

Stock Market Prediction Approach in United Kingdom

MSc Research Project

Sonali Pandhure

Student ID: x18137458

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing



Student Name:	Sonali Pandhure
Student ID:	x18137458
Programme:	MSc. Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	12/12/2019
Project Title:	Stock Market Prediction Approach in United Kingdom
Word Count:	5501
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission , to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Stock Market Prediction Approach in United Kingdom

Sonali Pandhure
x18137458

Abstract

Now a day's Stock market prediction has become an attractive and interesting thing for data analysts. In this research most of the models such as Time series and machine learning techniques are used for stock market prediction. London Stock market is the trending topic for all human beings and it will remain trending in future because each and every person has started investing in stocks. In this research London stock market data is analysed and predicted depending on Brexit discussion impact happened in month October 2019 on London stock market. Also Brexit is a very trending thing in UK and it has very high impact on parliament and on entire European unions. Methodologies are used for this London stock market predictions are ARIMA time series, Random Forest, logistic regression, multiple regression, and Naïve Bayes model. ARIMA is used for forecasting the 3 months stocks, Logistic regression is used to find impact of Brexit discussion and other model are used to predict the accuracy of the results.

Keywords: *Random Forest, Logistic Regression, ARIMA time series, Naïve Bayes, multiple regression etc.*

1 Introduction

Stock market is one of the most important element for economy because it provides access to the customer to purchase shares and ownership from the big budget companies. Stock market is the collection of market where other securities, bonds, and trending of equities are takes place. Stock market arena keeps developing constantly under refinement process. Depending on variation depending on every day, investors need to do each and every investment very carefully to gain the profit. Updated London Stock market research data is publicly available on London stock exchange website on daily basis. You can get the updated information at each second. This data is timely basis and gets updated on each second. Stock market forecasting or prediction is very difficult task because of timely basis changing data. This data always gives the technique which gives good stock exchange. motivation to predict and develop new forecasting model in return. Stock data and prices are not random points, these points are collected or treated as timely basis discrete model and this is depending of set of well-defined numerical data items which are collected at successive points on consistent interval time. Brexit discussion happened in October 2019 month and the prediction of this month is compared with other months. Depending on this prediction we can interpret the results. This Brexit discussion month and other month prediction comparison is shown by using logistic regression and next 3 months

prediction is shown by using ARIMA time series model. Also Naive Bayes, multiple regression model and Random forest models are applied to check accuracy of the data. Here R programming language is used to discover the predictions and statistical data which helps to give better predictions and visualizations. R language is gives access to predict the stock market data by using different models. Proposed research gives, London stock market predictions are implemented depending on Brexit discussion happened in the month of October 2019. Where it can be proved that this data is clearly declared and predicted by using R language. Also these predictions are giving better clarity to the users for their stock investments in different and trending companies. By using this visualizations and predictions we can give the better results or predictions of what will be the stocks if Brexit happens. This is a nice little introduction with some figure in Figure 16

1.1 Research Question

RQ: "To what extent can prediction of Brexit discussion (October 2019) using logistic regression machine learning technique used and also prediction of stock market using ARIMA time series (November, December and January 2020), Random forest, multiple regression and Naïve Bayes"

1.2 Objectives and Contributions

By considering the research question, following objectives are created.

Objectives 1: Data cleaning and removing unwanted rows and columns by using R language.

Objectives 2: Evaluation and Implementation of machine learning models for predicting impact of Brexit discussion on stock market.

Objectives 3: Comparison of developed prediction models and calculate ARIMA time series, Logistic regression, Random forest, multiple regression and Naive Bayes models.

Contribution

Major contribution to the results is ARIMA time series model, Logistic regression, Random Forest, multiple regression and Naïve Bayes model. All models are used to predict the stock market predictions. By using these models we can predict the 3 months future data and also different models for correct accuracy. Basically these models are used to detect the future stock market income prediction and others are used to predict the accuracy of the particular model. Logistic regression model is used to predict the data of October 2019 over other month's data for brexit discussion impact on stock market income. All models are implemented and resulted from objective two and objective three which gives high contribution in terms of predicting the stock income. Remaining technical report is listed as follow: Chapter 1 gives the detailed idea about stock predictions abstract, introduction, research question and objectives. Chapter 2 gives the detailed idea about technologies and models are used previously to predict stock market income and returns from year 2002 to 2019. Chapter 3 gives the idea about overall methodology and process flow. Next chapter gives the idea about implementation process where detailed idea is given in terms of models used, implementation, model accuracy, evaluation and results. Next chapter that is chapter 5 shows the concluded result by summarizing

each result depending on objectives listed above, in detailed discussion and future work for other people to give short idea what should happened in future and key to their next research project.

2 Literature Review of stock market prediction from year (2004 to 2019)

2.1 Introduction

In this given chapter, literature related overview is given in terms of stock market prediction opportunities and challenges faced during implementation. It stated different types of models, methodologies, and predictions which are submitted to the IEEE site. This literature survey/ review is further divided into different sections by considering types of models used which are: section 2.2 states challenges faced during predictions, section 2.3 states the ARIMA time series model which gives future prediction, section 2.3 stated different models and technologies used during predictions. Section 2.4 gives the critical review technique, methods to predict the time series model, problems faced and gaps identified.

2.2 Critical Review of Stock Market

Li et al. (2018) In this article Stock market prediction is given based on sentiments of polarities on news-driven, information sources, financial news, and social media. Stock market skewed news distribution gives few news with lower accuracy of predictions. In given tests firstly both target and source stocks mapping is done which is later on built by using sentiment dimensions. In second case 3 different scenarios are taken in consideration which are: target and sources should highly correlated, target and source in one sector and former news rich in other, and source stock has maximum prediction rate in terms of validated dataset. And third one is depending on principles the design of majority principles are implemented. In this research analysis, data retrieval and data extraction is done effects on stock market news. It includes growth of sentiment analysis distribution for evaluation of effects of news gauging models, development of dictionary based model (Sentiment analysis model), and financial sector model for the pharmaceutical stock market. Zhang et al. (2018) In previous research all the prediction has done by using one dataset but because only one dataset all features won't be covered fully. This research holds different dataset assets and by using this multiple instance models are created which combined sentiments, events, and quantitative data in compressed framework. Here Novel event extraction and representation method is applied and used. Soni et al. (2018) In this given research different modules like PSO, Naïve Bayes, Decision tree, and Black Hole techniques are used. After applying these algorithms some new natural techniques are used where accuracy calculation and comparison is done over multiple techniques and algorithms. Hence the salient features and proposed models are then created. Here big dataset is used to predict the outputs. Different predefined techniques are tested over same dataset.

Yoo et al. (2005) This research is based on a survey of stock market predictions. Basically the survey of challenging work of financial stock market predictions. In the given

research research discussion of disadvantages and advantages of stock market and its predictions are given. By combining prediction model with event information plays good role in terms of more accuracy. Because of stable automated extraction system and accurate event weighting method are necessary to give good enactment in financial stock market time series prediction. Liu and Wang (2019) Proposed research states the attention based method to effectively explore the complementary between numerical data and news while predicting the stock prices. Stock trend information is itself present in the news and then is transformed in the form of numerical distribution data. News are encoded with the numerical data. Proven method gives the full numerical data from the news.

