A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

7/24/2021

Project

A Quantitative Analysis of the Stocks of Different Sectors; Comparing the Ease of Prediction of One Sector from another.

Several thin, curved lines in shades of blue and grey originate from the bottom left and sweep upwards and to the right.

Kashfay Haider Naqvi
21389466

Acknowledgements

Writing this report was a very long and thought-provoking process, requiring me to emerge my mind into new ways of thinking and analysing. This would not have been possible without some of the people around me who helped me continuously.

Firstly, I would like to thank my lecturers and specifically my supervisor, Dr Scott Yang, who put in a lot of hard work, supporting me, all the way from the beginning, reading my early drafts and helping me improve in our regular meetings.

Another thanks go out to my family, who have, over the course of the last few years, given me a platform to study, whilst sacrificing their own needs.

And finally, to my friends at university who have helped get me through the years with their uplifting personalities and vibrant character. Some of you, I cannot mention by name, but you know who you are. A special thanks goes out to one of my closest friends, Dhiraj Khullar, who encouraged me to get interested in the Stock Market initially. Without his help, this report would have been impossible to complete.

Contents

Acknowledgements.....	1
List of Figures and Tables.....	4
Introduction	5
Stock Market Background.....	5
Stock Sectors.....	5
Machine Learning.....	6
Relevance	6
Research Questions	7
Main Question	7
Sub Questions	7
Project Aims & Objectives.....	7
Literature Review.....	9
Algorithms.....	9
Evaluation Metrics	10
Methodology.....	11
Data Collection.....	11
Data Process and Analysis.....	12
Python for Machine Learning.....	12
Microsoft Excel.....	12
Algorithms.....	13
Evaluation Metrics	14
Results.....	16
RQ 1 - How do the different sectors within the stock market compare when it comes to ease of prediction?	16
RQ 2 - Which algorithm produced the best predictions?	19
Discussion.....	22
RQ 1 - How do the different sectors within the stock market compare when it comes to ease of prediction?	22
RQ 2 - Which algorithm produced the best predictions?	24
RQ 3 - Does the effect of Covid-19 play a part in stock prices of companies?	28
Conclusion.....	30
Limitations / Future Works	30
Appendix	32
Python Code.....	32
SVR and LR (svr_fb.py)	32

ANN (svr_ann_collab.ipynb)	35
Excel Merge (merge.py)	37
Link	40
Raw Data	41
Forms	42
Ethics Form.....	42
Project Supervisor Form.....	49
Project Progress Form 1	50
Project Progress From 2	51
References	52

List of Figures and Tables

Figure 1	16
Figure 2	16
Figure 3	17
Figure 4	17
Figure 5	17
Figure 6	18
Figure 7	18
Figure 8	19
Figure 9	19
Figure 10	19
Figure 11	20
Figure 12	20
Figure 13	20
Figure 14	21
Figure 15	21
Figure 16	25
Figure 17	25
Figure 18	26
Figure 19	26
Figure 20	28
Figure 21	29

Introduction

Stock Market Background

A stock market is an exchange platform where sellers and buyers of stocks can directly interact with each other. Sellers put shares of their business on the market for the public to purchase. The prices of these shares can vary, and these shares can be traded for money in the stock market. To be able to predict the trend of a stock, therefore, is quite appealing. However, it is practically impossible to achieve such prediction with a 100% accuracy, due to the various of factors which may influence the index.

Within the stock market itself, there is a variety of different sectors in which each stock can be categorised in. The current sectors consist of Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate and Utilities. A company would publish their stock within these sectors too and the growth of a sector is seen by a certain number of top stocks within that sector.

As we know, there are many different stock market exchanges which are based all around the world. In this dissertation, the main focus will be on one of the world's most popular exchanges, the NASDAQ. The main reason for this is the availability of resources and information regarding this stock exchange as well as it being relevant to more people around the globe.

Stock Sectors

Communication Services – this is one of the newer sectors which includes some telecommunication service providers such as wireless telecoms and landline services. This sector also includes some media and entertainment companies. Facebook, Alphabet and Netflix are amongst the companies within this sector.

Consumer Discretionary – this sector encompasses stocks for which the demand relies upon the consumer's financial status. The companies would sell luxury items as well as leisure items. Within this sector, you would find companies such as Amazon and McDonald's.

Consumer Staples – this sector includes companies which consumers go to regardless of their financial status. In simpler terms, items that are a 'must-need' such as food and drink as well as supermarkets. In this sector, you would find Coca Cola and Walmart.

Energy – this sector covers companies dealing with oil and natural gas. As well as this, it includes companies dealing with other consumable fuels such as coal. Related businesses to these companies are included within this sector too i.e. a company who provides equipment, materials and services to gas producers. Chevron Corporation is within this sector.

Financials – this sector relates to businesses which deal with money. You would find major banks alongside insurance companies and a few more. Some of the companies you would find are Berkshire Hathaway and Morgan Stanley.

Health Care – this sector has two distinct sections, one including companies which develop and sell drugs and the other covers companies which sell health related equipment. Health insurance companies are also included within this sector. Johnson & Johnson are one of the companies present in the Health Care sector.

Industrials – companies which range from business that use heavy equipment as well as larger transportation services such as airlines, trains etc. Those companies which build large machinery as well as aerospace and defence are included in this sector too. You will find Union Pacific in this sector.

Information Technology – this sector covers the majority of companies within the technological world. Companies that create software and hardware related to IT are included. Some examples of companies in this sector are Microsoft and Apple.

Real Estate – this sector covers companies which deal with real estate and real estate management. The companies related to these types of companies are also included within this sector. Simon Property Group is in this sector.

Utilities – this sector includes all the companies which provide availability of essential services such as electrical power and gas distribution. Also, companies which provide water nation wide are included in this sector.

Machine Learning

Machine learning (ML) is one of the branches of Artificial Intelligence (AI) and it consists of algorithms which learn from the data being fed into them, by improving the accuracy without necessarily being programmed to do so. In other words, Machine Learning promotes self-learning for the machine.

Machine learning uses a data set and splits it into training and test data. The training data is used to feed the algorithm and train it to fit the whole dataset, as well as any future, unseen data. The test data is then used after to examine how well the model was at predicting unseen data.

We must also run an algorithm on the training data. The main type of algorithm we will be using in this study will be regression algorithms. This is due to the fact that with our datasets, we will be dealing with continuous data rather than discrete data. Regression is used to predict the value of a dependent variable using an independent value.

Machine learning is preferred to other statistical models due to its ability to be able to handle larger datasets better. Statistical models are generally better than ML when it comes to predicting/forecasting using datasets with a small amount of data and features however, when the dataset increases in size and the features within the dataset increase, ML starts to outperform other statistical models very consistently.

Relevance

The last couple of years have been hugely influential on the stock market, as well as other trading markets, such as the Cryptocurrency market. The interest that these markets have accumulated has reached an all-time high, with people who do not even have much financial knowledge beginning to invest into them. They see the potential reward for doing so and believe it can help their financial status.

One of the main reasons can be put down to the current situation of the Coronavirus. This has left some people redundant as well as making others realise that their job may not be as secure as they

had initially thought. So, these markets become a lucrative opportunity to make some more money on the side.

In this study, we shall stick to the stock market prediction. As experts say, it is important to know and understand the company before you invest and buy some shares of it. My study will concentrate on these issues and giving individuals a better overview on which sectors of the stock market will be easiest to predict, and it will allow investors to consider the risk of investing into the sector of their liking.

Research Questions

The study will focus on one main question as well as a few sub questions which corroborate the main question. The sub questions will provide some alternative path to answering the main question too.

Main Question

RQ 1 - How do the different sectors within the stock market compare when it comes to ease of prediction?

This will be the main basis of the study. The prediction for each sector within the stock market will be analysed and then compared to other sectors. The sector which is the easiest to predict will be stated with clear reasoning as to why this is.

Sub Questions

RQ 2 - Which algorithm produced the best predictions?

RQ 3 - Does the effect of Covid-19 play a part in stock prices of companies?

These two questions will be supporting the main question by providing alternative reasoning for some of the prediction values. We will also delve into these questions separately.

Project Aims & Objectives

The aim of this study is to predict the closing price of a stock a certain number of days in the future. The stocks will be separated into sectors and the evaluation metrics for each sector will be calculated and compared with the others. This will allow investors to have a better understanding on the stability of the stocks in a sector; something which is important when making decisions regarding buying and selling. It is also good investing practice to understand the stock before you invest, so the aim will be to provide a broader understanding of the volatility of the closing price in the sector.

The Objectives of the study would be:

- Identify all the sectors within the Stock Market and use a certain number of stocks from them to use as representation of that sector.
- Download the datasets for these stocks and format them in a way which can be read by the libraries used in Python.
- Decide on a few algorithms which will be used for prediction and code for them in Python.

- Feed the algorithms the dataset for a range of days: 1,2,3,4,5,10,15,30 and compute their results. Extract these results into a worksheet and have them formatted in a way which is readable.
- Gather and analyse the results.
- Compare the results of the stocks in the sectors using the evaluation metrics outputted by the Python scripts.
- Compare the different algorithms against each other.

Literature Review

Algorithms

In a study regarding the Karachi Stock Exchange (KSE), the researchers found that there are certain factors which have an impact on the market performance. They narrowed them down to six factors: Market History, The News, General Public Mood, Commodity Price, Interest Rate and Foreign Exchange. After these factors were taken into consideration, the ML techniques applied to the dataset were Support Vector Machine (SVM) and three variants of Artificial Neural Network (ANN): Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF). “Applying SVM algorithm on the training and test set gives different results”; producing 100% accuracy on the training set but only 60% accuracy on the test set, SLP gave 83% and 60%, MLP gave 67% and 77% and RBF gave 61% and 63% respectively for training and test set. On average, MLP had the highest accuracy as well as the highest amongst all the ML techniques for the training set. Hence, MLP is more effective at predicting the market performance as the test set is what the model needs to be tested on for unknown instances. We must note that Support Vector Regression (SVR) and SVM are similar. SVM is used for classification, which is what was done in this study, and SVR is a regression algorithm which can be used with continuous values [1].

In another paper, the researchers stress on the use of Regression Models for predicting continuous values with given independent values. This paper uses two models. One of the models is a basic Linear Regression Model which uses factors of the dataset such as open, close, low, high and volume values of each stock. The second model is a Recurrent Neural Network (RNN) which is the regression version of a classification Artificial Neural Network. The version of RNN used is Long Term Short Memory (LSTM). A sequential model is used in this paper which stacks two LSTM layers on top of each other. In this study, the LSTM model came up on top [2].

In another article, we understand the use of regression analysis. Regression analysis is used to estimate the relationship between a dependant variable and some independent variables. Due to this, it is widely used for prediction and forecasting, which enables it to be implemented within the field of machine learning. To carry out regression analysis, there are many different techniques including linear regression and ordinary regression. These are known as parametric. There are also nonparametric regression techniques, one of which is SVR. Polynomial, sigmoid and RBF regression are also mentioned within the article [3].

In a recent study done in November 2020, D. Madhusudan used solely SVM regression (SVR) to predict the outcome of two stocks, Tesla Inc. and Reliance NS. SVM regression algorithm was used with different kernels. These kernels transform the input data into its required form. Three different kernels were used in the study: linear kernel, polynomial kernel and radial basis function kernel. The researcher then continued to take historical data from Yahoo Finance regarding the two companies and used this to produce a prediction. The conclusion was that “all the SVR kernel methods perform differently on different data points”. However, RBF SVR kernel ended up on top with the best prediction for both Tesla Inc. and Reliance Ltd. The graphs produced also show that RBF is the better SVR kernel as it predicted the stock prices to the closest degree with the original graph compared to the other kernels. It was also mentioned that in the future to consider the view of people as this can also affect the stock price of a company [4].

On a paper there are three different algorithms which are compared: SVM, Random Forest (RF) and ANN. RF comes up on top with the best prediction accuracy of 52.9331% whereas SVM has 51.7287% and ANN has 52.3216%. Even though RF does have the overall better prediction accuracy, it is clear that the accuracies for all the algorithms are very close and that this only applies for the dataset chosen for this study, which is aimed at a Chinese A-share market. Due to how close these values are, we cannot say there is a significant advantage of using RF over the others [5].

To further back up the point above regarding RF, another study was conducted which was investing different strategies. In this study, three different algorithms were used: Linear Regression (LR), RF and SVR/SVM regression. For the stock predictions analysis, on average, the SVR performed best at predicting stock price over a certain number of weeks. For example, SVR gave an average accuracy for predicting 4 weeks in advance of 66.6%, LR gave an average of 64.5% and RF gave an average of 58.2%. This is of course one example but the general trend within the study is that SVM outperforms the other two. LR also marginally outperformed RF within this study [6].

Evaluation Metrics

In a study which proposes a stock price prediction model, extracting data from historical data as well as social networks, a set of different evaluation metrics were used to test and compare the different prediction models against each other. These measures were Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) [7].

In this paper, soft computing techniques are used to gain forecasting results. The researchers also analysed 100 different published papers which focus on neural and neuro-fuzzy techniques. These researchers then use the input data, forecasting methodology, performance evaluation and performance measures to compare and classify the techniques. That statistical measures used were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Squared Prediction Error (MSPE) [8].

In another paper discussing evaluation/performance metrics in machine learning regression and forecasting, surveys were conducted over a 25-year period. The researchers were surveying organisations involved within this industry to see which metrics they use the most. A large number of metrics were initially identified and the topmost metrics were discussed. These metrics were MSE/RMSE, MAE and MAPE. In general, these metrics should not be used as singles but rather together to see an overall picture instead of just one aspect [9].

Methodology

Data Collection

The data required for this research was quantitative data. Quantitative data is data which the value is in the form of a number. This data can then be utilised in mathematical calculations as well as statistical analysis. The output of this data can be used to draw conclusion regarding the hypothesis or research questions [10].

