times 100}$$

- Volatility:

  A practical explanation of volatility is that it is a measure of the uncertainty or risk associated with the size of changes in a security’s value.It is a measure of the fluctuation of the stock price.This metric reflects the average amount a stock's price has differed from the mean over a period of time. It is calculated by determining the mean price for the established period and then subtracting this figure from each price point. It is calculated as the standard deviation of the stock price over a given time period.

III. Model Building

- Model Description

- Support vector machine classifier:

  Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for either classification or regression challenges. The coordinates of

individual features (support vectors) are used to construct a frontier (hyper-plane) which segregates the two classes of samples. Given a set of training data, each data points as belonging to one or the other of two categories, an SVM training algorithm can build a model that assigns new data points to one category or the other, making it a 2-sides classification. The most important part of SVM model is to find the optimal hyperplane. To do so, we need to compute the distance p between two points A and B and a hyperplane. The hyperplane which gives us the minimum p is the optimal one. Then, we can find target value of new data by using optimal hyperplane.

- Decision tree classifier:

  Uses a decision tree (conditional statement) to go from observations about an item to a conclusion about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a separation set of values are called classification trees; in these tree structures, we use the leaves to represent class labels and branches represent conjunctions of features that lead to those class labels. The best part of decision trees in data science is that it is able to give people a view of their machine learning models. Decision trees provide a way to approximate discrete valued functions and are robust to noisy data. Decision trees can be represented using the typical Tree Data Structure.

- Random forest classifier:

  An ensemble learning method for classification which operates by constructing a multitude of decision trees. Based on the training data, we need to build several different decision trees since there will always have some errors by bias in training data. Every decision trees will calculate the target value for new data which we use as input. Set the most appeared value as the final random forest value.

Since a model's default parameters do not always necessarily yield the best possible solution. We decided to use some Machine Learning tuning technique. Sci-kit learn's GridSearchCV method selects the best parameters for a particular model. The method also implements the k-fold cross-validation technique which is a way of saying how well the model will generalize to new data.

We got the scores for 30 companies using each model to find model performed best. The model comparisons and results is discussed later in the results and conclusion section.

- Process roadblocks faced:

Initially, we started out with different equity fundamental indicators such as price to earnings ratio, earnings before income tax depreciation amortization(EBITDA), equity trade

volumes. The model feature selection and weighing proved to be too cumbersome for the models that we ran. The accuracy prediction scores that were obtained as a result were far from significant.

Also, in our initial approach, we had considered predicting the actual stock price itself. So essentially our business problem was that of a regression problem. After getting really low r2 scores and by going through different scientific journals we realized that it was better to predict price trends than actual prices itself.

IV. Modeling Outcomes

- Results for the choice of time period:

The accuracy scores were compared in order to find the best time interval to predict the trend in stock price change. Tableau software was used for making the below chart.