2.3 A Critical Review of Techniques, Algorithms and Methodologies used for Stock Market Predictions

Sharaff and Choudhary (2018) Given research gives idea about different stochastic models like ANN , ARIMA model, RNN , and winters model to predict and analyze poor Mumbai stock exchange Sensex and new stock guides of standards. In given research Artificial neural network (ANN) model is compared with other models to make predictions. This model proved robust while predicting future stock market indices. Attigeri et al. (2015a) In given research different supervised machine learning techniques are applied for stock market predictions which overcomes difficulties like impact of factors on stock by which it becomes hard to predict the stock prices. 5 models such as Random forest, SVM , KNN, Softmax and Naïve Bayes are predicted and compared with each other to get better model and its predictions. By the experiment it's concluded that Random forest model has better accuracy than all other model and it gives more accurate predictions than others. After reducing half of the data Naïve Bayes model performed better than others. Kumar et al. (2018) In the given research research Taiwan stock exchange prediction is calculated recurrent neural network and dataset is trained by using ARIMA model analysis. Second difference data is used and gives better prediction if we train neural networks using it rather than raw data. Here at the time of back-propagation training and addition to the traditional modification term, more 2 prediction errors are stated in terms of adjustment of weight connections. Elagamy et al. (2018) Proposed research Random forest model is combined with text mining approach to give novel approach to classification of critical indicators and news articles. 3-8 classes' current classification of critical indicators is found and random forest gives the better performance in terms of high accuracy. Random forest is the best classifier to classify the news articles which are depending on bigram features.

Idrees et al. (2019) This research has predicted the stock market data by using time series analysis for good profit and future stocks. In this Forecasting, ARIMA, Time series analysis, Box Jenkins method and Stock market forecasts are used. Over all models ARIMA model has proved best predictions in terms of future stocks and profits. Chen (1994) In this project, multiple neural networks are applied and evaluated to predict. Here following neural networks are used: GRNN (General Regression Neural Network), CGNN (Conjugate Gradient Neural Network), BPN (Back-propagation trained network), CSNN (Class Sensitive Neural Network). From the predictions it stated that CSNN network is the best performing network amongst all of them. In this they showed that major indicators are leading over the indicator that is not used. Chen et al. (2015) In the proposed research RNNs (Recurrent neural networks) are powerful and good models for process

the linear data such as written neural language, time-series data, and sounds. LSTM (Long Short-Term Memory) is best amongst the RNN architectures. Long Short-Term Memory (LSTM) includes computational unit which replaces traditional artificial neurons, memory cell in the hidden layer of the RNN network. Oncharoen and Vateekul (2018) In this proposed research LSTM deep learning model approach is used to predict short and long term financial time series predictions. With the traditional neural network, without disappearing gradient problem LSTM holds time steps of arbitrary sizes. Both stacked and bidirectional LSTM models are used to benchmark with simple LSTM n/w and shallow neural n/w. Hassan and Nath (2005a) In this given research numerical information like technical indicators and historical prices, and textual information with headlines and news contents are used to predict the stock market prices. Where textual data contains text representation construction on the other side word embedding known as traditional method is not good match to represent semantics of financial news because of word sparsity problem in dataset.

2.4 New Techniques, Models and Algorithms to Predict Stock Market Trends

Cheng et al. (2018) HMM (Hidden Markov Models) used previously to predict the stock market. In this approach, the posterior Hidden Markov model is used to predict the stock value of coming day depending on the historical data. In this Intra-day low and high value of stock to train continuous Hidden Markov models and fraction of changes in stock values are taken into consideration. This HMM then used to make the maximum priority decisions on all possible stock values for the next coming day. Hassan and Nath (2005b) Attention based mechanism model is used to calculate stock prediction depending on attention based short and long term memory model to build trading strategies and to predict stock price movements. Most of the studies are there to by using deep learning model to predict the stock exchange. Attention mechanism from neural network has become popular in neural machine translation.

Attigeri et al. (2015b) Given research states that the hidden Markov models approach to predict the stock prices based on interrelated markets. HMM are used to forecast some airline stocks. Hidden Markov models are basically used to recognize patterns and classification problems because of its suitability which is proven to build dynamic modelling systems. We cannot use HMM to predict future events straightforward. HMM can be used only on past trained dataset's of the chosen airlines.

Karmiani et al. (2019) In given research stock analysis is done by using 2 techniques like: fundamental and technical analysis. By applying machine learning techniques on historical data, technical analysis is done and by using sentiments from social media fundamental analysis is done. Here correlation of sentiments and stock values are analyzed. And trained models are used to analyze and predict future stock values. Future work is to use summative assessment of sentiments for particular tweet or news for better prediction. Vui et al. (2013) In this research, it is stated that financial stock market prediction may be predicted better when we predict it using number of markets instead of one single market. Here for good predictions, selection criteria are used. This criterion is only based on RNN prediction accuracy, previous years recorded changes and observed back-propagation. Results are obtained for four consecutive years by using data of five

international stock markets. Nikfarjam et al. (2010) In this research, ANN is used to give the review of the ANN applications in prediction of stock market in order to predict what can be done in future. Also in this research neural network with hybrid method is used to predict the stock market and to compare with ANN. Schierholt and Dagli (1996) By taking consideration of news and its impact in analytic's of stock markets behavior, gives better predictions with more profit to the traders. In this given research stock market forecasting systems are introduced and on the same way implemented components are compared. Predictions are done by using different models which are applied to find out the impact of news on stock market analysis. Here labels are used to label news, and during time interval of news publication the prices are getting monitored though more accurate labels can get after analyzing stock market around the news publication.

Li and Liu (2009a) In this given research the standard and poor's 500 indexes are trained using multiple neural network classification architectures. Older Multilayered perceptions used for stock market predictions. In given research the probabilistic neural networks and perception architecture are used to forecast the steadiness, decline and incline of indexes. Kimoto et al. (1990a) In the given research different exterior stock price systems and interior complexity of stock market price system are shown, this research shows the analysis of stock market prediction depending on BP neural network where the predictions are calculated by using three-layered neural networks (feed-forward), principles depending on number of hidden layers, and current network topology. Li and Liu (2009b) In this research timing prediction of buying and selling is discussed for stock market on Tokyo stock exchange. By using different prediction methods and learning algorithms the stock predictions are visualized. Predictions are calculated based on modular neural networks. This research gives the predictions of stock market and also gives the best time to sell stocks. By using cluster analysis, price of stock fluctuation rule is extracted.

Kimoto et al. (1990b) The proposed research is based on WNN and GA (genetic algorithms) which states that the combination of this gives the better accuracy than other algorithms and existing approaches in terms of stock market predictions. The trading volume provided as inputs and instead of Morlet Gaussian wavelet function, Morlet wavelet function is used for the prediction model. Also here small number of hidden nodes is built as a new and additional work to the project. Normal topology of general algorithms with the max no of generations is chosen. Iteration of optimization process is chosen until both error / weights are optimized, to make complete use of genetic algorithm. Fang et al. (2014) Given research shows the predictions by using probabilistic lexicon generated by taking consideration of Stock market closing prices and Thai financial news are examined. By using PLSP (Probabilistic lexicon based stock market prediction) algorithm, relevant term is extracted and probabilistic values are assigned to it.