The data obtained is also secondary data. As we know, primary data is data which the researcher has themselves collected through primary sources. For our research, we could have also collected primary data by looking at a stock index and deriving stock prices daily however this was not necessary as there are many resources which can provide us with accurate and reliable readings of the market as secondary data. Using secondary data allows us to conduct statistical analysis on this data without worrying about the physical data collection from a primary source.

To obtain this data, we used Yahoo Finance. This network is a part of one of the largest Media outlets in the world, in Yahoo, and has a very respectable reputation. They collect financial data and news, stock related news as well as reports, and allow the public to view and download them.

As Yahoo Finance has such a large database of data, we had to obtain data which met our certain criteria. As the main research question required me to collect data regarding sectors of the stock market, a decision had to be made about how many datasets we would need to represent a whole sector. There can be numerous stocks within a sector, and it would not be efficient to represent a sector using all of the stocks within it.

I decided to rank the stocks in each sector in decreasing order of Market Cap, with the stock with the highest Market Cap value being at the top. I then filtered out and selected the top 10 stocks of each sector. This way we would be able to compare the strongest stocks in each sector. Also, these are the stocks which most investors would know and decide to invest in, making this study more relatable.

Another criteria was the time range from which we gathered data. As we are using historical data, a decision had to be made about how far back we must take the data from. Do we go back to the day the company was made public on the stock market, giving us different data ranges for each stock, or do we set a date from when we take data. Also, looking at previous works, they use a timescale of going back around 10 years, such as the work done by Jigar Patel et al [11]. The decision was made to use data from the 1st of January 2010 all the way up to 17th May 2021. This was due to the fact that taking uneven data for each stock could result in a difference between stock prices, and this is something we are not analysing. On top of that, the market as a whole had just slowly started to recover after the 2008 recession, so the use of data after would give us more stable data. Different companies were affected in a different way during the recession. An article shows that building materials and supplies dealers had a sales loss of around -3.28% due to the recession, whereas furniture stores had a sales loss of around -0.54% [12]. This again shows how different industries were affected differently so it would be unfair to include data during the recession.

During this collection process, there were some stocks which were not public in 2010 and became public later. This resulted in some stocks having slightly less datapoints, however as a whole there wasn't many which may have drastically affected the result of the algorithms.

The data taken was also daily stock data. We did not focus on the hourly/weekly data as the research objectives required us to forecast days into the future rather than hours or weeks.

Data Process and Analysation

As we download the datasets from Yahoo Finance, for each stock we have access to different features. These features are: Open, High, Low, Close, Adjusted Close and Volume. I decided to look at the close and adjusted close values for each stock. The close price itself is the literal cash value of the stock at the end of the market day whereas the adjusted close price takes into account more stock attributes such as dividends, stock splits and some new stock offerings. This shows that the adjusted close price will give us an overall better understanding of the value of the stock compared to just its closed price which could be largely affected by a stock split [13]. Due to these reasons, the basis of this research was around the adjusted close price of each day.

To further prepare the data, we created another column with predicted adjusted close price for a certain number of days in advance. For example, if we are trying to forecast out only one day, in the predicted adjusted price column would be the adjusted price for one day ahead. If we are attempting to forecast 15 days, then the predicted adjusted close price column will contain the adjusted close price for 15 days in advance. This was however not done manually and was done using Python. This coding language was used to prepare the data as well as analyse the data. Microsoft Excel was also used to view and play around with the data.

Python for Machine Learning

Python is a programming language which allows the coder to create and deploy high level data structures with a simple approach to object-oriented programming [14]. It is especially useful when it comes to Machine Learning (ML). The core algorithms of this research are ML algorithms, so it would make sense to utilise Python for this. As ML algorithms can get very complicated, Python offers a simple and consistent approach, making the code much more human readable as well as taking the focus off learning the language and purely on enhancing the algorithms [15].

Python also offers countless libraries related to ML. Due to the complexity of ML algorithms, some of them can take time and as a coder it is not efficient to get caught up coding for other solutions, such as Mean Squared Error, so Python allows the user to download and import pre-made libraries with snippets of code and methods already created. Some examples of these libraries are NumPy, which can be used for data analysis, and Scikit-Learn, which is a ML related library including concepts such as K-Folds and training data splitting etc. [15]

Microsoft Excel

Excel is a spreadsheet software, produced by Microsoft, which allows the user to organise and format data [16]. This was used in the research in a few different ways alongside Python. Firstly, the ML algorithms run in python outputted statistical errors into an Excel workbook for each stock. Another Python script was the run which gathered all the data from the Excel workbooks and putting them into another workbook. Finally, we had an Excel workbook which contained the formatted data in a readable way, as well as containing data analysis such as bar charts and graphs.

Algorithms

According to a source book, there are a few different types of algorithms which can be found in ML. Two of the main ones involve supervised and unsupervised learning [17].

Supervised Learning (SL) is a subcategory of Machine Learning where labelled datasets are used to train algorithms. The training of the algorithm then allows the model to predict and classify future data points according to the training dataset. SL can adjust the parameters within the algorithm, such as weights, as data is inputted during the training stage, until the model is fitted suitably. SL is separated into two components: Classification and Regression. Classification is used to categorise data into clusters/categories and regression is used to manipulate data between an independent and dependent variable and output a value, rather than assign a category [18].

For our research, we will be using Regression as our aim is to produce a prediction for a certain number of days into the future rather than a classification. In an article, we understand the use of regression analysis. Regression analysis is used to estimate the relationship between a dependant variable and some independent variables. Due to this, it is widely used for prediction and forecasting, which enables it to be implemented within the field of machine learning. To carry out regression analysis, there are many different techniques including linear regression and ordinary regression. These are known as parametric. There are also nonparametric regression techniques, one of which is SVR. Polynomial, sigmoid and RBF regression are also mentioned within the article [3].

Unsupervised Learning is another subcategory of ML where unlabelled datasets are used to train algorithms. In unsupervised learning, clusters are created based on similarity and relationship between the outputs [19]. For this study however, we will only be using supervised learning as our datasets include all labels.

Within SL, there are many different algorithms which could be used. After researching and reading articles, there is a major trend of using algorithms such as Support Vector Regression (SVR) models and Artificial Neural Networks (ANN). Consequently, I decided to implement these two models as well as a simpler Linear Regression (LR) model.

Linear Regression (LR)

Linear Regression is a very common model which is used to identify a relationship between variables. As the name suggest, LR produces a straight line which attempts to fit onto the data with the independent variable being the adjusted close price of the date and the dependent variable being the predicted adjusted close price for the next 'n' number of days [20].

A line is plotted on the graph and distance between the line and the datapoints is calculated, known as the residual, and summed up. Note that the distance is squared so there are no negative values. The line is then further rotated multiple times and the line that has the lowest squared sum is the line which fits best to the data. This line is then used as the Linear Regression line.

Support Vector Regression (SVR)

The concept of SVR is very similar to that of SVM (Support Vector Machine). As we know, a SVM tries to distinctly classify data points using a hyperplane. Along with the hyperplane, we have decision boundaries either side of the hyperplane, where the distance between the two boundaries is known as the margin. To build a better SVM, a larger margin is required as this means the data has been

split with more confidence due to the larger gap between the distinct groups of data. A hyperplane can be a linear line or non-linear, as well as being more than just 2 dimensional [21].

The difference when it comes to a support vector regression algorithm is that the margin must be as small as possible which allows the production of a hyperplane to fit the data the best. Contrary to a simple LR model, SVR can support non-linear hyperplanes too. This gives the SVR more functionality as well as predicting to a higher degree of precision [22].

Artificial Neural Network (ANN)

An ANN consists of a few different required components/layers. There are input layers, hidden layers and output layers. The input and output layers are quite straightforward where the input takes in the raw dataset and the output is the predicted outcome after the data has gone through the neural network and the hidden layers. The hidden layers are the main component of the ANN as that is where all the functions take place. There is backpropagation going on between the hidden layers which help to optimise the weights between the input and other layers [23].

Evaluation Metrics

A combination of evaluation metrics are used for the research. After inspecting the articles, we picked out some of the most commonly used metrics, which are the most useful when it comes to giving the user a picture of the data being analysed, especially related to stock prediction. We used four metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

Mean Squared Error (MSE)

This tells you how close the regression line is from certain points. This is done by the distance between the point and the line being calculated. The distance is calculated along the y axis and not perpendicular to the line. This distance is then calculated between every point and the regression line, squared, and then the average distance is calculated. The distance/difference is squared because this eliminates any negative differences between the actual and forecasted values. Here is the formula:

$$MSE = \frac{1}{n} \sum (actual - forecast)^2$$

We can see that the difference between the actual and forecasted values is being squared and then summed up. The average is then taken by dividing by n, which is the number of datapoints.

Root Mean Squared Error (RMSE)

This evaluation metric is very similar to MSE. The only difference is that after the MSE has been calculated, the value is then squared rooted, making it considerably smaller. This makes it easier to compare values as extremely high MSE values will have a much smaller value for RMSE. Here is the formula:

$$RMSE = \sqrt{\frac{1}{n} \sum (actual - forecast)^2}$$

Mean Absolute Error (MAE)

This tells you the size of the error without considering the direction. For example, in MSE actual – forecast may give a negative number, so it is squared to eliminate this. In MAE however, an absolute value is used so the difference between the actual and forecast has no direction meaning the value will always be ≥ 0 . This means the value does not need to be squared, giving smaller values in general compared to MSE. Here is the formula:

$$MAE = \frac{1}{n} \sum |actual - forecast|$$

We can see the difference between is being actual and forecasted is being summed up for all the points. The absolute value is taken, making even the negative values into positive values. The mean is then taken to give us MAE.

Mean Absolute Percentage Error (MAPE)

The issue with MAE is that the value given cannot definitively define the size of the error as it is relative. A way in which this problem could be dealt with is by taking the MAPE as this will allow you to compare results with different sets of errors. The only difference from MAE is that we are calculating the percentage here so the difference is calculated and then divided by the actual value. Here is the formula:

$$MAPE = \frac{1}{n} \sum \left| \frac{actual - forecast}{actual} \right|$$

Overview of Evaluation Metrics

We see how the different metrics have their own advantages and disadvantages and how they can portray the data in a different way allowing the user to manipulate the results in a way of their liking. So, we decided to use all four different metrics and the results can be discussed, analysing the different aspects of the forecast algorithm.

MSE and RMSE do give more weighting to larger errors due to that fact that the difference between the actual and forecast is squared and as this difference increases, the squared difference increases in an exponential manner. This in return gives more weight to the larger errors within the set.

MAE can ignore larger errors as the difference will not be significant enough to have a major impact on the final error value. So, chunks of data at certain times which give larger errors compared to the rest of the data may be slightly disregarded. This can be an issue if these larger errors require more attention.

All in all, a mix of all these four errors will give the user a broader understanding in which they will be able to understand the algorithm to a better degree than if just one or two evaluation metrics were used [24].

Results

RQ 1 - How do the different sectors within the stock market compare when it comes to ease of prediction?

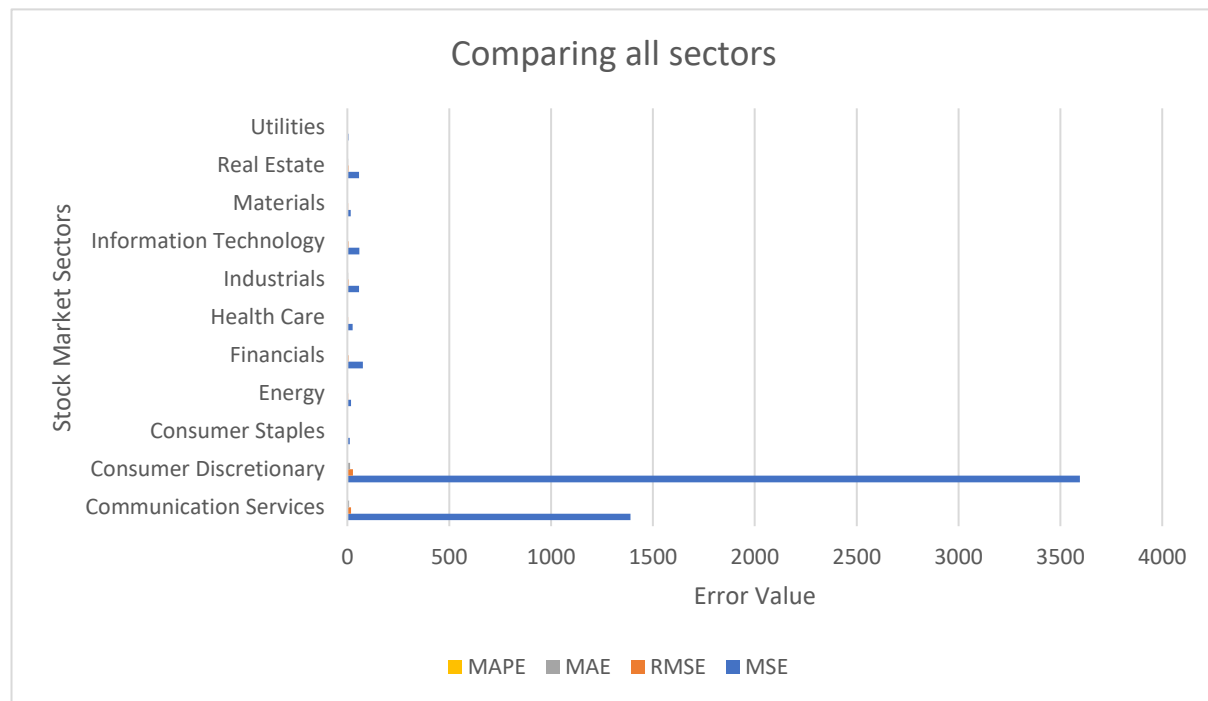


Figure 1: All evaluation metrics have been averaged over all forecasted days. This bar chart shows the evaluation error metrics for each sector.