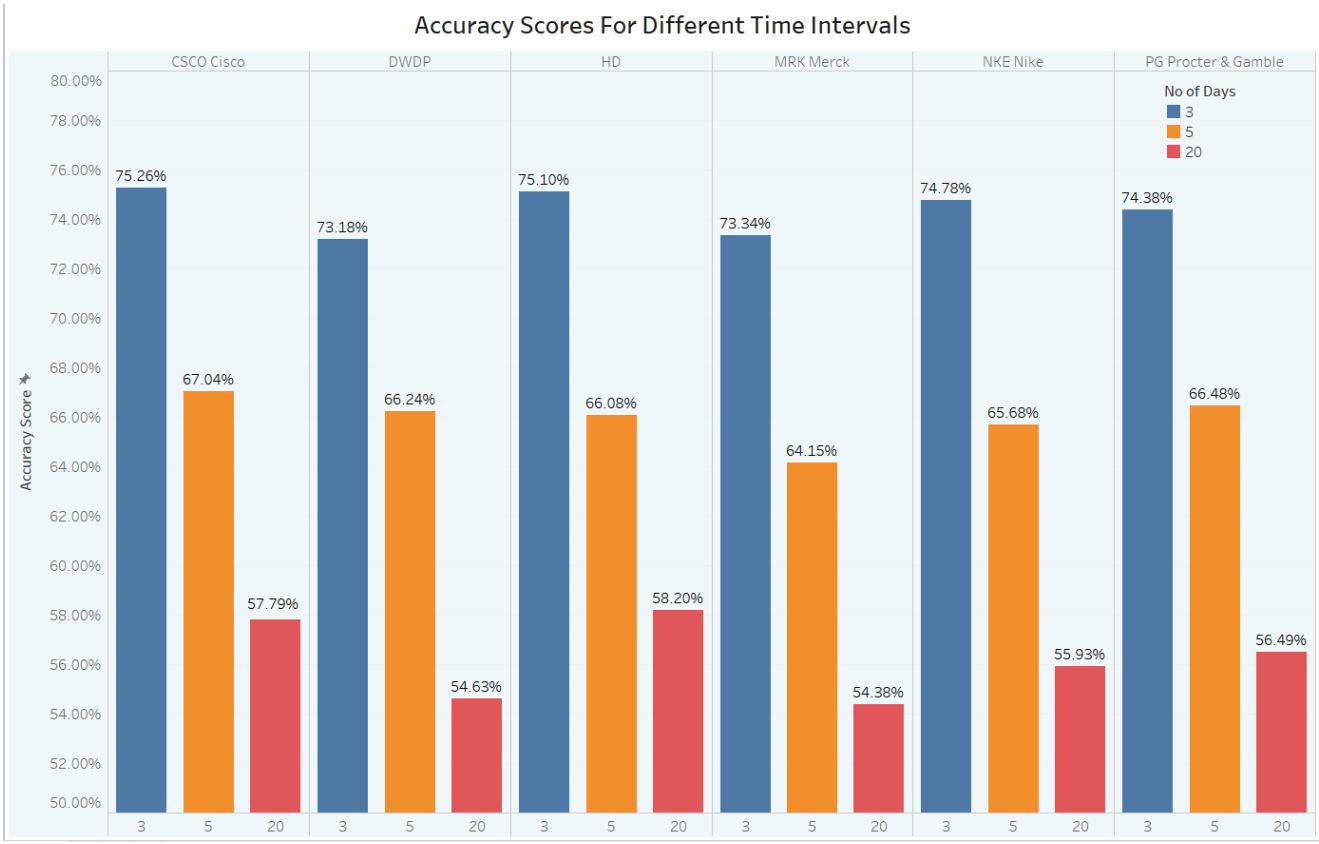

Figure 6

The above graph shows that the time interval period of 3 days results in the best accuracy scores possible. This is in line with the fact that it is better to predict the trend in a shorter time interval as compared to a large time interval. The model we used for comparing the scores is the SVC.

- Choice of machine learning algorithm:

We used 3 machine learning models to predict the trend in stock prices. We wanted to first choose the model which gave the best results. Below chart represents the comparison between 3 different machine learning algorithms. The time interval chosen is 3 days since it results in the best-predicted results across all 3 classifiers.