2.5 Conclusion

Depending on literature survey its clearly seen that there is clear results that brexit discussion and its impact is new for stock market prediction and also ARIMA time series prediction of 3 months is also helpful for it. The reviewed models, techniques and methodologies shows good data related to stock market prediction using machine learning techniques.

3 Data Preparation, Methodology, Design and Feature Selection

3.1 Introduction

Given chapter is the basic in detailed information of methodologies used in the research project and it's stated depending on the research question. Section 3.2 gives the detailed flow of Implementation flow which is used during the implementation. Section below 3.2 shows the information about methodologies, process, models, collection of data, input levels, methods and output levels.

3.2 Methodology used Approach

Crisp DM methodology is used in this research. Here the methodology is an iterative process which includes backup of the data and repeating approach. In every step it seems like an experiment whose outcomes are utilized successively to eventually accomplish the last results. Given figure shows the step by step flow of scientific technology used to Use the "Insert Citation" button to add citations to this document. Figure 1

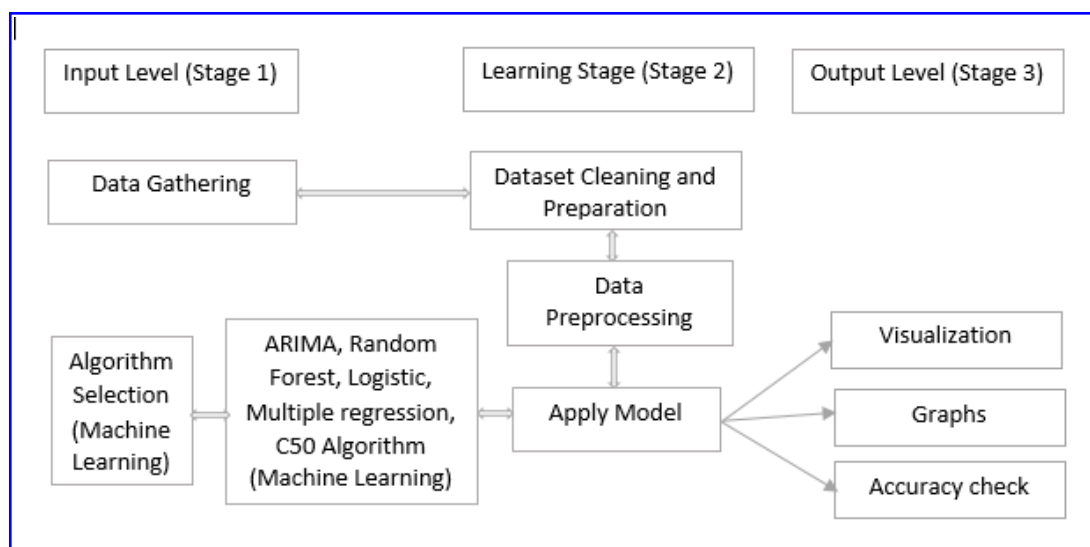


Figure 1: Methodology approach process flow diag.

In this given overall process flow the above methodology is divided into 3 different stages: Input stage, Learning stage and Output stage. This first stage includes dataset gathering, collection and scraping of data. Where stock market data is collected from London stock market website. Second stage is learning stage which involves learning phase where dataset cleaning, merging, processing, pre-processing has done where we can repeat each stage at any point of time to resolve the further obstacles. Most of the major decisions takes place in this stage. Final outcome contains in Output stage which gives the final results in terms of stock market predictions.

3.3 Architecture and Design of London Stock Market

Figure 2 shows the architectural design of implementation flow of stock market predictions. This architecture is made up of 2 tiers. Client tier and second is Business tier. Client layer is made up of client and visualization and results where, client is getting these results from tier 2 i.e. Business layer. Business layer consists of Data fetching/scraping, then by using R language that data gets cleaned. Here in data cleaning unnecessary data is removed and then took the necessary data as per requirements. After that, given data is preprocessed by using R language and SWOT function. Clean data is used to apply model on it. Here models applied are: ARIMA time series, Logistic regression, Multiple regression model, Random Forest, and Naive Bayes. Final outcome i.e. results and visualizations are used to predict the all models and send to the client for further use. Figure2

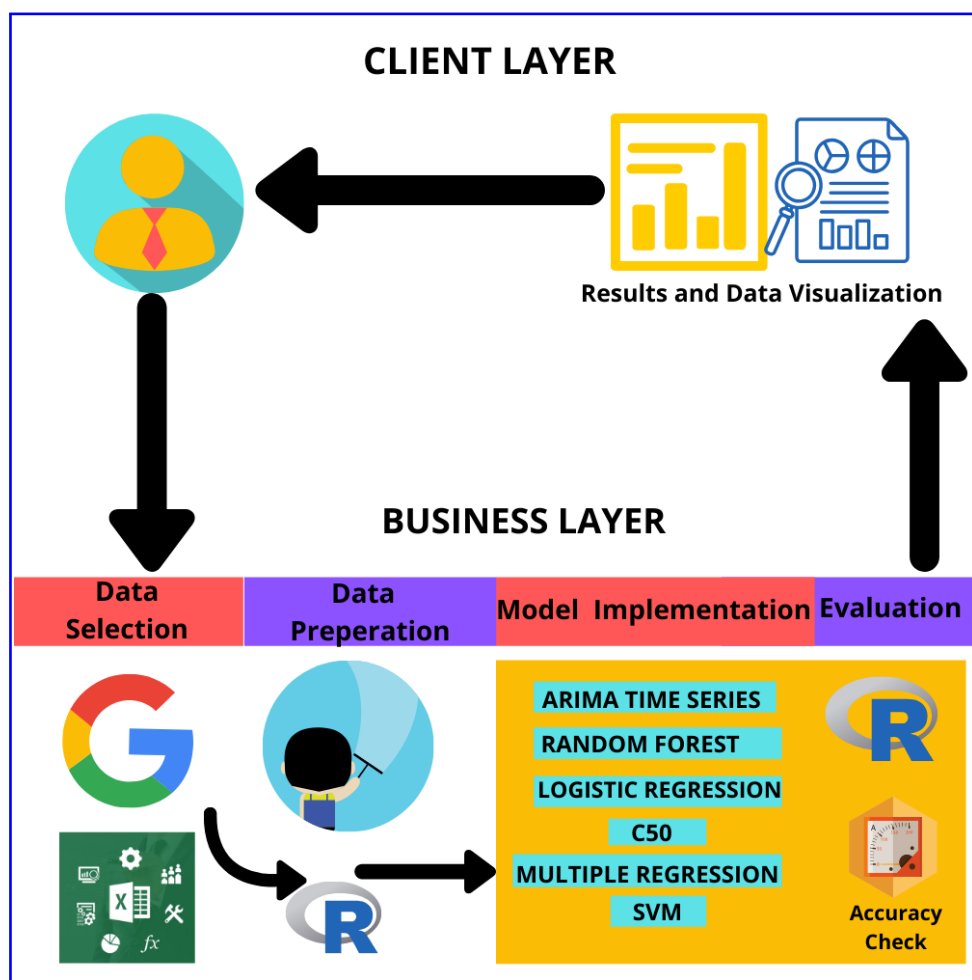


Figure 2: Methodology approach process flow diag.