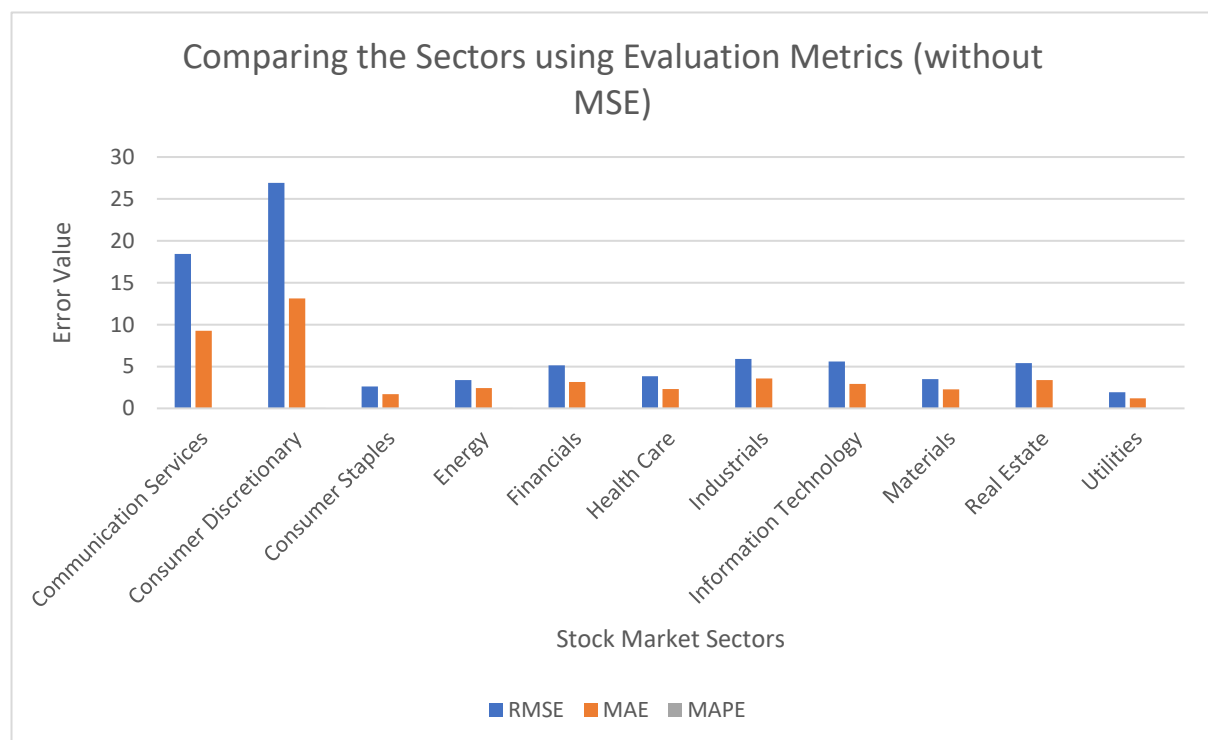


Figure 2: MSE removed for each of the sectors to show a better comparison between the smaller error values.

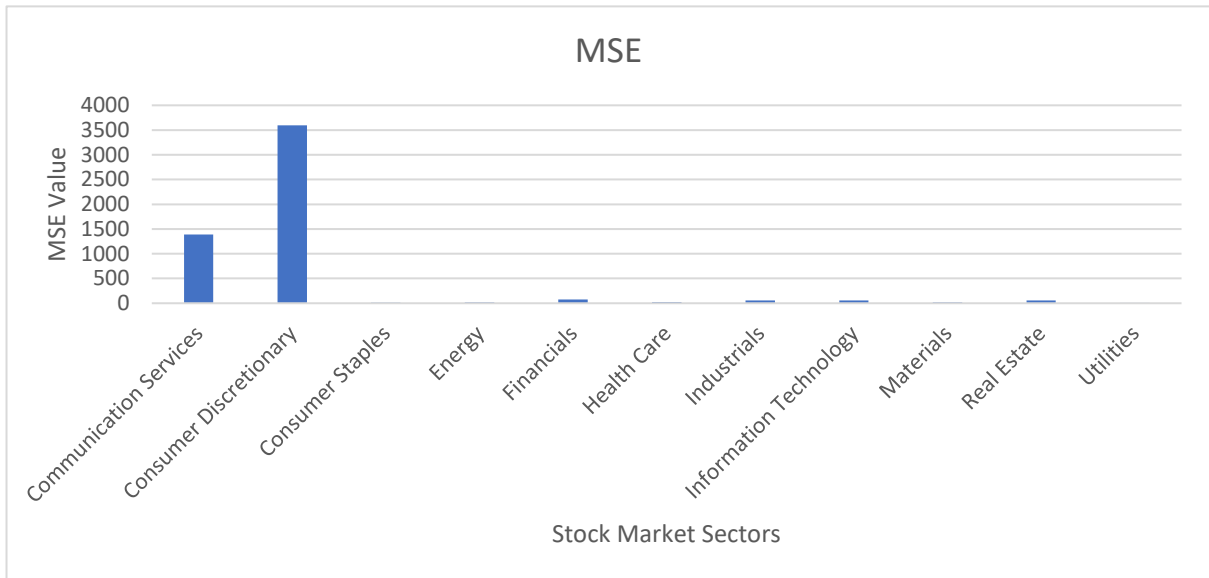


Figure 3: MSE only for each sector.

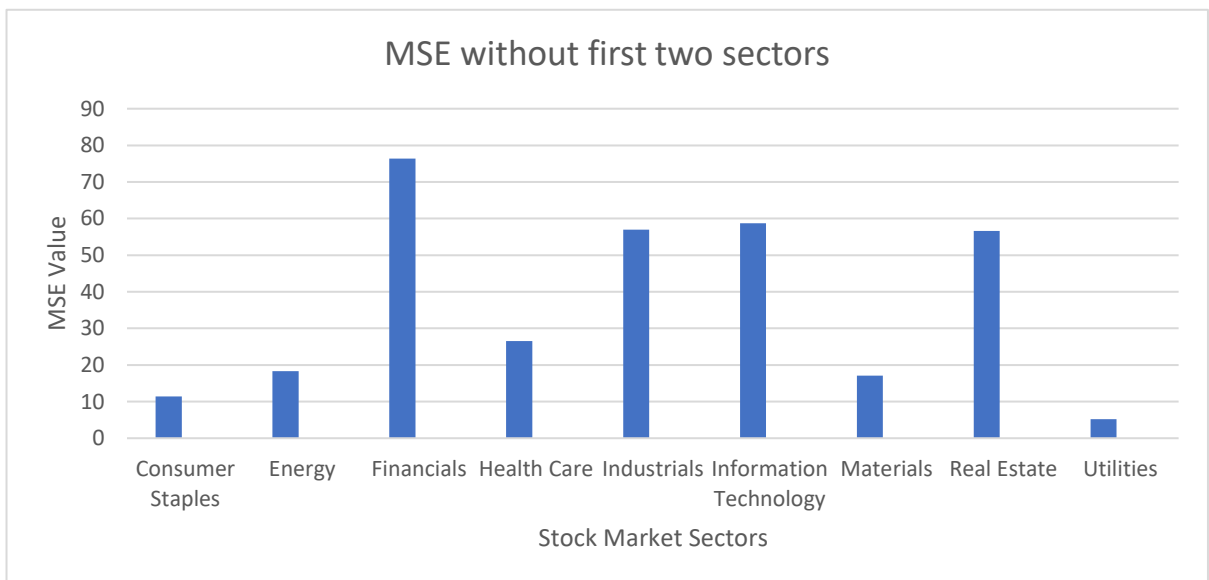


Figure 4: MSE only removing Consumer Discretionary and Communication Services as their MSE values were too high compared to the rest, making it difficult to make a comparison.

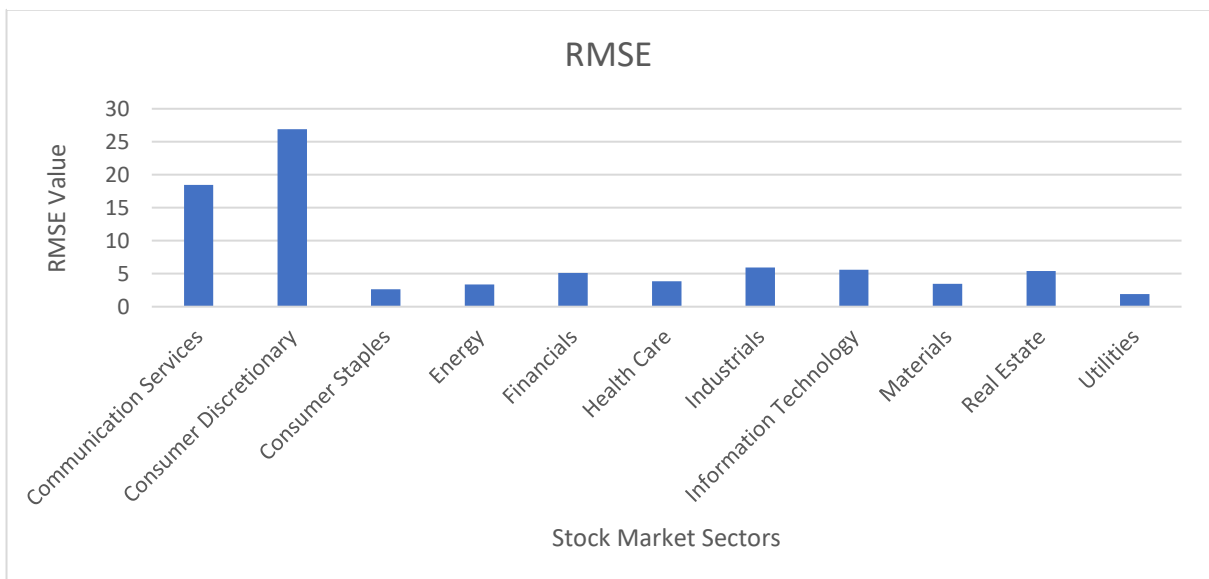


Figure 5: RMSE only for all the sectors.

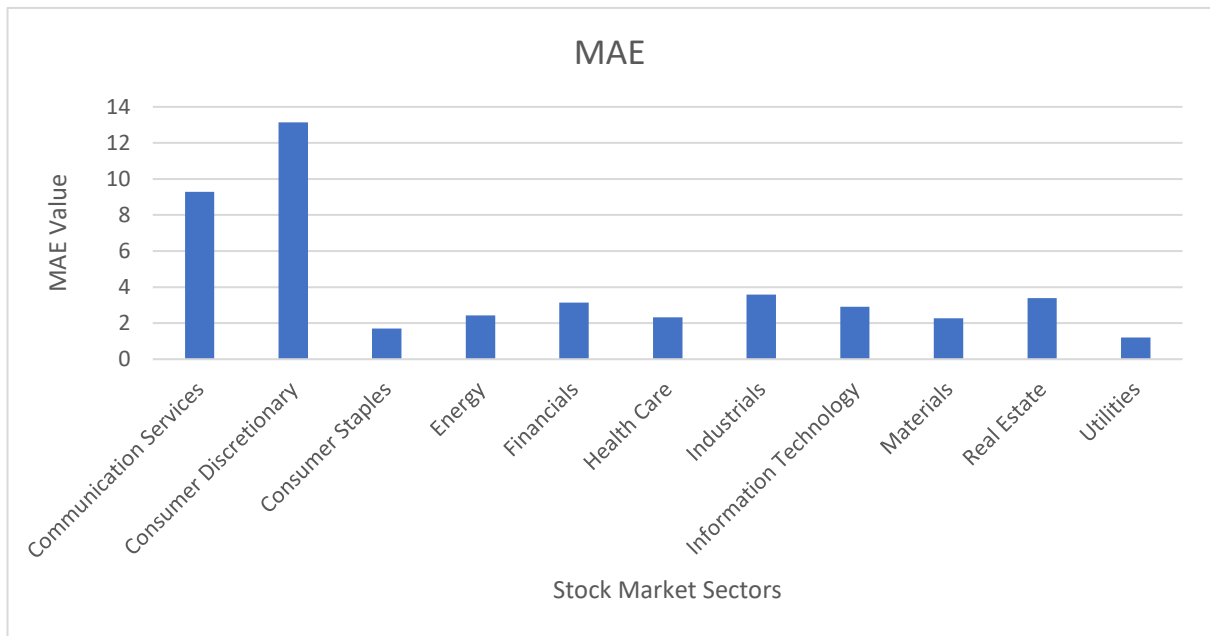


Figure 6: MAE only for all the sectors.