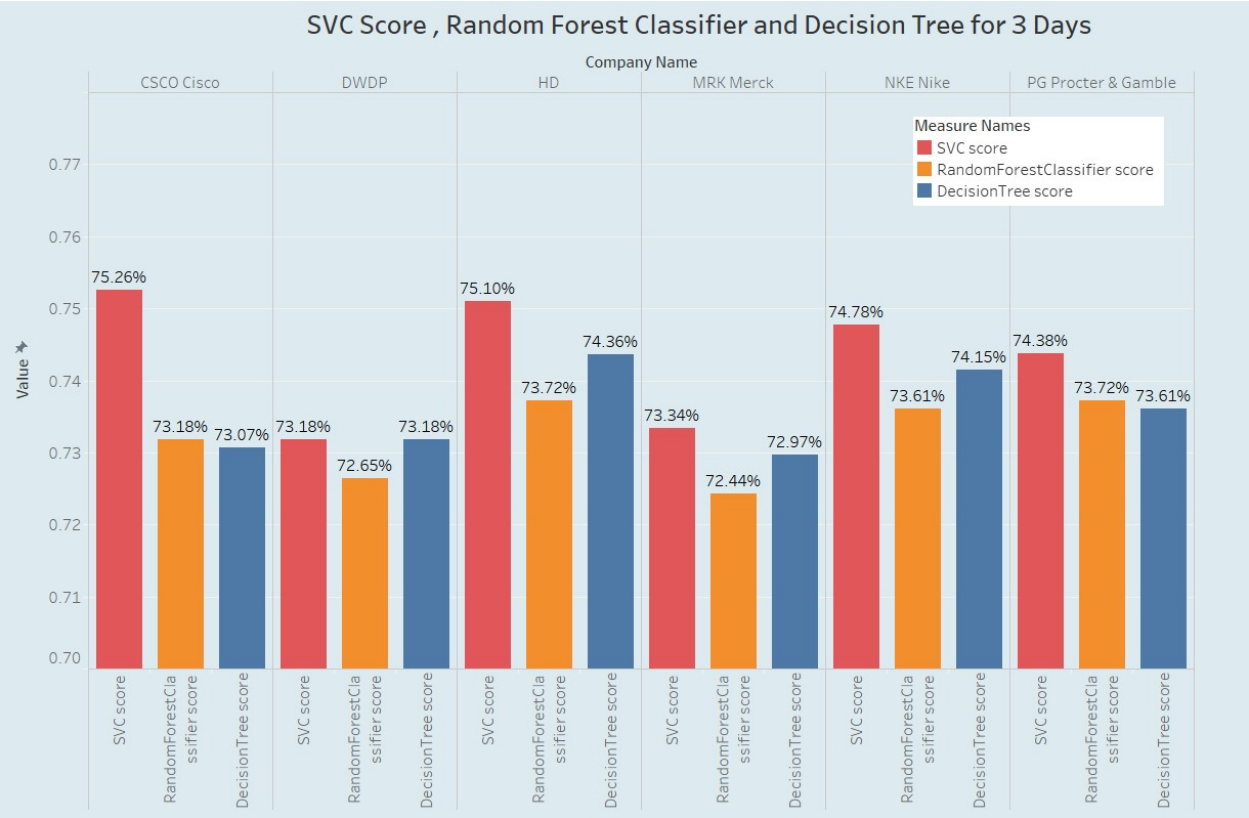

Figure 7

As seen above from the above chart, the support vector machines classifier gives the best scores. On closer examination, we see that there is not really a significant accuracy score difference between SVC and Decision tree scores for some companies such as Home Depot(Ticker symbol: HD) or Nike(Ticker symbol: NKE). We then need to make further quantitative analysis to choose the best machine learning algorithm.

- Analysis using precision-recall curve:

The recall versus precision curve was plotted for to compare the results of different machine learning models. Below figure illustrates one of the comparison of precision-recall curve for the support vector classifier.