3.4 Technical Requirement, Feature Selection, Data Preparation

Architecture level design includes different technologies and to predict stock market advanced libraries from R language are used. These libraries are used to create machine

learning models/ algorithms. Following are the methodologies and technologies which are used to predict London stock market.

1. Tier 2 – Business layer

Data Collection In order to predict London stock market, firstly all data is gathered from data sites. Where data is taken from ¹. All excel files are collected from this data source. Data collected is on monthly basis. All excel files including data from each year located on different sheets month wise. Data Preparation: Collected data is cleaned by using R language. Where all unnecessary data is removed by using R language commands. Also all Null and NA values are removed by using the same. Programming language: Programming language R is used to prepare models. These models are prepared by using advanced R libraries which are later on used to predict the London Stock market. Models Used are ARIMA time series model is used to predict the next 3 months stock market data. Where advanced ARIMA command is used. Logistic regression is used to predict what brexit discussion impact is happened on overall stocks. Random Forest, Multiple regression and C50 models are used to check accuracy. Data Evaluation: In data evaluation process the overall accuracy check is done. Total implemented results are checked in this section. Where each and every model go through this process.

Device specifications

Device Name: SONALI

Processor: Intel(R) Core(TM) i3-3217U CPU @ 1.80GHz

Installed RAM: 8.00 GB (7.98 GB Usable)

System type: 64-bit operating system, x64-based processor

Edition: Windows 10 Pro The numbers starts at 1 with every call to the enumerate environment.

2. Tier 1- Client Layer

Results: Final results collected in the form of Graph, Visualizations, and charts. The final outcome is collected from data evaluation where accuracy check has done.

Client: Final results which are collected from data evaluation are send to the user. Here London stock exchange data is predicted for users who want updates on overall London stock market. London stock market prediction gives better accuracy results in terms of company shares and company status. Also it gives idea to the client that in which company they should invest.

3.5 Feature Selection

The feature selection and research question is based on stock market industry. Primary objective of this topic is to find is there any impact happened on stock market during brexit discussion or not? Also by using ARIMA time series 3 months stock prediction is calculated. For accuracy check, Random forest, multiple regression, and Naive Bayes.

¹www.londonstockexchange.com

3.6 Data Preparation

Raw data is taken from ². Where all .xlsx files are stored for each month of the year. These files contained all the information about stock for the particular company. Each .xlsx file contained monthly data sheets. These all datasheets are merged together and combined in 1 dataset. All un-necessary columns and un-necessary data is removed by using R language. This data is later on pre-processed by using SWOT function in R language. This Trained data is used to apply model on it. Figure3

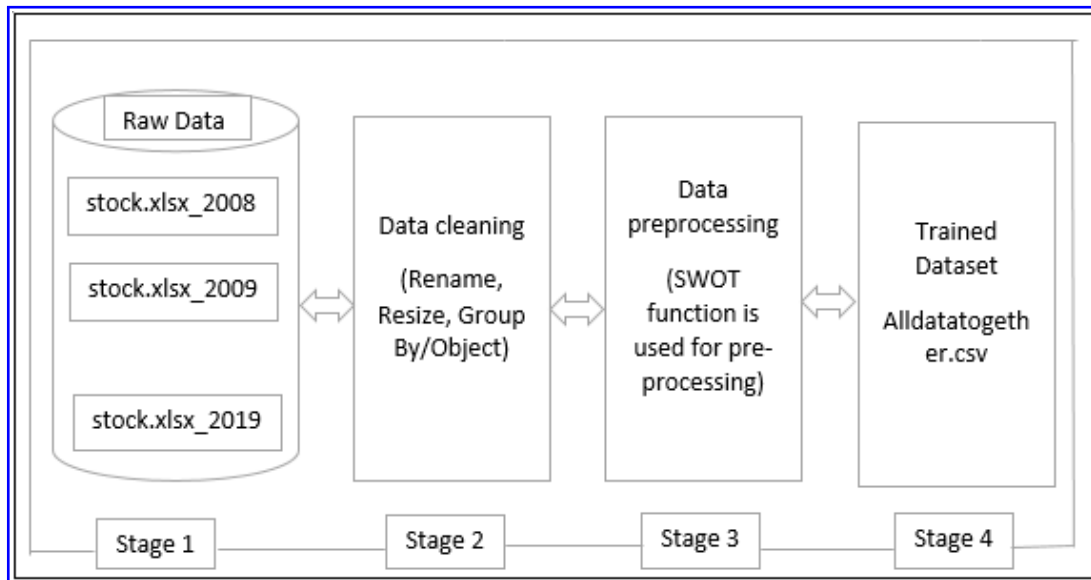


Figure 3: Data Preparation Process Flow

In this chapter methodology approach is presented. Entire process-flow Design and architecture, Data preparation process is briefly explained in implementation and solution.

4 Implementation, Evaluation, and Results Prediction Models for Stock Market Prediction

In this chapter overall implementation, results of ARIMA, Logistic regression, Random forest, multiple regression and c50 and evaluation is provided in detailed which is listed in Objective 2. Given chapter is also gives in detailed implementation, understanding of design, and development. Front-end implementation is given and named as user dashboard objectives. Major objective of this process is ARIMA model and classification of regression model. Hence aim of this major objective is to get predictions depending on the models. Entire implementation is discussed in sub objective 2(a) to 2(e). Next objective is further discussion about development process of stock market prediction. By implementing these objectives gives the idea about gross rated companies in which user can invest their money and brexit discussion impact review, which further going to comes in consideration for actual Brexit.

²londonstockexchange.com

4.1 Implementation, Evaluation and Results of ARIMA Model

ARIMA stands for Auto Regressive Integrated Moving Average model. Time series model forecasting is basically multidisciplinary scientific tool. Time series models are used for historical data. Different prediction problems are going to predict by using these models. This model only need necessary variables from historical data. ARIMA model equation is provided below:

$$\Delta^d X_t = (1 - B)^d X_t$$

As considering y , ARIMA will work as follows,

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q}$$

Following table shows the acf and pacf for arima model,

	ACF	PACF
AR	Geometric	Significant till p lags
MA	Significant till p lags	Geometric
ARMA	Geometric	Geometric