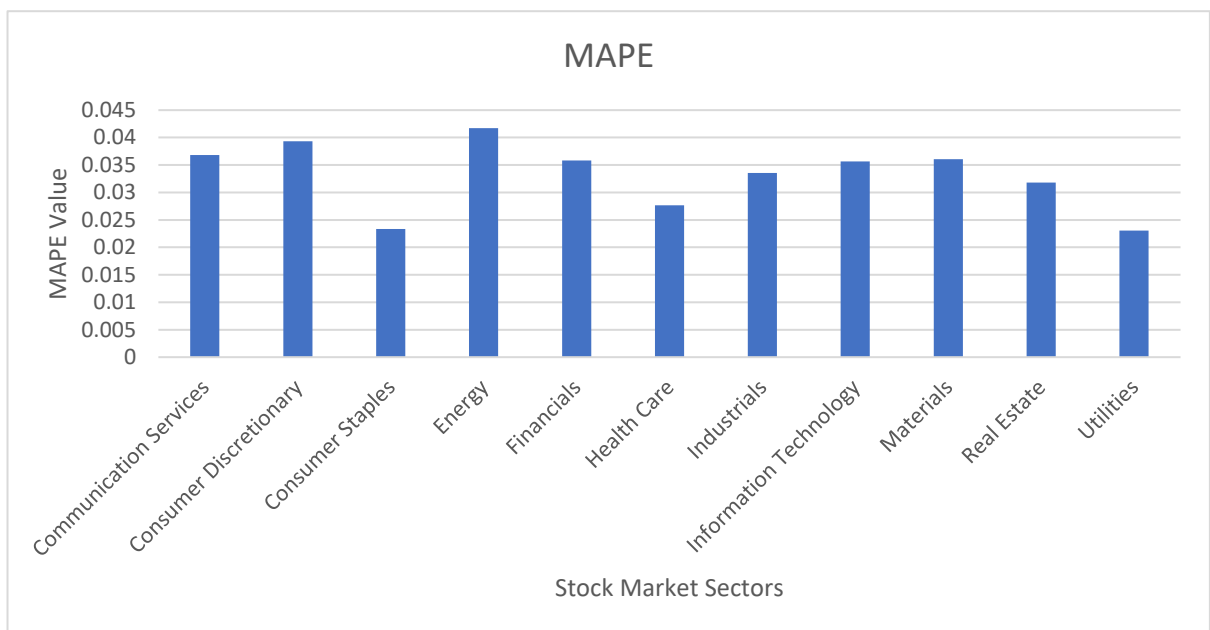


Figure 7: MAPE only for all the sectors.

RQ 2 - Which algorithm produced the best predictions?

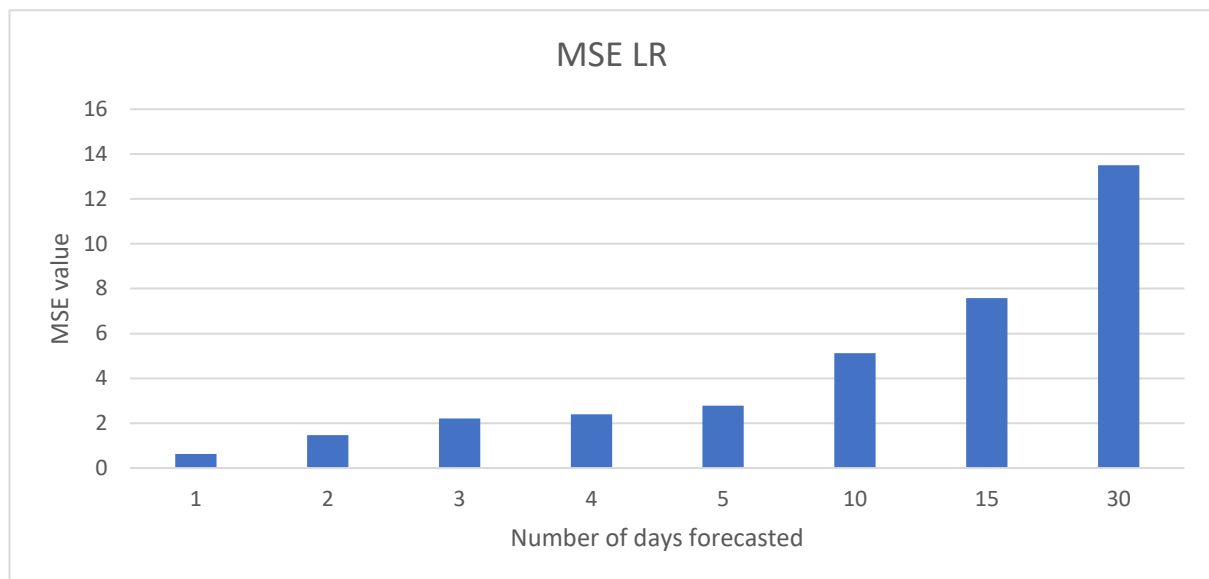


Figure 8: MSE for LR over 8 forecasted days in Utilities sector, showing the increase of error as the days progress.

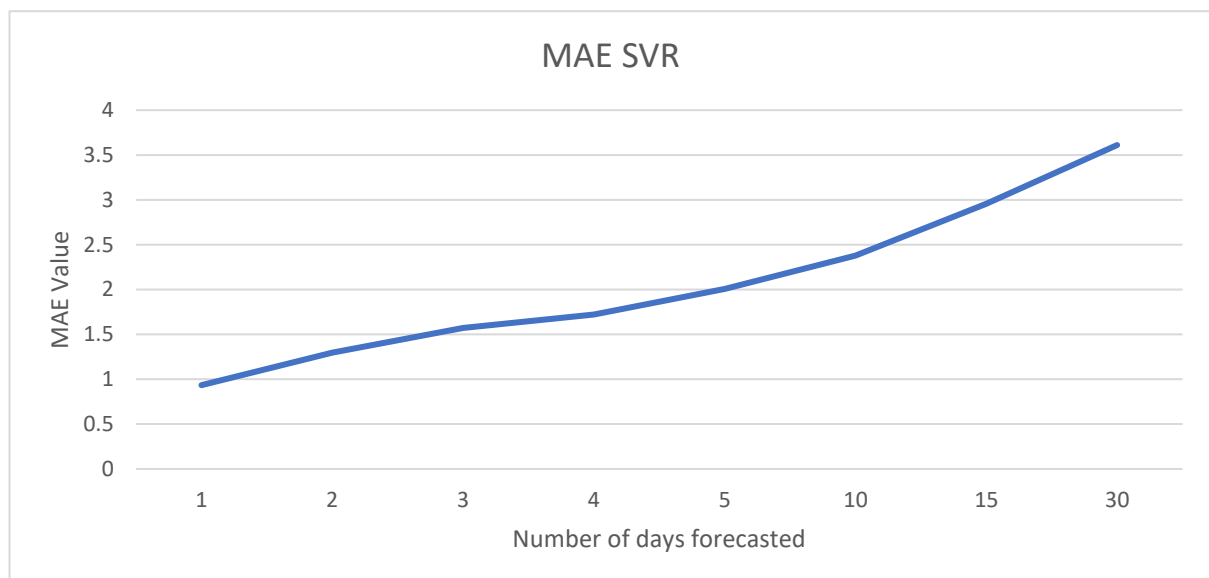


Figure 9: MAE for SVR over 8 forecasted days in Health Care sector, showing the increase of error as the days progress.

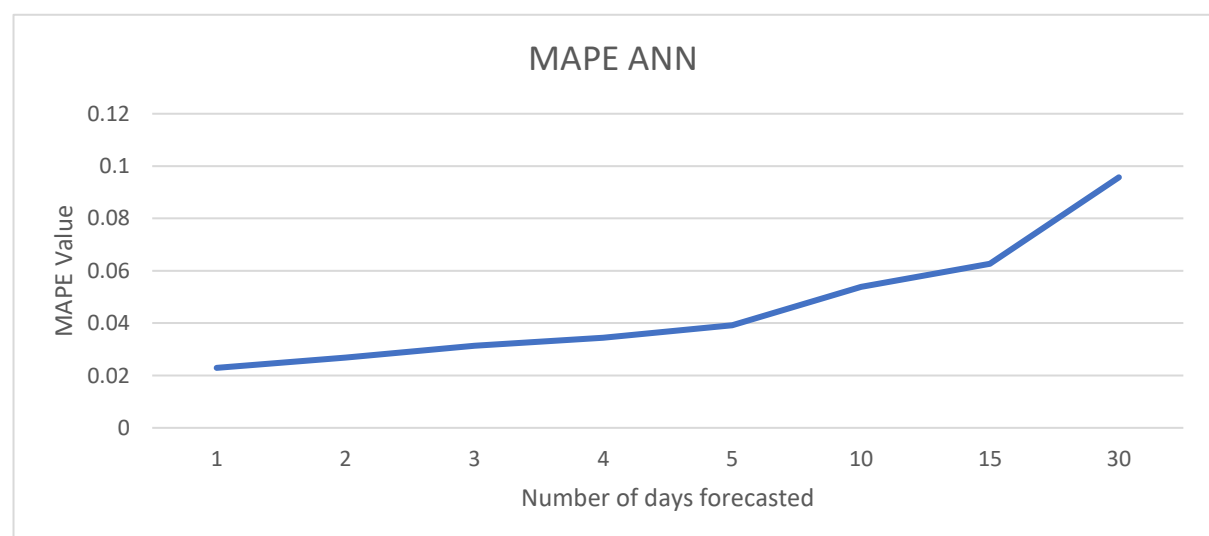


Figure 10: MAPE for ANN over 8 forecasted days in Energy sector, showing the increase of error as the days progress.

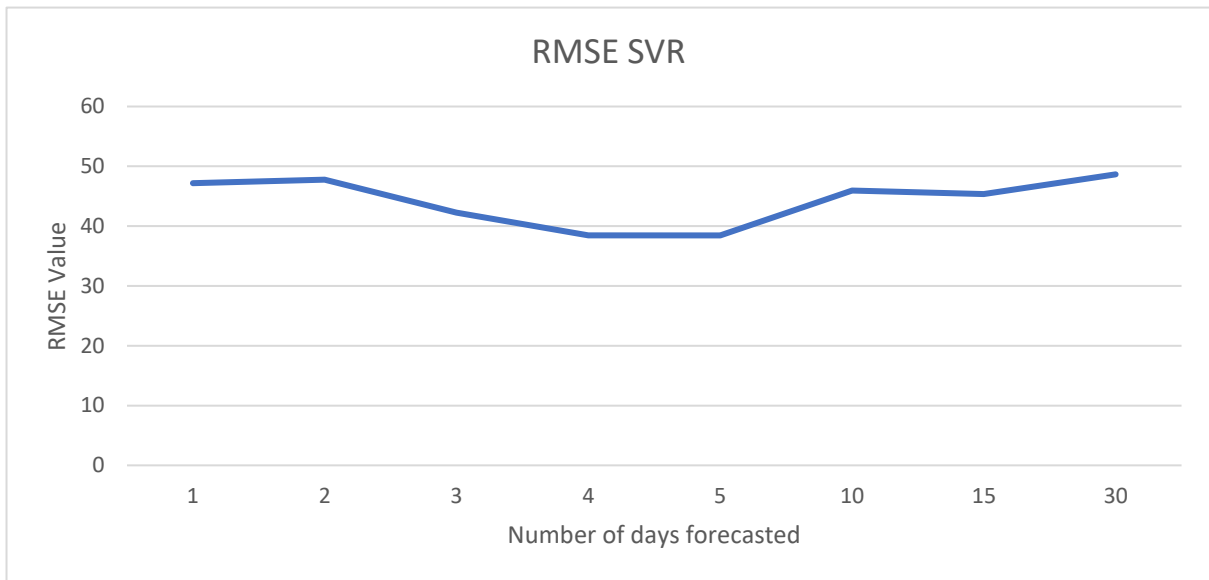


Figure 11: RMSE for SVR over 8 forecasted days in Consumer Discretionary sector, showing a consistent error as days progress.

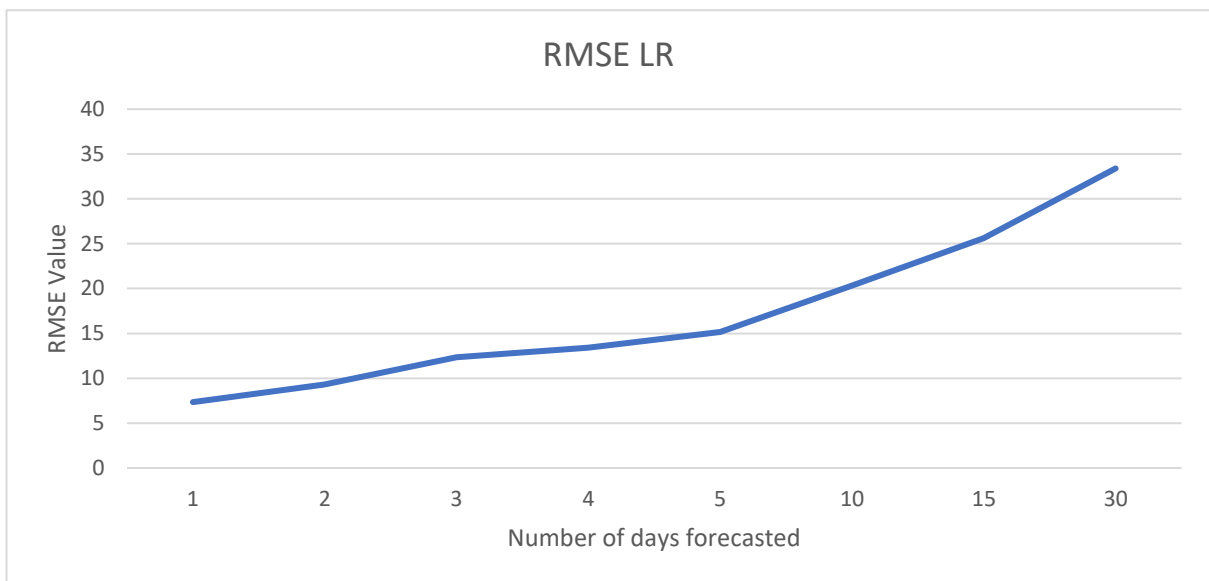


Figure 12: RMSE for LR over 8 forecasted days in Consumer Discretionary sector, showing the increase of error as the days progress.