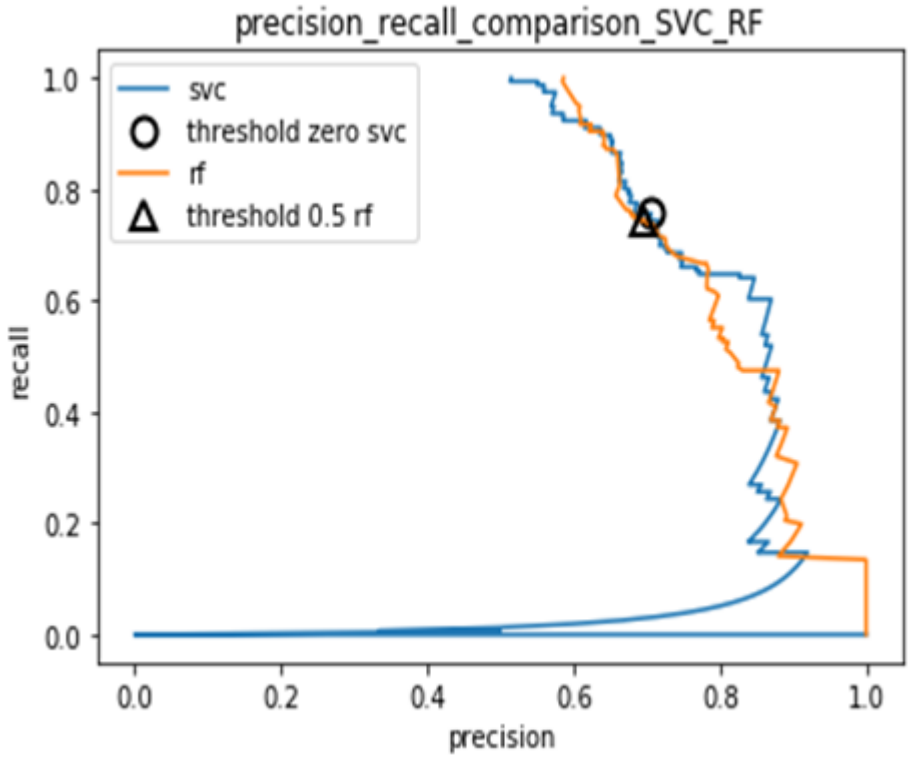

Figure 8

As you can see we cannot make any decisive conclusions about which model is performing better from the above figure. The other plots also resulted in the same figures.

Analysis using the area under the curve method for a R.O.C curve:

The receiving operator characteristic area under the curve is a measure of classifier performance, which is widely used in machine learning. It is basically a plot of true positive rate (or recall) versus the false positive rate. The area under the curve for each of the model gives its AUC score. Below figure illustrates the ROC curve for the 3 machine learning models that we used which can be used for comparison.