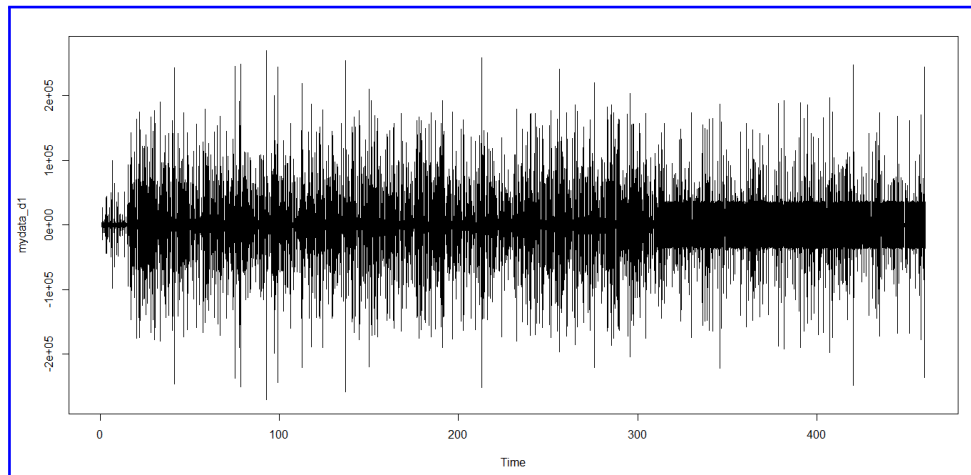


Figure 4: GGPlot for stock market and date

By using ARIMA time series model stock market income plotted against monthly data (in months), where time span is given from year 2010-2019 and by using this data above output were obtained. Figure4

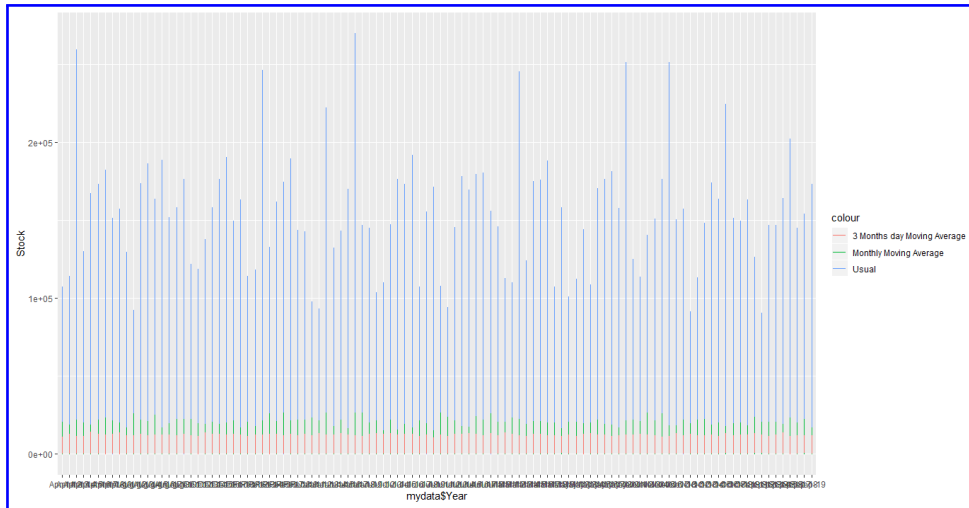


Figure 5: Moving average for 3 months stock

This moving average is evaluated and calculated to predict stock market data for 3 months. Where moving average is calculated on the basis of monthly data presented from year 2010 to 2019. Figure 5

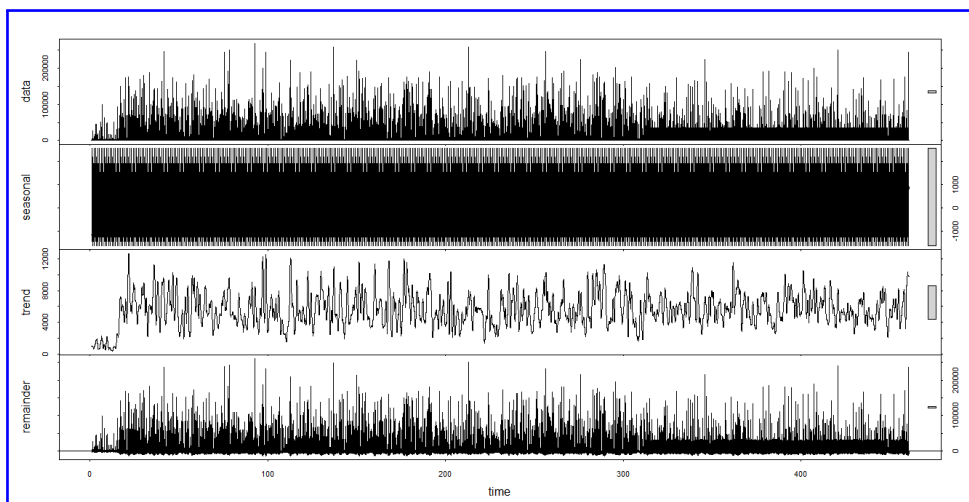


Figure 6: 3 months stock moving average

Here 4 components are present in this time series model. Which are Remainder, Trending, Seasonal (monthly), and data. These are plotted to check whether this data is non stationary or stationary. From the graph and component it is stated that seasonal or monthly components doesn't fulfil stationary assumptions hence those will be omitted. Figure 6

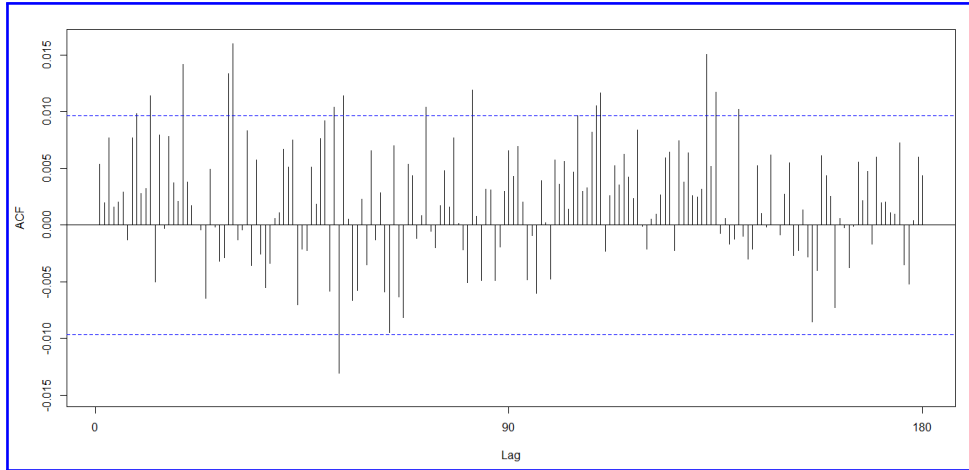


Figure 7: Co-relation using ACF

Co-relation of ARIMA time series is seems to be equally correlated because from the graph it can be stated that the lines because significant co-relation is seems to be equal to the non-significant co-relation in the residual lags.Figure7