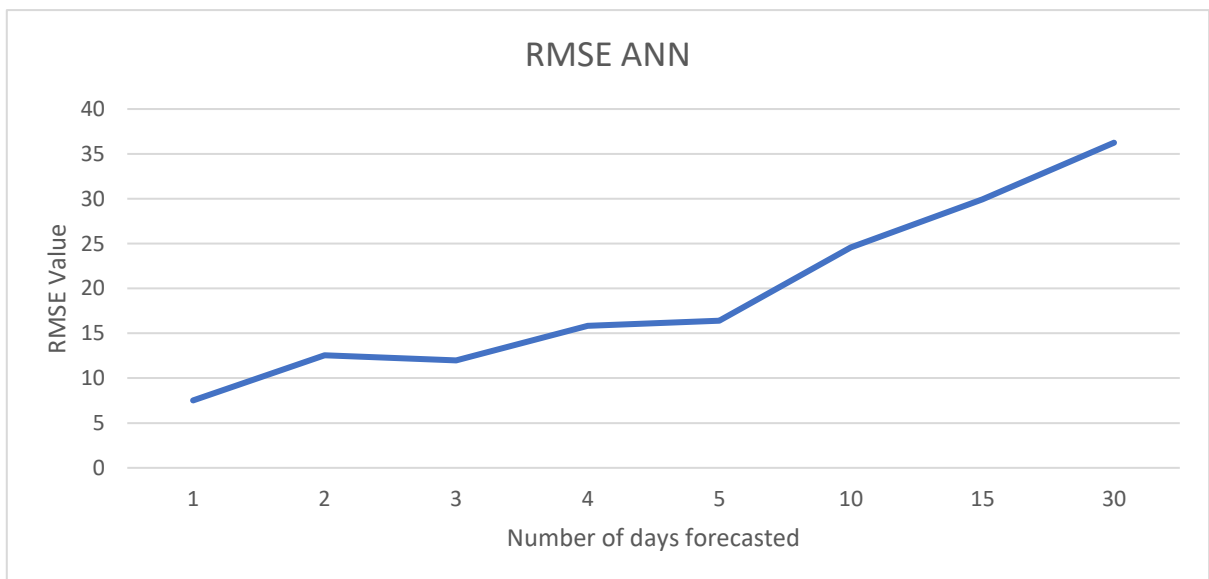


Figure 13: RMSE for ANN over 8 forecasted days in Consumer Discretionary sector, showing the increase of error as the days progress.

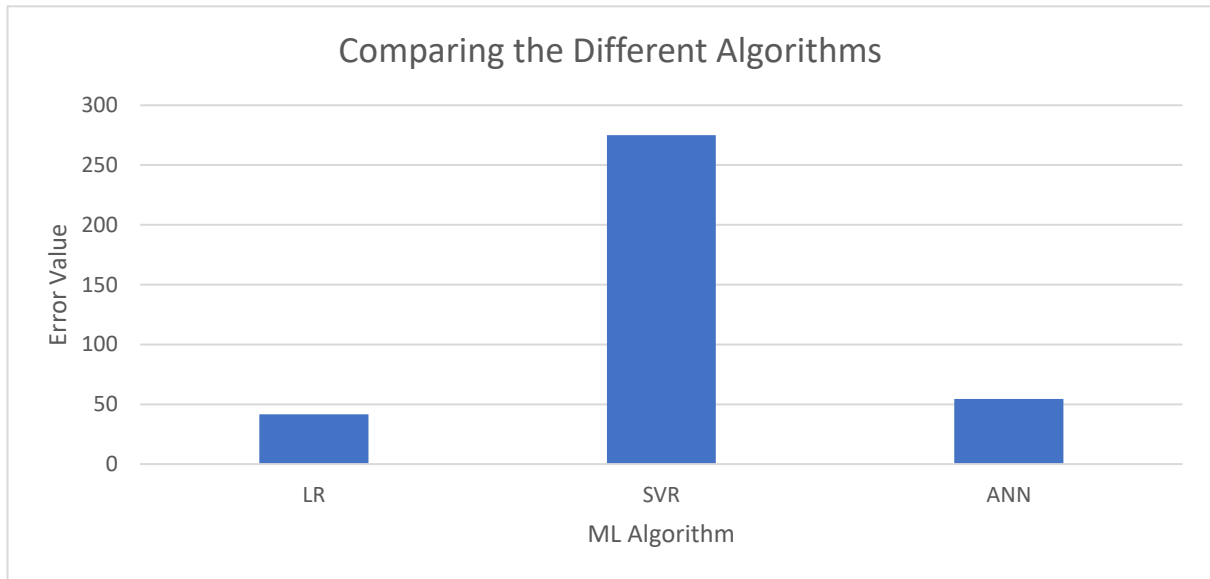


Figure 14: Comparing the different ML algorithms where the error values have been summed up and averaged for each algorithm. This includes all of the error metrics (MSE, RMSE, MAE and MAPE).

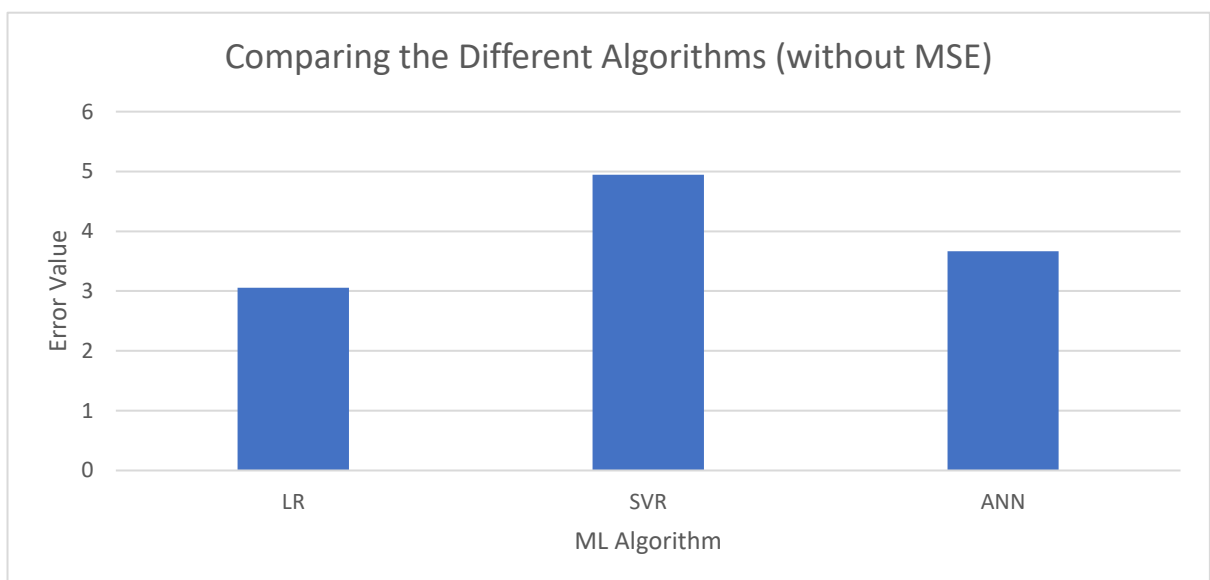


Figure 15: Comparing the different ML algorithms where the error values (without MSE) have been summed up and averaged for each algorithm. This includes three of the error metrics (RMSE, MAE and MAPE).

Discussion

RQ 1 - How do the different sectors within the stock market compare when it comes to ease of prediction?

Firstly, to understand this question we must look at the different sectors as a whole. We cannot be looking at the individual stocks within each sector or the different error values after different forecasted days into the future. To do this, we have summed up all the error values for all the stocks, over all the forecasted days and then worked out the average. We then came up with a singular value for each evaluation metric. If we look at figure 1, our first sight goes to the very high MSE values for two of the sectors: Consumer Discretionary and Communication Services.

As we know, MSE values are relative in a sense that you cannot pick out a MSE value and then compare to other MSE values outside of the set of results. We can only compare these values within the results produced. As we see, we have very large values for those two sectors showing that they produce the highest MSE values. From this alone we can make a judgment that for these two sectors, predicting them accurately can be rather difficult compared to some of the other sectors.

Due to the MSE values being very large, figure 1 does not give much other useful information. We can also see the sectors Real Estate, Information Technology, Industrials and Financials have higher MSE values compared to the others but making any other judgements from this chart is challenging.

One solution to this problem is removing the MSE value from the chart and looking at the other evaluation metrics and comparing them. The concept is still the same of having the evaluation metrics summed up and averaged over all the forecasted days. In figure 2, we can again see the Communication Services and the Consumer Discretionary sectors have the highest RMSE value. Of course, this value will be in relation to MSE as it is the root of it however, we also notice the higher MAE value for these two sectors.

We can see the highest MAE value is for Consumer Discretionary of around 13.13 and Communication Services has a MAE value of ≈ 9.28 . The next highest MAE value is of the Industrials Sector which is ≈ 3.85 . This again shows the difference between the former two sectors compared to the remaining sectors. Due to this, we now know that both MSE, RMSE and MAE show us that the sectors Consumer Discretionary and Communication Services are difficult to predict compared to the others.

Using the same chart, we can't compare the different MAPE values due to their very low numbers. MAPE are percentage values so they will range from 0-1 only. To closely analyse the individual evaluation metrics for each sector, we created different charts for each metric (MSE, RMSE, MAE, MAPE).

Figure 3 is a chart of MSE values only. Again, due to the ridiculously high values of the first two sectors, it is very difficult to compare the other sectors so in figure 4, we removed the first two sectors. This made the difference between the MSE values more visually clear. In this chart, we see that the Financials sector has the highest MSE value at ≈ 76.34 and the lowest MSE value was that of the Utilities sector at ≈ 5.18 . This means there is 15 times more risk in investing in the Financials sector compared to the Utilities sector in accordance with MSE.

Figure 5 then shows us again how drastically higher the first two sectors are. In this chart, RMSE values are being compared for each sector. The difference may not be as much compared to MSE but that is due to the squared function being exponential and the root in RMSE lowers the impact of larger MSE values. The trend, however, is the same between RMSE and MSE. The Utilities sector again has the lowest value.

Looking at the MAE value chart, figure 6, we see how Consumer Discretionary has the highest error value, closely followed by the Communication Services sector. The Utilities sector has the lowest MAE value at ≈ 1.20 . This means that it is around 11 times more difficult to predict the stock prices for the stocks in the Consumer Discretionary sector (sector with highest MAE value) than the Utilities sector.

Another evaluation metric we have gotten the results for is MAPE. These results were interesting compared to the other three metrics. In figure 7, we see that all the sectors give a result much more tightly packed compared to the previous charts. The average MAPE value is ≈ 0.03316 . The highest value for MAPE is from the Energy sector which is ≈ 0.0417 and the lowest is again the Utilities sector giving ≈ 0.0230 .

From all of this, we have a clear answer on which sector seems to be the easiest to predict stock prices. For all evaluation metrics, we see that overall after the values have been summed for all forecasted days and then averaged, we get the lowest error values for the Utilities sector. We know about the stocks in this sector and that they are majorly companies for electric, gas, water and forms of power.

Using our results, we see that these stocks go up and down in a very stable manner compared to some other stocks from different sectors. One of the reasons for this can be that in most cases these companies are protected by the government regulations. This means that other companies which try to emulate these companies already in the market find it very difficult to meet these regulations and in essence this decreases any competition for these companies by a large amount.

According to Investopedia [25], the most popular type of investors into this sector are income investors. These investors try to find companies to invest in that pay dividends. Dividends are a small portion of the profit the company gets. Quarterly, dividends are paid to shareholders. The companies within this sector usually outyield the companies from other sectors when it comes to dividends and this is again due to the stability of these stocks, having no major competition in most cases. This allows the stock to increase in value in a very stable manner. The demand for these utility stocks does not change too much, even during economic cycles and recessions, making them much steadier.

Having a sector which is easiest to predict has its benefits however, there are also some disadvantages. When a stock is stable, it becomes more easier to predict but a stable stock can mean that the price is going up or down at a slow rate. There are no quick jumps which go up and the overall profit an average investor can make from these stocks is lower than what they could if they invested in some other, less stable stocks. Another thing to note is that these results only give details on how easy it is to predict the stock. The prices of these stocks can also be decreasing at a steady rate, making it easy to predict.

The results are in fact in line with the above theory which shows us that the algorithms did work to some degree of accuracy.

RQ 2 - Which algorithm produced the best predictions?

As mentioned before, three different algorithms have been used to build a model for stock price prediction. The three models are LR, SVR and ANN.

Firstly, to see how these models worked, we can look at each individual sector and see the error across each of the forecasted days. We see in figure 8, which is for the Utilities sector, that as the forecasted days increase, there is an increase in error too. For this figure, it is the MSE value but this trend is visible for all of the other metrics too for most of the sectors (Figure 9 and 10). This supports the theory for the prediction as of course when you are predicting further into the future, the accuracy of the prediction should decrease and the error should increase, which is shown by the results.

Figure 11 is a RMSE chart for SVR for the Consumer Discretionary sector. Remember, this sector was giving ridiculously high errors for most of the metrics. This chart shows a very unusual trend where the error across all the forecasted days is high. There is no exponential trend, unlike most of the other sectors. This goes to show that the SVR had issues predicting stocks in this sector as the RMSE error value for only 1 day forecasted is ≈ 47.205 and for 30 days forecasted is it ≈ 48.658 . This is showing how close these values are to each other. What is also interesting is that the error for forecasted days 3, 4, 5, 10 and 15 are all lower than for 1 day. This could either be due to outrageous number within the data set or an error with the algorithm.

What we must bear in mind is that this is only for the SVR model. Looking at figure 12 and 13 for the same sector, using the same evaluation metric, the graph trends look similar to those previously discussed.

To answer the research question, we amalgamated the evaluation metrics for each algorithm for all of the forecasted days, summed them up and then averaged them across all the sectors. In the end we are left with three values, corroborating to each algorithm.