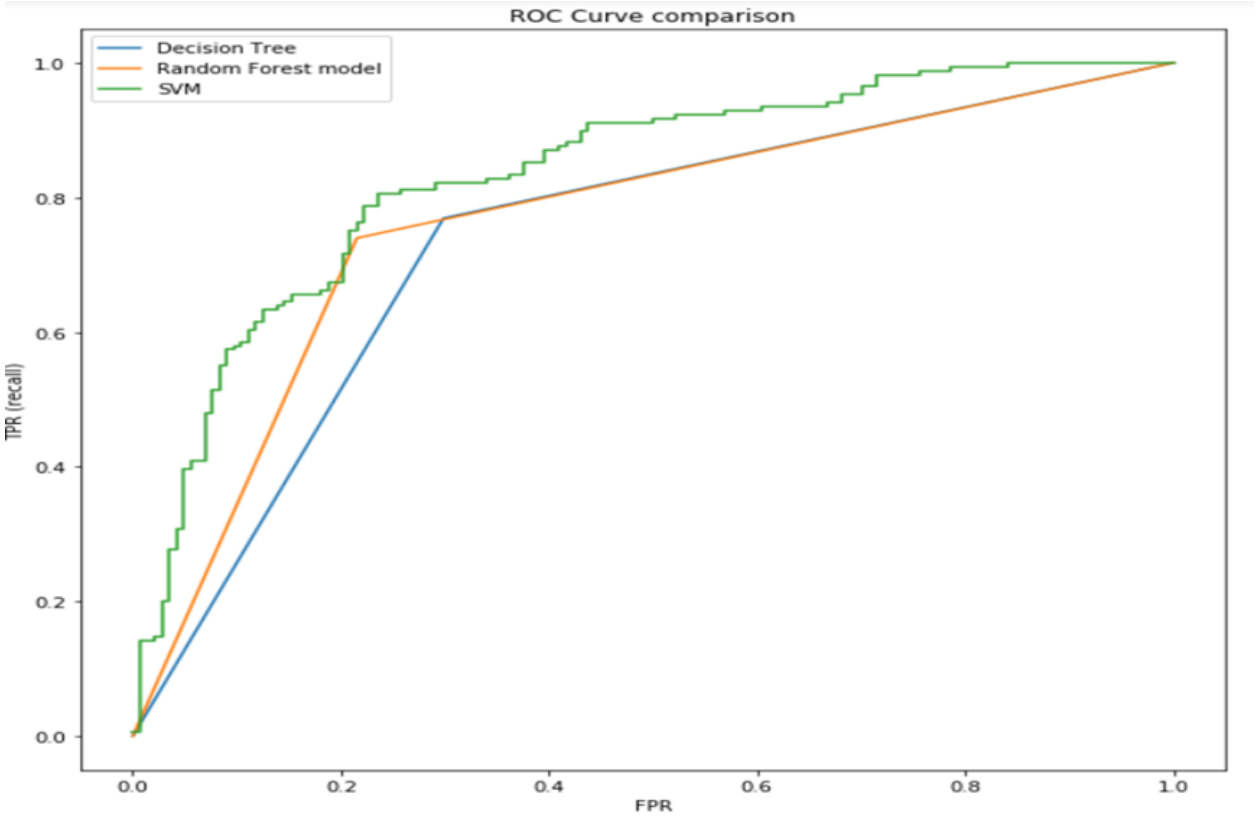

Figure 9

This figure was plotted using matplotlib and scikit learn package. The ROC curve gives us some sort of an idea that the area under the curve for SVM model would be greater than the other two classifiers.

- AUC score comparison:
  The following illustrates the AUC scores for different classifiers.

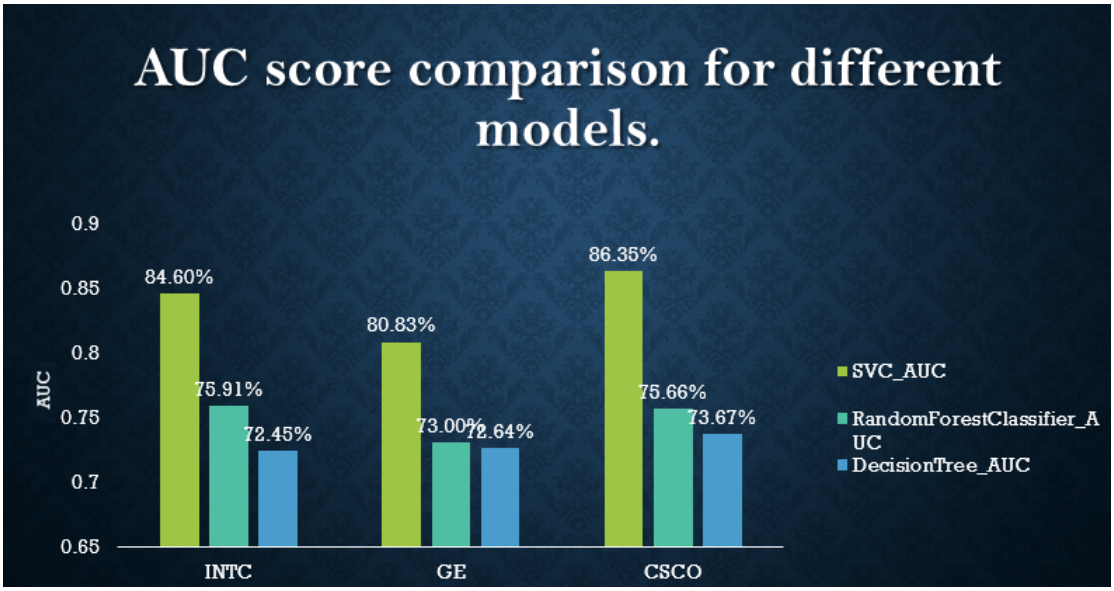

Figure 10

This figure was generated using Microsoft Excel visualization tool. The above bar charts show the AUC scores for different companies when each company stock price is predicted for all the three models. It shows that the AUC score is significantly higher for the SVM model when compared to the random forest classifier model.