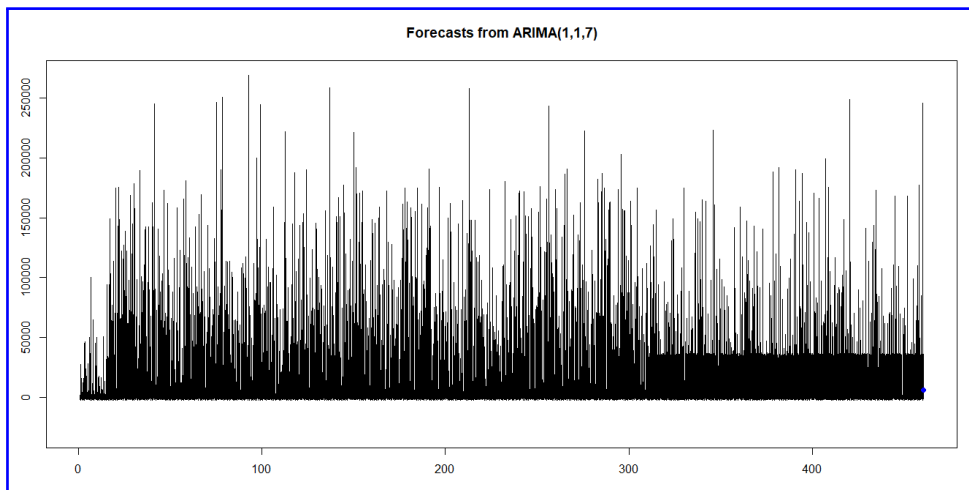


Figure 8: Forecast from ARIMA

Here the ARIMA model is plotted like that it is having 1 non-seasonal difference and one auto-regressive model and also 7 forecasting errors from the log itself. Blue dot indicating the predicted output of the model.

4.2 Implementation, Evaluation and Results of Random Forest

Random forest basically used to predict time series model. It is the most popular algorithm of data mining technique. This algorithm is known as ensemble learning algorithm. This algorithm is used for high dimensional classification as well as it's used to handle regression models. This algorithm ensemble tree based algorithm where collection of random variable trees are enlisted. Once all trees are collected, it got placed by

using or taking the help of bootstrap methods. Final outcome is taken by using achieved collected samples. Figure8

To make a prediction at a new point x :

$$f(x) = \frac{1}{K} \sum_{k=1}^K T_k(x)$$

Where K is number of trees,
 F is number of random i/p variables which are chosen during each split.
 The accuracy obtained from ARIMA time series model predicting
 Random forest model was applied on the given dataset which is used for the prediction of
 multiple different time series forecasting fields. Here Total number of trees were created
 was 500 and OOB error rate is 1.19%.

Confusion Matrix and Statistics		
	Reference	
Prediction	0	1
0	300	16
1	130	11962
Accuracy : 0.9882		
95% CI : (0.9862, 0.9901)		
No Information Rate : 0.9653		
P-Value [Acc > NIR] : < 2.2e-16		
Kappa : 0.7984		
McNemar's Test P-Value : < 2.2e-16		
Sensitivity : 0.69767		
Specificity : 0.99866		
Pos Pred Value : 0.94937		
Neg Pred Value : 0.98925		
Prevalence : 0.03466		
Detection Rate : 0.02418		
Detection Prevalence : 0.02547		
Balanced Accuracy : 0.84817		
'Positive' class : 0		

Figure 9: Accuracy check of Random Forest


```

Call:
  randomForest(formula = X ~ ., data = Trainset, importance = TRUE)
  Type of random forest: classification
  Number of trees: 500
  No. of variables tried at each split: 2

  OOB estimate of error rate: 1.19%
Confusion matrix:
  0    1 class.error
0 665  288 0.302203568
1  56 27942 0.002000143

```

Figure 10: Confusion matrix of Random forest

Accuracy for random forest tree was obtained: 98.79% Sensitivity was obtained was 69.07% and specificity given by this model is 99.85%. Figure9 Figure10

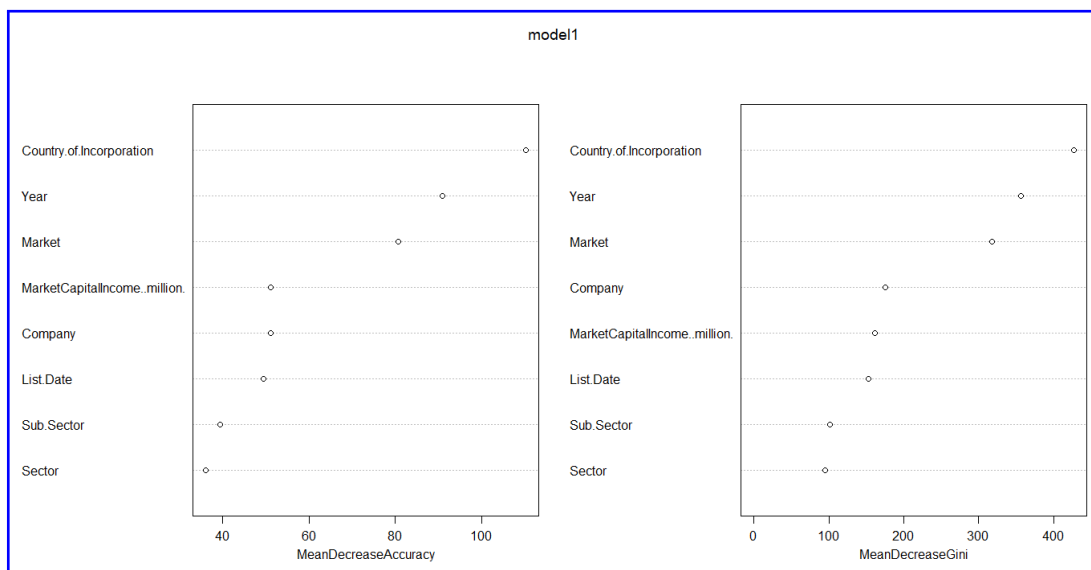


Figure 11: Random Forest Classification

Number of variables are split into 2 at each and every split. Figure11 Here above variable measure importance is measured by using varimplot function which is described above. Figure12

4.3 Implementation, Evaluation and Results of Logistic Regression

This algorithm is machine learning algorithm. This algorithm is mainly used for classification problems, it is based on probability concept and predictive analysis. Logistic regression is used if the Y variable gives values either in 0's or 1's. The predictions are basically determination of mathematical equation which is used to predict the value or probability of event 1. Figure8

$$1 / (1 + e^{-\text{value}})$$

Logistic regression formula in term of Y is,

$$y = e^{(b_0 + b_1 \cdot x)} / (1 + e^{(b_0 + b_1 \cdot x)})$$

Where Y: Output (Prediction)

B0: Intercept or bias term

B1: Coefficient

In logistic regression Residual Histogram shows the allocations of residuals for all given observations.