Figure 14 shows us the total error value for each algorithm. What we can see is that SVR has a value of ≈ 274.96 , LR has a value of ≈ 41.54 and ANN has a value of ≈ 54.5 . This shows us clearly that SVR worked the worst, and LR came up on top with the lowest error values. SVR gave a value ≈ 6.6 times larger than that of LR and ≈ 5.05 times larger than that of ANN. From previous figures we see that MSE gave a very high value and this could be affecting the results here. Even though we used MSE to answer our research questions, it may be a good idea to remove that and see if the results still have a similar trend when it comes to picking the best algorithm for prediction.

Looking at figure 15 we see how even without MSE, we see there is still a difference and the results follow the same as that of figure. The difference between SVR and the other two algorithms has been reduced where SVR the error value of SVR is less than 2 times compared to both LR and ANN.

LR comes out on top with the lowest value of error out of all three algorithms which is rather surprising at first look due to the more advanced algorithms within SVR and ANN. But then if you consider the larger picture, LR is in essence a line of best fit through the dataset and due to the large number of points, a straight line may actually have the least error. Having a line which is polynomial and bends through the data, like that of SVR and ANN, can result in a prediction value further away from the actual point. For example, if the model predicts that the price will increase for the next day, the line will bend upwards, but then in reality the price dropped, this will show a larger error compared to a line of best fit which took the middle point.

The image below shows the SVR MSE value for Amazon, Tesla Inc and Home Depot which are all a part of the Consumer Discretionary sector:

Forecasted Days	AMZN	TSLA	HD
1	72345.43	1029.878	6.543546
2	85195.19	804.1668	13.36034
3	64472.38	718.1016	17.74914
4	57271.03	542.8592	20.30473
5	49683.94	1187.619	46.25607
10	53970.76	1011.613	47.87049
15	62004.2	1449.14	57.14601
30	56292.41	1676.383	147.9563

Figure 16: Taken from Results sheet on Microsoft Excel which compares the raw values for each day forecasted. This table shows the SVR MSE values for AMZN, TSLA and HD.

What we can get from this data is that the algorithm itself does not produce the large values as we can see the low values for HD, but the stock's volatility itself. To further understand this, we can look at the charts for these stocks:



Figure 17: Amazon stock price from 2010 to present taken from Yahoo Finance online.

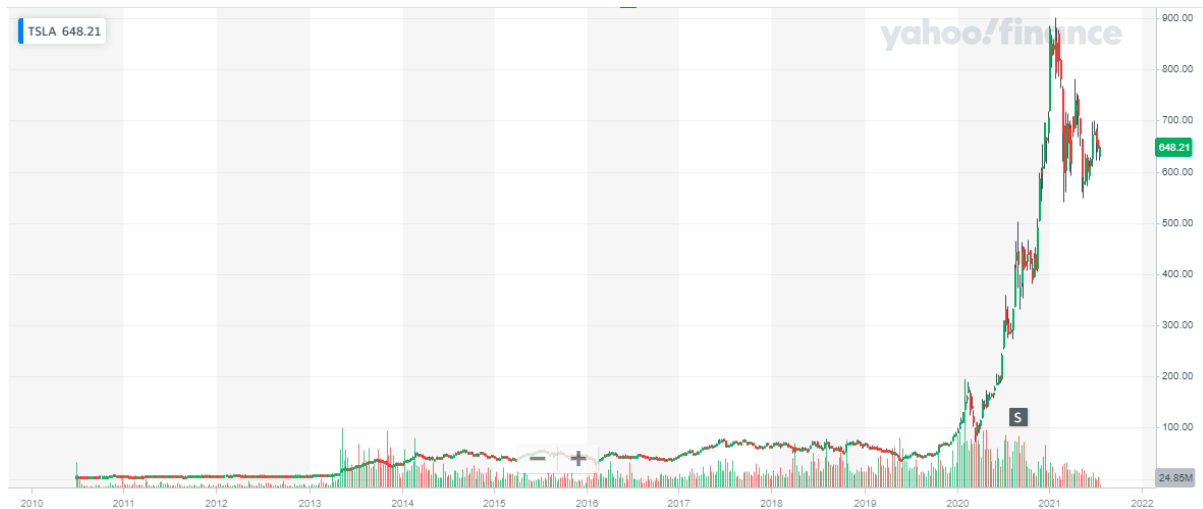


Figure 18: Tesla stock price from 2010 to present taken from Yahoo Finance online.



Figure 19: Home Depot stock price from 2010 to present taken from Yahoo Finance online.

Our first conclusion from this we can get is that looking at Tesla up until 2020, the curve was very flat in a relative sense. But as soon as 2020 hit, the curve just went higher and higher. Our models were learning from a date in 2010 up until a date in May 2021. What this does is that model only has a smaller dataset to learn the quick rise and is already trained on the slower more stable and flat stock prices. That large rise leaves the model predicting values which are way off, giving larger errors.

Looking at the Amazon stock price, the curve was relatively flat from 2010 to 2013 and then there is gradual increase in the stock price up until late 2018 where there is a larger drop. Now this drop will again cause issues to the algorithm because it would not have dealt with a drop like this before in the dataset. After this drop there is then another rise from the start of 2019 and the stock prices fluctuate from here until early 2020. There is again a small drop but then a rapid rise. This rise is in

line with effect of Covid on the market. Amazon, as we know, is an online shop. Due to coronavirus, people preferred to buy their items online compared to going out. This may be due to personal preference of government guidelines, resulting in Amazon becoming rapidly more popular within households. However, again this causes issues to the algorithm as there is only a set which the model can be trained on.

The final stock we are looking at here is Home Depot, where there is gradual rise from 2010 to present. Yes, there are some drops, again, the biggest one being in line with the Coronavirus outburst in early 2020, however in comparison, the stock price increases in a very linear fashion. After the drop in early 2020, the stock price returns to its price before the coronavirus rather quickly, meaning the drop did not have that much of an effect on the algorithm itself. This then results in smaller error value when predicting.

RQ 3 - Does the effect of Covid-19 play a part in stock prices of companies?

As briefly discussed in the section above, we can see that covid did have an effect on some companies. The degree to which this effect was felt by companies was very different however. As we saw with Amazon, it incurred a small dip and then rapidly rose. Some other companies unfortunately did not recover.



This chart is for an Airline company. Due to the outburst of the virus, there was a ban on travel to and from many countries. This meant that a lot of air travel companies took a hit.

Delta Air Lines were one of those companies which took a hit. We can see the rapid decrease in stock price around February/March 2020.

By looking at the stock price before the virus, it was fluctuating around 60USD. The drop made it go down to 20USD. This meant that the value of the company was down by 3 times.

Even now, the stock price is around 39USD. This goes to show that after a year and a half of the virus, the company has still not recovered to the price before.

Figure 20: Delta Air Lines stock price taken from Yahoo Finance online.



This chart is for American Water Works Company. Now, we see that when the virus outburst occurred, there was a drop in the stock price, but that was only due to a lot of investors panicking and selling. The business itself was not heavily impacted on by the virus.

This led to investors buying back into the company rapidly. Also, as soon as the drop happened, investors saw this as an opportunity to buy into the company as they knew it will not be damaged by the effect of the virus, making it a stock which was in essence on a discount.

We see just a few months after the drop, the stock price was back on track and increasing at a steady rate.

Figure 21: American Water Works Company stock price taken from Yahoo Finance online.

There can be many more examples of this, but in short, we see that many companies were affected by the virus. The depth of the effect varied from company to company. Some were affected in a positive way and some in a negative way. What we can derive from this is that investing into a single stock and holding high hopes for a single stock can be a bad strategy for investment. Having a more diversified portfolio will help the investor get past these small losses in some stock companies.

Conclusion

This study focused on the ease of prediction of different sectors where the main objective was to find a sector which is the easiest to predict stock prices for. We looked at a number of different algorithms to use which predicted the price of stocks.

The first research question was mainly regarding the main objective of finding a sector which is easiest to predict. After using a variety of algorithms and evaluation metrics, we came up with an answer to the question. We saw that across the board, the Utilities Sector had the lowest error values. We also saw that two sectors, Communication Services and Consumer Discretionary, had very high error values compared to all the other sectors. The remaining sectors were all quite similar, giving very similar error values for the different algorithms used too.

We understood that the reason for Utilities being the easiest to predict was due to the nature of the stocks present in this sector. The stocks were very reliable and competition was limited due to these stocks being endorsed by local authorities and governments, giving these companies the upper hand. Also, the fact that these companies are not affected much by economical issues and recessions, they make very stable stocks with the stock prices easier to predict compared to some other companies outside of this sector.

The second research question focused on the different algorithms and which algorithm was the best to use. What we found was that LR came up on top. SVR was the worst at predicting. This was surprising due to the complexity of SVR and ANN but due to the stocks chosen, LR suited best as it gave the line of best fit for the datapoints, which seemed to give the lowest error values. This points towards the idea that the stock prices were in fact more volatile than expected, and maybe the SVR and ANN needed some more training.

The final research question was to do with the Coronavirus impact on companies within the stock market. We looked at how some companies were affected. WE also saw that a majority of companies were affected however the degree to which they were affected varied. Some companies had a better response and some had a worse.

Limitations / Future Works

One of the limitations was that we only used historical data to come up with conclusions and results. When investing, there is a lot more to take into consideration such as social media influence, the current change of leadership of a company, the company structure, and these were of course neglected in this study. Also, an event like coronavirus could be incorporated so the model would understand why this dip happened for many companies and may disregard the few months in which the company was struggling due to worldwide pandemic etc.

Another limitation is that we only used the top 10 stocks to represent the whole sector when there can be hundred of different stocks within the sector itself. This was again due to convenience and due to the fact that the machine would not be able to efficiently produce results for such a large number of stocks. In the future, more stocks could be used or a different variety of stocks. The top 10 stocks would only represent the best stocks in the market and some of the other stocks may have a large influence on the sector as a while yet they were not included within this study.

Some of the algorithms were not optimised completely. This was due to algorithm optimisation being out of the scope of the project where only some aspects of optimisation were used. This especially affected SVR and ANN as they both are able to be optimised to a much higher degree than what was currently used. To optimise for this study, a mix of previous work was used to find the best blend of metrics for prediction. In the future, more time and care should be taken into optimising the algorithms as I am sure this will yield much better results.

The dataset used was from the 1st of January to 10th May. Most of the stocks were able to give data for this time frame however there were some who did not. Some of the newer companies did not exist on the stock market at the time so they had fewer data points. In the future this could be altered too. Maybe broken down into sections so that large dips do not affect the overall performance of the algorithm or even the range being increased.

Another piece of work which can be derived from this study is to create algorithms which group stocks into ease of predictability. Instead of having pre-determined groups such as the sectors, the algorithm itself would group individual stocks and create new groups which can be a better way of investing for investors. For example, stocks from the Utility and Health Care sector could be merged into a new group if they have the lowest error values.

Appendix

Python Code

SVR and LR (svr_fb.py)

This code is for the SVR and LR algorithms with an example of the Information Technology sector:

```
import numpy
import quandl
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.model_selection import train_test_split, cross_val_score,
KFold
import pandas as pd
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_absolute_percentage_error
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import mplfinance as mpf
import os

# Get the stock data

sectorname = 'Information Technology'
stockname =
['AAPL', 'ADBE', 'CSCO', 'INTC', 'MA', 'MSFT', 'NVDA', 'ORCL', 'PYPL', 'V']

for s in stockname:
    pathname = f'dataset/Sectors/{sectorname}/{s}.csv'
    df = pd.read_csv (pathname)

    # Grab the filename from the path
    filename = os.path.basename(pathname)
    filename_grab = os.path.splitext(filename)[0]

    # Get the Adjusted Close
    df = df[['Adj Close']]

    # A variable for predicting 'forecast_out' days out in future
    for i in [1,2,3,4,5,10,15,30]:
        forecast_out = i
        # Create another column (dependent variable) shifted 'forecast_out'
units up
        df['Prediction'] = df[["Adj Close"]].shift(-forecast_out)
        print(f'{sectorname} {s} {i}')
    # Create the independent dataset
    # Convert the dataframe to numpy array
    X = np.array(df.drop(['Prediction'],1))
    # Remove the last 'forecast_out' rows
    X = X[:-forecast_out]

    # Create the dependent dataset
    # Convert the dataframe to numpy array (All of the values including the
NaNs)
    y = np.array(df['Prediction'])
    # Get all of the y values except the last 'forecast_out' rows
    y = y[:-forecast_out]
```

```

# Split the data into 80% training and 20% testing
x_train, x_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)

#K-Fold
cv_kfold = KFold(n_splits=5, random_state=1, shuffle=True)

# Create and train the Support Vector Machine (Regressor)
svr_rbf = SVR(kernel='rbf', C=1e3, gamma=0.1)
svr_rbf.fit(x_train, y_train)
cv_scores_svr = cross_val_score(svr_rbf, X, y, cv=cv_kfold)
kfold_final_svr = cv_scores_svr.mean()
print(f'Final KFold Average for SVR: {kfold_final_svr}')

# Create and train the Linear Regression model
lr = LinearRegression()
# Train model
lr.fit(x_train, y_train)
cv_scores_lr = cross_val_score(lr, X, y, cv=cv_kfold)
kfold_final_lr = cv_scores_lr.mean()
print(f'Final KFold Average for LR: {kfold_final_lr}')

print('\n')

#Set x_forecast equal to the last forecast_out rows of the original dat
set from Adj. Close column
x_forecast = np.array(df.drop(['Prediction'],1))[-forecast_out:]

# Print the SVR prediction for the next forecast_out days
svm_prediction = svr_rbf.predict(x_forecast)
print(f'Prediction of the Adj. Close price for the next
{forecast_out} day(s) by the SVR model: \n{svm_prediction}')

# Print the Linear Regression prediction for the next forecast_out days
lr_prediction = lr.predict(x_forecast)
print(f'Prediction of the Adj. Close price for the next
{forecast_out} day(s) by the LR model: \n{lr_prediction}')
print('\n')

# Testing Model: Score returns the coefficient of determination R^2 of
the prediction
# Best possible score is 1.0
# svm_confidence = svr_rbf.score(x_test, y_test)
# print("svr confidence: ", svm_confidence)

# Testing Model: Score returns the coefficient of determination R^2 of
the prediction
# It provides a measure of how well observed outcomes are replicated by
the model, based on the proportion of total variation of outcomes explained
by the model
# Best possible score is 1.0
# lr_confidence = lr.score(x_test, y_test)
# print("lr confidence: ", lr_confidence)

print('\n')

## Creating predicted values for both models
y_pred_svr = svr_rbf.predict(x_test)
y_pred_lr = lr.predict(x_test)