- Section summary:

In this section, we first chose to determine which time frame will give us the best prediction accuracy for the trend in the stock price on the following day. Then we compared different models by using precision-recall curve, ROC AUC method. The summary of the conclusions and its business implications are listed in the next section.

V. Conclusions and Business Implications

- Conclusions:
  - Accuracy tells us how precise our model is. We conclude from our results that the accuracy for predicting the trend is higher for small time intervals i.e. 3 days or 5 days in comparison to a longer time interval i.e. 20 days or 90 days. This implies that predicting the trend of stock price for the next day is more accurate if we take the prices of previous 3 or 5 days instead of taking a longer time interval. This happens because the stock prices are very erratic and getting good accuracy from a long-time span is indecisive.

- Although accuracy scores for our three models were very close, we saw that Support Vector Machine classifier outperformed. This conclusion was made by taking the Average Under the Curve (AUC) score of the three models. AUC score gives us the amount of all positive points that have a higher score than all the negative points. So, a perfect AUC score of 1 tells us that the model is 100% accurate.

- When predicting the trend of the equity price for any company, the accuracy score was not adversely affected by the industry sector to which a company belongs. Stock prices of companies belonging to different industry sectors such as retail, telecommunications, aviation, etc. were used and yet accuracy scores did not have a significant variation in them.

- Business implications:

We believe that our business problem will help the mutual funds in raising their net asset value which in return will give more profits to our company and its investors. No financial statements were used to find our training features. Through stock prices, we derived all of them and used them for our model, which in return gave a slight improvement in the accuracy of our models.

Using the data science team's tuned model, the fund manager of each mutual fund can take an informed decision whether to buy a particular equity or sell it. The fund manager can also take the decision to whether take a long or a short position on a particular equity. Long position is when the buyer expects that the stock price will rise in the future. The short position is where the stock owner sells the stock hoping that the stock price will decrease and buy the stock at a lower price.

VI. Future Work

We would like to expand our business further by incorporating more ideas into this model to make it more accurate and successful. Since we tried to predict the trend of the stock prices for the following day, another important feature that we can work upon is to try and predict the degree to which the prices will increase or decrease using artificial neural networks. This would give us more insights into the stock prices which will help our business to grow and in return make higher revenues for the company.

Incorporating additional features such as Price to Earnings ratio, stock dividends and volumes will digress from the technical features and give us results using different features. In our model, we have used all three features in finding our predictions. We plan to use two out of these three features to see which would give us the best results. Many papers in scientific

literature also used the use of artificial neural networks for prediction. Since these methods were outside of what was taught in class, we will also try and build a model around these concepts. Lastly, we plan to develop a hybrid model to predict stock market trends. For e.g. using a combination of principal component analysis and support vector machines.

VII. References

1. https://finance.yahoo.com/
2. https://www.bloomberg.com/
3. https://finance.google.com/finance
4. Machine Learning Techniques for Stock Prediction - Vatsal Shah
5. Predicting Stock Price Direction using Support Vector Machines - Sahil Madge
6. Predicting Financial Time Series Data Using Hybrid Model- Bashar Al-hnaity and Maysam Abbod
7. Stocks Market Prediction Using Support Vector Machine- Zhen Hu , Jie Zhu, and KenTse.
8. G. F. Bjoern Krollner, Bruce Vanstone, “Financial time series forecasting with machine learning techniques: A survey,” in European Symposium on Artificial Neural Networks: Computational and Machine Learning, Bruges, Belgium, April 2010.