Pattern	Pattern Indicators
Bar which is faraway from the all other plotted bars	Significant till p lags
An outlier	Skewness

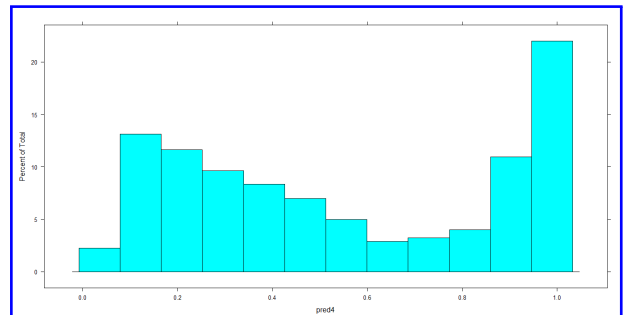
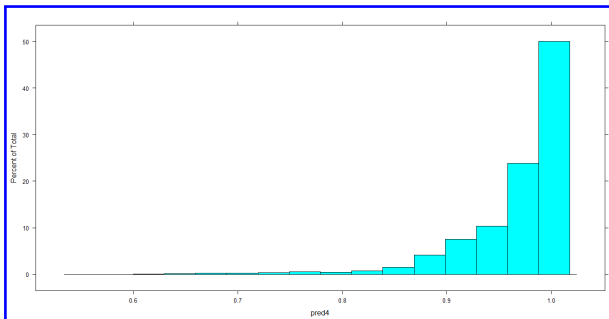


Figure 12: 2 Figures side by side

When we compare both histograms it's clearly seen that there is drastic change from the year 2010-2019 to Brexit month discussion October 2019. As shown in Figure, the histogram is gradually increasing every year in terms of stock market Company's status (International user/not). As seen in graph it shows tremendous impact in stock market. Figure

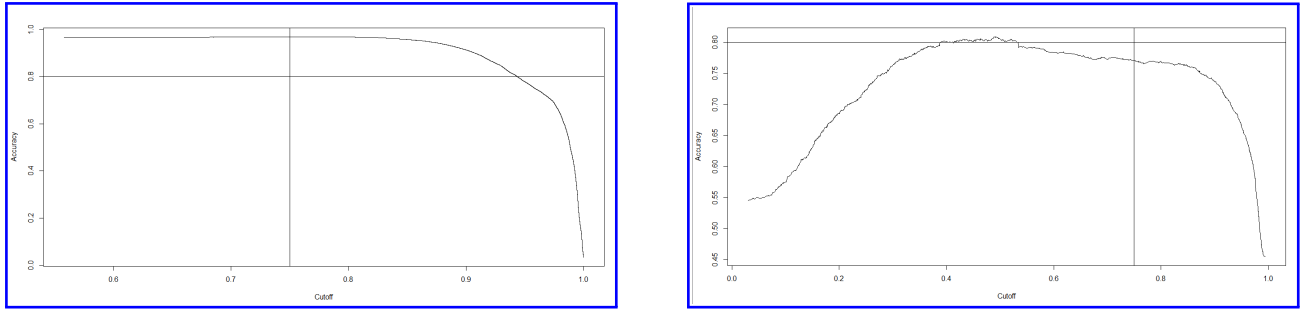


Figure 13: Histogram of year 2010-2019 and October 2019

Here both curves have tremendous impact on stock market during brexit discussion. In Accuracy curve 2010-19 it was steady from 0-0.75 but after cutoff it got decrease drastically. The other curve from fig.19 there are ups and downs where it got reached to the accuracy point and after that during the period of Brexit discussion it got decreased. Hence from the curve it can be noted that there is impact of Brexit discussion on stock market.

```
> histogram(pred4)
> pred4 <- prediction(pred4, mylogistic$x)
> eval <- performance(pred4, "acc")
> plot(eval)
> abline(h=0.8, v=0.75)
> max <- which.max(slot(eval, "y.values")[[1]])
> max
[1] 38377
> acc <- slot(eval, "y.values")[[1]][max]
> acc
[1] 0.9689789
> cut <- slot(eval, "x.values")[[1]][max]
> cut
      5629
0.7659658
> print(c(Accuracy=acc, Cutoff = cut))
      Accuracy Cutoff. 5629
0.9689789    0.7659658
```

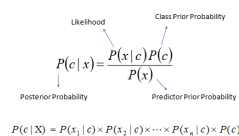
```
> histogram(pred4)
> pred4 <- prediction(pred4, mylogistic$x)
> eval <- performance(pred4, "acc")
> plot(eval)
> abline(h=0.8, v=0.75)
> max <- which.max(slot(eval, "y.values")[[1]])
> max
[1] 998
> acc <- slot(eval, "y.values")[[1]][max]
> acc
[1] 0.8091385
> cut <- slot(eval, "x.values")[[1]][max]
> cut
      2099
0.4913529
> print(c(Accuracy=acc, Cutoff = cut))
      Accuracy Cutoff. 2099
0.8091385    0.4913529
```

Figure 14: Accuracy vs Cutoff curve of year 2010-2019 and October 2019

Depending on the above pictures when we consider accuracy from year 2010-19 it seems more i.e. 96.897 while on other hand accuracy of Brexit month discussion is 80.91. Accuracies are calculated on the basis of market users (International/Normal).

4.4 Implementation, Evaluation and Results of Naïve Bayes

Naïve Bayes is classification algorithm which gives hypothesis of individuality among predictors. It shows that the presence of feature which is present in a class is not related to the presence of any other feature. Figure8



Naïve Bayes algorithm is basically used to predict the predictors independently and its relations. In simple terms it assumes that the existence of certain feature in a class is not related to the existence of any other feature.

```

> #confusion matrix
> mat <- confusionMatrix(p1_n, train_nb$X)
> mat
      predictions
targets 1      2
1      966 1589
2      151 30384
> #checking accuracy, precison, recall
> (accuracy <- sum(diag(mat)) / sum(mat))
[1] 0.9474161
> accuracy
[1] 0.9474161
> (precision <- diag(mat) / rowSums(mat))
      1      2
0.3780822 0.9950549
> precision
      1      2
0.3780822 0.9950549
> (recall <- diag(mat) / colSums(mat))
      1      2
0.8648165 0.9503018
> recall
      1      2
0.8648165 0.9503018
> #AUC/ROC
> n <- roc(X ~ p1_n, data = train_nb)

```

Figure 15: Accuracy check

From the above accuracy code, accuracy is calculated which is 94% And precision is coming 37%. Naïve Bayes model is basically used on big data.

4.5 Implementation, Evaluation and Results of Multiple Regression

This model is basically works exactly like its name, where this model is used to predict multiple variables to predict the results of different variables. Main goal of this model is to predict the correlation between independent and dependent variable.

$$Y = \beta_0 + \beta_1 X$$

Where it shows the relationship between Y (continuous variable) and x1, x2, x3... etc.(2 or more independent variables.)

```

> summary(model)$coefficient
      Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 1171.061   498.7627  2.347932 1.888266e-02
X            4631.680   507.3168  9.129757 7.166540e-20
> model <- lm(MarketCapitalIncome..million. ~ X, data = newstock)
> summary(model)

Call:
lm(formula = MarketCapitalIncome..million. ~ X, data = newstock)

Residuals:
    Min       1Q   Median       3Q      Max
-5803  -5800  -5687  -4353  263862

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1171.1      498.8   2.348  0.0189 *
X            4631.7      507.3   9.130 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18550 on 41357 degrees of freedom
Multiple R-squared:  0.002011, Adjusted R-squared:  0.001987
F-statistic: 83.35 on 1 and 41357 DF, p-value: < 2.2e-16

> confint(model)
              2.5 %    97.5 %
(Intercept) 193.4753 2148.646
X           3637.3278 5626.031
> #Residual Standard Error (RSE), or sigma
> sigma(model)/mean(newstock$MarketCapitalIncome..million.)
[1] 3.284135

```

Figure 16: Accuracy check

Above diagram shows the accuracy of multiple regression model which was applied on London stock market. Where as per calculation accuracy is measured as 97%.