```

```

# Mean Squared Error for SVR
mse_svr = mean_squared_error(y_test, y_pred_svr)
print(f'The MSE for the SVR algorithm is: {mse_svr}')

# Mean Squared Error for LR
mse_lr = mean_squared_error(y_test, y_pred_lr)
print(f'The MSE for the LR algorithm is: {mse_lr}')
print('\n')

# Root Mean Squared Error for SVR
rmse_svr = mean_squared_error(y_test, y_pred_svr, squared=False)
print(f'The RMSE for the SVR algorithm is: {rmse_svr}')

# Root Mean Squared Error for LR
rmse_lr = mean_squared_error(y_test, y_pred_lr, squared=False)
print(f'The RMSE for the LR algorithm is: {rmse_lr}')
print('\n')

# Mean Absolute Error for SVR
mae_svr = mean_absolute_error(y_test, y_pred_svr)
print(f'The MAE for the SVR algorithm is: {mae_svr}')

# Mean Absolute Error for LR
mae_lr = mean_absolute_error(y_test, y_pred_lr)
print(f'The MAE for the LR algorithm is: {mae_lr}')
print('\n')

# Mean Absolute Percentage Error for SVR
mape_svr = mean_absolute_percentage_error(y_test, y_pred_svr)
print(f'The MAPE for the SVR algorithm is: {mape_svr}')

# Mean Absolute Percentage Error for LR
mape_lr = mean_absolute_percentage_error(y_test, y_pred_lr)
print(f'The MAPE for the LR algorithm is: {mape_lr}')

n_array = numpy.array([])
new_array = numpy.append(n_array, [kfold_final_svr, kfold_final_lr,
mse_svr, mse_lr, rmse_svr, rmse_lr, mae_svr, mae_lr, mape_svr, mape_lr])
# print(new_array)

new_df = pd.DataFrame (new_array)
## save to xlsx file

filepath =
f'results/Sectors/{sectorname}/{forecast_out}Days/{filename_grab}_results.x
lsx'

new_df.to_excel(filepath, index=False)

```

ANN (svr_ann_collab.ipynb)

This code is for the ANN algorithm with an example of the Real Estate sector:

```
pip install mplfinance
pip install -U scikit-learn

# import quandl
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.model_selection import train_test_split
import pandas as pd
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_absolute_percentage_error
from sklearn.datasets import make_classification
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import mplfinance as mpf
import keras
from keras.models import Sequential
from keras.layers import Dense
import os

from google.colab import drive
drive.mount('/content/gdrive')

sectorname = 'Real Estate'
stockname = ['CCI', 'DLR', 'EQIX', 'PLD', 'PSA', 'SBAC', 'SPG', 'WELL']

for s in stockname:
    pathname = (f'/content/gdrive/My Drive/Colab
Notebooks/dataset/Sectors/{sectorname}/{s}.csv')
    df = pd.read_csv (pathname)
    filename = os.path.basename(pathname)
    filename_grab = os.path.splitext(filename)[0]

    df = df[['Adj Close']]

    # A variable for predicting 'forecast_out' days out in future
    for i in [1,2,3,4,5,10,15,30]:
        forecast_out = i
        # Create another column (dependent variable) shifted 'forecast_out'
units up
        df['Prediction'] = df[["Adj Close"]].shift(-forecast_out)

        print(s,i)
        # Create the independent dataset
        # Convert the dataframe to numpy array
        X = np.array(df.drop(['Prediction'],1))
        # Remove the last 'forecast_out' rows
        X = X[:-forecast_out]
        #print(X)
        y = np.array(df['Prediction'])
        # Get all of the y values except the last 'forecast_out' rows
        y = y[:-forecast_out]

        x_train, x_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)
        x_forecast = np.array(df.drop(['Prediction'],1))[:-forecast_out:]
```

```

ann_relu = Sequential()
ann_relu.add(Dense(200, input_dim=1, activation='relu'))
ann_relu.add(Dense(200, input_dim=200, activation='relu'))
ann_relu.add(Dense(200, input_dim=200, activation='relu'))
ann_relu.add(Dense(1, activation='linear'))

keras.optimizers.Adam(lr=0.1, beta_1=0.9, beta_2=0.999, amsgrad=False)
ann_relu.compile(loss='mean_squared_error', optimizer='RMSprop',
metrics=['mean_absolute_percentage_error'])

ann_relu.summary()

history = ann_relu.fit(x_train, y_train, epochs=200,
batch_size=32, validation_split=0.15, verbose=1)

y_pred_ann = ann_relu.predict(x_test)

mse_ann = mean_squared_error(y_test, y_pred_ann)
print(f'The MSE for the ANN algorithm is: {mse_ann}')

rmse_ann = mean_squared_error(y_test, y_pred_ann, squared=False)
print(f'The RMSE for the ANN algorithm is: {rmse_ann}')

mae_ann = mean_absolute_error(y_test, y_pred_ann)
print(f'The MAE for the ANN algorithm is: {mae_ann}')

mape_ann = mean_absolute_percentage_error(y_test, y_pred_ann)
print(f'The MAPE for the ANN algorithm is: {mape_ann}')

n_array = np.array([])
new_array = np.append(n_array, [mse_ann, rmse_ann, mae_ann, mape_ann])
print(new_array)

new_df = pd.DataFrame(new_array)

filepath = f'/content/gdrive/My Drive/Colab
Notebooks/results/Sectors/{sectorname}/{forecast_out}Days/{filename_grab}_a
nn_results.xlsx'
new_df.to_excel(filepath, index=False)

keras.backend.clear_session()

```

Excel Merge (merge.py)

This was used to grab all the results and merge into one excel file:

```
from os import listdir
from os.path import isfile, join
from glob import glob
import openpyxl
from openpyxl import Workbook

def get_all_value_excel(file_path):
    wb_obj = openpyxl.load_workbook(file_path)
    sheet_obj = wb_obj.active
    values = []
    for i in range(1, sheet_obj.max_row+1): # Ignore first row 0 value
        cell_obj = sheet_obj.cell(row=i, column=1)
        value=cell_obj.value
        if type(value)==type(0) and value==0: # ignore first cell if its 0
            continue
        values.append(value)
    wb_obj.close()
    return values

def get_multiple_excel_values(file_paths):
    values=[]
    for path in file_paths:
        values+=get_all_value_excel(path)
    return values

folder_names=[]
selector='Sectors'
excel_file_map={}
i=1
while True:
    folder_name=input(f"Pleasse enter {i} folder relative path : ")
    if folder_name.strip() != '':
        folder_names.append(folder_name+'\\Sectors')
    else:
        continue
    i+=1
    more_folders=input("Do you have more folders (Y/N) ? ")
    if more_folders.strip().lower()=='y':
        continue
    else:
        break

for folder in folder_names:
    sectors=glob(folder+'*/') # get sub folders of Sector. * means to get
    matching folders
    sectors=[s.replace('\\', '/') for s in sectors]
    for sector in sectors:
        sector_name=sector.split('/')[-2]
        sector_folders=glob(sector+'*/') # get day folders in side sector
        sector_folders = [s.replace('\\', '/') for s in sector_folders]
        for day_folder in sector_folders:
            print(day_folder.replace('\\', '/'))
            print(day_folder)

        days=int(day_folder.split('/')[-2].lower().replace('days',''))
# get days
        onlyfiles = [join(day_folder, f) for f in listdir(day_folder)
if isfile(join(day_folder, f))] # files in days folder
```

```

        print(onlyfiles)
        for file in onlyfiles:
            file_name=file.split('/')[-1]
            if ord(file_name[0])<65 or ord(file_name[0])>122: # if
excel open a file it creates temp file. to skip that
                continue
            if '.xlsx' in file_name and '_test.xlsx' not in file_name:
# remove non excel files and excel files with test postfix
                stock=file_name.split('_')[0].upper()
                if sector_name in excel_file_map.keys():
                    if days in excel_file_map[sector_name].keys():
                        if stock in
excel_file_map[sector_name][days].keys():

excel_file_map[sector_name][days][stock].append(file)
                        else:

excel_file_map[sector_name][days][stock]=[file]
                        else:

excel_file_map[sector_name][days]={stock:[file]}
                        else:
                            excel_file_map[sector_name]={days:{stock:[file]}}

print(f"File map : {excel_file_map}")
wb = Workbook()
sheets = list(excel_file_map.keys())
sheets.sort()
wb.active.title=sheets[0]

i=0
for sheet in sheets:
    days=list(excel_file_map[sheet].keys())
    days.sort()
    if i==0:
        ws=wb.active
    else:
        ws=wb.create_sheet(sheet)
    i+=1
    stock_column_map={}
    last_column=65
    row=2 # keep track of start of day should be row
    for day in days:
        stocks=list(excel_file_map[sheet][day])
        stocks.sort()
        for stock in stocks: # make header columns
            if stock not in stock_column_map.keys():
                ws[f'{chr(last_column)}1']=stock
                stock_column_map[stock]=chr(last_column)
                last_column+=1
        max_values=0
        for stock in stocks:
            temp_row = row
            print(f'Processing Files {excel_file_map[sheet][day][stock]}')

values=get_multiple_excel_values(excel_file_map[sheet][day][stock])
            for value in values:
                ws[f'{stock_column_map[stock]}{temp_row}']=value
                temp_row+=1
            if max_values<len(values)+1+row:
                max_values=len(values)+1+row

```

```
row=max_values  
wb.save('test.xlsx')
```


Link

To access the code for all of the above, alongside raw data files and datasets, please visit the GitHub repository:

<https://github.com/kashfay110/stock-market-python.git>

Raw Data

To access the raw data files, you can access them from the results and results-ann folder in the following GitHub repository:

<https://github.com/kashfay110/stock-market-python.git>

To access the results Excel file which contains all the results merged together, you can access it here:

<https://github.com/kashfay110/stock-market-python/blob/d70f6535b0ff56243941f9659ec734fbb1772bd/All%20Results%20Final.xlsx>

Forms

Ethics Form



RESEARCH ETHICS AND INTEGRITY

RISK ASSESSMENT FORM

FORM A

For all undergraduate Dissertations and Research Projects

FOR OFFICE USE ONLY

To be completed by module leader/supervisor [subject to confirmation by SCREP]

In the opinion of the module leader/supervisor this application falls into

CATEGORY 1 []

CATEGORY 2 []

Please fill in this form, then SAVE IT AS A PDF and submit as instructed by your supervisor with your project proposal

Your name: Kashfay Naqvi

Student number: 21389466

Your email address: 21389466@student.uwl.ac.uk

Name of supervisor: Dr Scott Yang

Title of project: Analysis of different Stock Market Indexes and their ease of prediction

Date: 11.12.2020

SECTION A**PROJECT DESCRIPTION**

Please answer the following questions:

1. Do you intend to involve human participants in the conduct of your research?
If no, please skip questions 1a & 1b.

☐ Yes

☒ No

1a. Does your research involve vulnerable adults (who are or may be for any reason unable to take care of themselves, or unable to protect themselves against significant harm or exploitation) or under-18s?

☐ Yes

☐ No

1b. Could your research potentially expose you, anyone assisting you, or participants to physical, psychological and/or emotional harm? (see Section B, Question 9)

☐ Yes

☐ No

2. Will your research involve travelling to geo-politically unstable regions/countries (e.g. areas affected by war, civil unrest, natural disasters, or listed as inadvisable to travel by the UK government)?

☐ Yes

☒ No

3. Will your research involve access to security-sensitive material? (see the University's Research Ethics Code of Practice 2018 for a definition of security-sensitive materials and Section B, Question 9 of this form)

☐ Yes

☒ No

This proposal *must be completed with the assistance of your supervisor/module leader/tutor*. You can change the size of the boxes (below) by typing or deleting as necessary.

It is very important to convey with clarity:

- Your research questions/the problem/the theme or topic you are investigating (what you are proposing to do and to find out or to create)
 - The methodology or technical approach (for projects comprising in whole or in part the creation of an artefact) you will adopt – methods, number of participants, who the participants (if any) will be, survey instruments used, technology and equipment employed etc.; and what questions you are planning to ask your respondents (if applicable); how you will deal with technical challenges.
-

SECTION B

Only complete if you answered YES to Q1 in Section A.