4.6 Conclusion

The evaluation, implementation, development and results of all algorithms ARIMA time series, Random Forest, Naïve Bayes, Logistic regression, and multiple regression implemented and applied accordingly. The implementation results are also donated the body of information. Also it has equivalent solved and implemented the research question.

4.7 Discussion

This section is about the detailed discussion of impact of Brexit discussion on London stock market using logistic regression also it includes future 3 month prediction of stock market by using ARIMA time series and other models like Random forest, Naive Bayes and Multiple regression is predicted and accuracy is calculated on it. Brexit discussion is happened in October 2019 and it got delayed to the year 2020. This is very unique topic and never been used before by any researcher. Depending on the results Brexit discussion had tremendous impact on London stock market. And if we compare accuracy of each model then it seen that the Random forest model had 98%, Multiple regression model had 97% and Naive Bayes model accuracy calculated were 94% as well. Hence according to the results accuracy of Random forest model were more precise and higher than the others. Future work for this is to predict the impact of actual Brexit once it happen. Next Brexit discussion and the current deadline is given is 31st January 2020. Hence

there are lot of things to work on Brexit and its impact. Also we can do the prediction depending on the impact happened on each industry and each sector.

5 Conclusion and Future Work

This project primarily gives and resolves the customer doubts in terms of London stock market. It also gives in detailed idea about London stock market and impact of Brexit Discussion on it. Chapter 4 includes implementation, evaluation and results all together and shows solutions of enlisted research question. Chapter 2 provided includes the literature survey from year (2004 to 2019). Which includes the problems discussed regarding stock market issues and their predictions.

Detailed model explanation, implementation and evaluation of all algorithms is listed in chapter 4. This work is done by many researchers but Brexit discussion impact on London stock market has been never done by any researcher. Also prediction of 3 months of November, December and January 2020 is calculated by using ARIMA time series model. Future work is to predict the Brexit impact on London stock market and its outcomes. Now a days Brexit has become a trending topic in European unions.

6 Acknowledgement

I would love to thank my mentor Dr. Catherine Mulwa for her support and patience throughout the semester. She gave me moral as well as mental support regarding entire project. Her support and contribution in my research helped me to gain lot of knowledge in machine learning and the related research's. Today I am able to do this research just because of her support and contribution. Because of this research project i grasped lot of knowledge regarding Data Analytics. I would love to thank my parents and friends because, they also gave me lot of mental support to survive and to complete my project on time.

References

- Attigeri, G. V., Manohara Pai M M, Pai, R. M. and Nayak, A. (2015a). Stock market prediction: A big data approach, *TENCON 2015 - 2015 IEEE Region 10 Conference*, pp. 1–5.
- Attigeri, G. V., Manohara Pai M M, Pai, R. M. and Nayak, A. (2015b). Stock market prediction: A big data approach, *TENCON 2015 - 2015 IEEE Region 10 Conference*, pp. 1–5.
- Chen, C. H. (1994). Neural networks for financial market prediction, *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 2, pp. 1199–1202 vol.2.
- Chen, K., Zhou, Y. and Dai, F. (2015). A lstm-based method for stock returns prediction: A case study of china stock market, *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2823–2824.

- Cheng, L., Huang, Y. and Wu, M. (2018). Applied attention-based lstm neural networks in stock prediction, *2018 IEEE International Conference on Big Data (Big Data)*, pp. 4716–4718.
- Elagamy, M. N., Stanier, C. and Sharp, B. (2018). Stock market random forest-text mining system mining critical indicators of stock market movements, *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–8.
- Fang, Y., Fataliyev, K., Wang, L., Fu, X. and Wang, Y. (2014). Improving the genetic-algorithm-optimized wavelet neural network for stock market prediction, *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 3038–3042.
- Hassan, M. R. and Nath, B. (2005a). Stock market forecasting using hidden markov model: a new approach, *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 192–196.
- Hassan, M. R. and Nath, B. (2005b). Stock market forecasting using hidden markov model: a new approach, *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 192–196.
- Idrees, S. M., Alam, M. A. and Agarwal, P. (2019). A prediction approach for stock market volatility based on time series data, *IEEE Access* **7**: 17287–17298.
- Karmiani, D., Kazi, R., Nambisan, A., Shah, A. and Kamble, V. (2019). Comparison of predictive algorithms: Backpropagation, svm, lstm and kalman filter for stock market, *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 228–234.
- Kimoto, T., Asakawa, K., Yoda, M. and Takeoka, M. (1990a). Stock market prediction system with modular neural networks, *1990 IJCNN International Joint Conference on Neural Networks*, pp. 1–6 vol.1.
- Kimoto, T., Asakawa, K., Yoda, M. and Takeoka, M. (1990b). Stock market prediction system with modular neural networks, *1990 IJCNN International Joint Conference on Neural Networks*, pp. 1–6 vol.1.
- Kumar, I., Dogra, K., Utreja, C. and Yadav, P. (2018). A comparative study of supervised machine learning algorithms for stock market trend prediction, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1003–1007.
- Li, F. and Liu, C. (2009a). Application study of bp neural network on stock market prediction, *2009 Ninth International Conference on Hybrid Intelligent Systems*, Vol. 3, pp. 174–178.
- Li, F. and Liu, C. (2009b). Application study of bp neural network on stock market prediction, *2009 Ninth International Conference on Hybrid Intelligent Systems*, Vol. 3, pp. 174–178.
- Li, X., Xie, H., Lau, R. Y. K., Wong, T. and Wang, F. (2018). Stock prediction via sentimental transfer learning, *IEEE Access* **6**: 73110–73118.
- Liu, G. and Wang, X. (2019). A numerical-based attention method for stock market prediction with dual information, *IEEE Access* **7**: 7357–7367.

- Nikfarjam, A., Emadzadeh, E. and Muthaiyah, S. (2010). Text mining approaches for stock market prediction, *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, Vol. 4, pp. 256–260.
- Oncharoen, P. and Vateekul, P. (2018). Deep learning for stock market prediction using event embedding and technical indicators, *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pp. 19–24.
- Schierholt, K. and Dagli, C. H. (1996). Stock market prediction using different neural network classification architectures, *IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pp. 72–78.
- Sharaff, A. and Choudhary, M. (2018). Comparative analysis of various stock prediction techniques, *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 735–738.
- Soni, D., Agarwal, S., Agarwal, T., Arora, P. and Gupta, K. (2018). Optimised prediction model for stock market trend analysis, *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1–3.
- Vui, C. S., Soon, G. K., On, C. K., Alfred, R. and Anthony, P. (2013). A review of stock market prediction with artificial neural network (ann), *2013 IEEE International Conference on Control System, Computing and Engineering*, pp. 477–482.
- Yoo, P. D., Kim, M. H. and Jan, T. (2005). Machine learning techniques and use of event information for stock market prediction: A survey and evaluation, *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, Vol. 2, pp. 835–841.
- Zhang, X., Qu, S., Huang, J., Fang, B. and Yu, P. (2018). Stock market prediction via multi-source multiple instance learning, *IEEE Access* **6**: 50720–50728.