	WHERE APPROPRIATE TO YOUR CHOSEN TOPIC/RESEARCH:	YES	NO	N/A
1	Will you describe in writing the main procedures to participants in advance, so that they are informed about what to expect? A copy of this must be attached to this application			
2	Will you tell participants that their participation is voluntary?			
3	Will you obtain written consent for participation and include within this that they have a right to withdraw at any point? A copy of this must be attached to this application			
4	If the research is observational, will you ask participants for their consent to being observed?			
5	With questionnaires, will you give participants the option of omitting questions they do not want to answer?			
6	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs? This should be evidenced in the consent form and (if applicable) with a signed copy of UWL's data management form, attached to this application.			
7	Will you debrief participants at the end of their participation (i.e. give them a brief explanation of the study)? A copy of this must be attached to this application			
8	Will your project involve deliberately misleading participants in any way?			
9	If you answered YES to Question 1b (section A) give details on a separate sheet and state what you will tell your participants to do if they should experience any problems (e.g. who they can contact for help).			
10	Do participants fall into any of the following vulnerable groups? If they do, please and tick box 2 overleaf. Note that you may also need to obtain satisfactory DBS clearance (or equivalent for overseas students).			
	Schoolchildren (under 18 years of age)			
	People with learning or communication difficulties			
	Patients			
	People in custody			
	People engaged in illegal activities (e.g. drug-taking)			
	Any other groups who could be reasonably argued as representing any form of vulnerability – please specify			

SECTION C



	WHERE APPROPRIATE TO YOUR CHOSEN TOPIC/RESEARCH:	YES	NO	N/A
11	Will you be accessing materials which may be considered security-sensitive under the Counter Terrorism Act (2015)?		X	
12	Does your project involve work with animals? If yes, please tick box 2 below.		X	

[Note: N/A = not applicable]

There is an obligation on the researcher to bring to the attention of the School Ethics Panel any issues with ethical implications not clearly covered by the above checklist.

PLEASE TICK EITHER BOX 1 OR BOX 2 BELOW AND PROVIDE THE DETAILS REQUIRED IN SUPPORT OF YOUR APPLICATION. THEN SIGN THE FORM.

Please tick

1. I consider that this project has no significant ethical implications to be brought before the School Ethics Panel.	<input checked="checked" type="checkbox"/>
-----------------------------------------------------------------------------------------------------------------------	--------------------------------------------

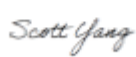
2. I consider that this project may have ethical implications that should be brought before the School Ethics Panel, and/or it will be carried out with children or other vulnerable populations.	<input type="checkbox"/>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------

I have received guidance on ethical research practices relevant to my subject as part of my preparation for this module.

Signed  Print Name: Kashfay Haider Naqvi

Date 11.12.2020

(UG Researcher(s))

Signed  Print Name: Dr Scott Yang

Date 11.12.2020

(Supervisor)

PROJECT OUTLINE

Your name: Kashfay Naqvi

Student number: 21389466

Your email address: 21389466@student.uwl.ac.uk

Name of supervisor: Dr Scott Yang

Title of project: Analysis of different Stock Market Indexes and their ease of prediction

Date: 11.12.2020

Introduction to the research

Background to research topic area (with references where applicable). 200 words approx. Include Aims and Hypothesis, Research Question(s) of the dissertation/project or an outline of the aims and the context of the creative artefact (where the project is a creative artefact)

The aim of the study is to investigate the ease of predictability of different Stock Market Indexes compared against each other.

Method

Outline all methodological issues for any projects requiring human participants and data collection - e.g. Design, participants, questionnaires, tests, method of data collection. *Who are you working with? How? What Measures? What interventions/manipulations? What controls?*



If your research entails exclusively the consultation of published documents, books, articles or other work in the public domain please state this under the heading 'materials'.

In the case of creative artefacts (such as audio-visual, audio or visual outputs or production such as a film, video, audio recording, composition, screenplay, piece of creative writing, performance, exhibition, screening, photograph, body of photographic work, painting, sculpture, installation, design or software) please use relevant sections of your dissertation/project proposal to place in the section below to which they most closely pertain.

Students and supervisors may also find it helpful to cross-refer to the Health and Safety clearance documents for any level 6 research projects which need to be completed by students *in certain fields*. Sections completed by students for the latter forms may be suitable to be repeated below.

Research design or schedule (for creative artefacts)

Participants, including (where applicable) collaborators in the making of creative artefacts

Materials (to include locations and objects/resources)

The materials used will be mainly articles regarding stock market prediction.

Procedure or details of technical aspects of creative production

To fit our idea, we will try to use hybrid methods of Machine Learning to predict the stock prices in different stock markets. We will then decipher which market gave use the best results, using performance metrics and other relevant metrics. The prediction of each markets will be compared and the market which is easiest to predict will be decided.

Analysis

Please complete this section only if your project requires written analysis to be submitted as the assessment or as part of it. If the project you are undertaking comprises a creative artefact such as a film or body of photographic work please type 'Not applicable' in this box

CLEARLY describe the method of analysis you are going to use. Is it *qualitative* or *quantitative*?

For students completing a Dissertation you should be able to refer to what you have learned in the Research Methods component of your study.

Project Supervisor Form



University of West London

School of Computing and Engineering

PROJECT MODULE

Module Code

CP6CS46E

Project Supervisor Form

Student Ref. No.
21389466

Submission Date
28.10.2020

Deadline Time

Student's Surname.
Naqvi

Student's Forename.
Kashfay Haider



Project Supervisor's Name.
Dr. Scott Yang

Project topic

The Analysis and Future Prediction of the Stock Market.

Comment by the supervisor

The provisional project topic is within the scope of machine learning and worth for further investigation.

	Signatures	Date
Student		26.10.2020
Supervisor		04/11/2020

Project Progress Form 1

University of West London



School of Computing and Engineering

PROJECT MODULE

Module Code

CP6CS46E

Project Progress Form 1

Student Ref. No. 21389466

Submission Date

Deadline Time

Student's Surname. Naqvi

Student's Forename. Kashfay

Project Supervisor's Name.
Dr Scott Yang

Please note: Work presented for the project must be the student's own. Plagiarism is where a student copies work from another source, published or unpublished (including the work of a fellow student) and fails to acknowledge the influence of another's work or to attribute quotes to the author. Plagiarism is an academic offence.

I confirm this is my own work

Student's signature

Project title

Analysis of the stocks of different sectors; comparing the ease of prediction of one sector from another.

Progress made since project was approved

Slight change in project hypothesis and methodology due to the nature of the idea. Also, the python code is nearing completion with only a few more additions that are required to be added. The writing of the final report has also slowly started to begin.

List of meeting dates with supervisor (to be completed by student)

4/12/2020, 11/12/2020, 8/01/2021, 15/01/2021, 22/01/2021, 29/01/2021, 05/02/2021, 12/02/2021, 19/02/2021, 26/02/2021, 04/03/2021. (Scheduled Meetings for every Friday unless prior agreement on change of date.)

Comments and further actions (to be completed by supervisor before submission)

The current progress is OK; hopefully will enter the experimental analysis stage in 1 or 2 weeks.

	Signatures	Date
Student		02/03/2021
Supervisor (before submission)	<i>Scott Yang</i>	02/03/2021

University of West London



School of Computing and Engineering

PROJECT MODULE

Module Code

CP6CS46E

Project Progress Form 2

Student Ref. No. 21389466

Submission Date
10.04.2021

Deadline Time

Student's Surname. Naqvi

Student's Forename. Kashfay

Project Supervisor's Name.
Dr Scott Yang

Please note: Work presented for the project must be the student's own. Plagiarism is where a student copies work from another source, published or unpublished (including the work of a fellow student) and fails to acknowledge the influence of another's work or to attribute quotes to the author. Plagiarism is an academic offence.

I confirm this is my own work

Student's signature

Project title

Analysis of the stocks of different sectors; comparing the ease of prediction of one sector from another.

Progress made since Project Progress Form 1

Python code is working correctly and have started to produce different results. Have started analysis on different stocks as well as some optimisation of the algorithms. Reading more articles to help with the final literature review on top of finalising the experimental points.

List of meeting dates with supervisor (to be completed by student)

4/12/2020, 11/12/2020, 8/01/2021, 15/01/2021, 22/01/2021, 29/01/2021, 05/02/2021, 12/02/2021, 19/02/2021, 26/02/2021, 04/03/2021, 12/03/2021, 19/03/2021, 26/03/2021, 9/04/2021. (Scheduled Meetings for every Friday unless prior agreement on change of date.)

Comments and further actions (to be completed by supervisor before submission)

Progress as planned till now.

	Signatures	Date
Student		09.04.2021
Supervisor (before submission)	Scott Yang	10/04/2021

References

- [1] M. Usmani, S. Adil, K. Raza, S. Ali, Stock market prediction using machine learning techniques, 2016.
- [2] Parmer, I., Agarwal, N., Saxena, S., Arora, R., Gupta, S., Dhiman, H. and Chouhan, L., 2018. *Stock Market Prediction Using Machine Learning*. [online] Ieeeexplore.ieee.org. Available at: <<https://ieeexplore.ieee.org/abstract/document/8703332>> [Accessed 23 July 2021].
- [3] A. Sharma, D. Bhuriya, U. Singh, Survey of stock market prediction using machine learning approach, 2017.
- [4] D. Madhusudan, Stock Closing Price Prediction using Machine Learning SVM Model, 2020.
- [5] Yuan, X., Yuan, J., Jiang, T. and Ain, Q., 2020. *Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market*. [online] Ieeeexplore.ieee.org. Available at: <<https://ieeexplore.ieee.org/abstract/document/8968561>> [Accessed 23 July 2021].
- [6] Lee, T., Cho, J., Kwon, D. and Sohn, S., 2018. *Global stock market investment strategies based on financial network indicators using machine learning techniques*. [ebook] ScienceDirect. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0957417418305761?casa_token=w2KL9Tc7dmIAAAAA:eHkAsWk6A2hOFCADLP5gqtmPy6ue1xnCRXzgJd0bmXn9nNi57qvIB_pAep00OqvNQ8wMkQ9tIPU> [Accessed 23 July 2021].
- [7] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T. and Sakurai, A., 2011. *Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction*. [online] Ieeeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/6118898?casa_token=IPigsCZrJA8AAAAA:iNmVx1B1XESbk2kJVBcMdbnGTEvfGJGj8kuxrx3-iYhH255HktUPUbX1qQl6S6ku45FrIPaKRXyMA> [Accessed 23 July 2021].
- [8] Atsalakis, G. and Valavanis, K., 2008. *Surveying stock market forecasting techniques – Part II: Soft computing methods*. [online] Available at: <<https://www.sciencedirect.com/science/article/abs/pii/S0957417408004417>> [Accessed 23 July 2021].
- [9] Botchkarev, A., n.d. *Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology*. [online] Arxiv.org. Available at: <<https://arxiv.org/ftp/arxiv/papers/1809/1809.03006.pdf>> [Accessed 24 July 2021].
- [10] QuestionPro. 2021. *Quantitative Data: Definition, Types, Analysis and Examples | QuestionPro*. [online] Available at: <<https://www.questionpro.com/blog/quantitative-data/>> [Accessed 24 July 2021].
- [11] J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock market index using fusion of machine learning techniques, 2015.
- [12] Bukszpan, D., 2012. *Industries Hit Hardest by the Recession*. [online] CNBC. Available at: <<https://www.cnbc.com/2012/06/01/Industries-Hit-Hardest-by-the-Recession.html>> [Accessed 24 July 2021].

- [13] Bischoff, B., 2019. *Adjusted Closing Price vs. Closing Price*. [online] Finance - Zacks. Available at: <<https://finance.zacks.com/adjusted-closing-price-vs-closing-price-9991.html>> [Accessed 24 July 2021].
- [14] Docs.python.org. n.d. *The Python Tutorial — Python 3.9.6 documentation*. [online] Available at: <<https://docs.python.org/3/tutorial/index.html>> [Accessed 24 July 2021].
- [15] Gupta, N., 2021. *Why is Python Used for Machine Learning? | Hacker Noon*. [online] Hackernoon.com. Available at: <<https://hackernoon.com/why-python-used-for-machine-learning-u13f922ug>> [Accessed 24 July 2021].
- [16] Techopedia.com. 2020. *What is Microsoft Excel? - Definition from Techopedia*. [online] Available at: <<https://www.techopedia.com/definition/5430/microsoft-excel>> [Accessed 24 July 2021].
- [17] Zhang, Y., 2010. *New Advances in Machine Learning*. IntechOpen.
- [18] Ibm.com. 2020. *What is Supervised Learning?*. [online] Available at: <<https://www.ibm.com/cloud/learn/supervised-learning>> [Accessed 24 July 2021].
- [19] Ibm.com. 2020. *What is Unsupervised Learning?*. [online] Available at: <<https://www.ibm.com/cloud/learn/unsupervised-learning>> [Accessed 24 July 2021].
- [20] Kavitha, S., Varuna, S. and Ramya, R., 2017. *A comparative analysis on linear regression and support vector regression*. [online] Ieeeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/7916627?casa_token=mv7LIPDtMTYAAAAA:XYIN3EOfpXgMnsJsLdxS0fiyy9XUR6Ogb46qGQmBPcCNU6xAoBxCfp2JsF_X4KsT_ZfwrkUIaoLBzw> [Accessed 24 July 2021].
- [21] Gandhi, R., 2018. *Support Vector Machine — Introduction to Machine Learning Algorithms*. [online] towardsdatascience.com. Available at: <<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>> [Accessed 24 July 2021].
- [22] Sethi, A., 2020. *Support Vector Regression In Machine Learning*. [online] Analytics Vidhya. Available at: <<https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>> [Accessed 24 July 2021].
- [23] Grogan, M., 2020. *Regression-based neural networks with TensorFlow v2.0: Predicting Average Daily Rates*. [online] towardsdatascience.com. Available at: <<https://towardsdatascience.com/regression-based-neural-networks-with-tensorflow-v2-0-predicting-average-daily-rates-e20ffa7ac9a>> [Accessed 24 July 2021].
- [24] Watson, N., 2012. *Using Mean Absolute Error for Forecast Accuracy*. [online] canworksmart.com. Available at: <<https://canworksmart.com/using-mean-absolute-error-forecast-accuracy/>> [Accessed 24 July 2021].
- [25] Ross, S., 2020. *What Kind of Investors Buy Utility Stocks?*. [online] Investopedia. Available at: <<https://www.investopedia.com/ask/answers/122314/what-kind-investors-buy-utility-stocks.asp>> [Accessed 24 July 2021].