**Stock Market Analysis Using Linear Regression**

**Team Members**

1. **Name :** Harshitha Kolli

**St id :** 11521716

**Mail id:** [harshithakolli@my.unt.edu](mailto:harshithakolli@my.unt.edu)

1. **Name :** Gowthami Mulpuri

**St id :** 11600593

**Mail id:** [gowthamimulpuri@my.unt.edu](mailto:gowthamimulpuri@my.unt.edu)

1. **Name :** Naga Tulasi Parvathi Veerla

**St id :** 11626471

**Mail id:** [NagaTulasiParvathiVeerla@my.unt.edu](mailto:NagaTulasiParvathiVeerla@my.unt.edu)

**Introduction:**

"Others' thought process" has always been an important piece of information for the great majority of us while making decisions. The Web and the Internet have made it possible to learn about the perspectives and experiences of people who are neither our personal acquaintances nor well-known expert commentators, that is, people we have never heard of. As a result, an increasing number of people are using the Internet to share their ideas with others. Individual consumers' advantage in web-based speculations about products and administrations, as well as the potential influence such conclusions have, is a driving force behind this field of inquiry. Furthermore, there are many challenges associated with this cycle that must be overcome. to be trampled to achieve the best results We focused this research on the fundamental philosophy that occurs in these systems and the procedures that must be done to address the challenges that arise.

**Motivation**

Organizations are primarily concerned with customer satisfaction and item surveys. Changes in web-based entertainment industry have been compared to changes in stock marketplaces. Distinguishing and resolving consumer complaints improves client satisfaction and an organization's dependability. As a result, an unbiased automated strategy is intended to describe customer surveys concerning any concern. Organizations may have gathered mountains of consumer feedback in the present atmosphere, when people are typically suffering information overload (albeit this does not indicate better or more meaningful bits of knowledge). Nonetheless, it is impossible for ordinary people to break it down physically without error or bias. The best-intentioned organizations are frequently stuck in a data vacuum. You are aware that you anticipate events to guide your path, and although you are You are conscious of your lack of them, but you are confused how to address the matter. Opinion research provides some insight into the primary concerns, primarily from the perspective of clients. Because opinion assessment may be automated, conclusions can be based on a large amount of information rather than pure instinct.

**Objective**

Time series estimation and presentation is fundamental in data analysis. Time series analysis is a subset of measures that is often used in areas such as econometrics and tasks research. Time series are commonly utilized in testing and information science. Stock costs vary and are still determined by a variety of factors. The fundamental goal of this project is to assess stock quality utilizing Linear Regression.

**Background**

**Using machine learning, predict stock market movements**

The exam work was completed by V Kranthi Sai Reddy, a student at the Sreenidhi Foundation for Science and Innovation and an ECM.[1]

Hyderabad is a city in India. Global One of the most significant financial developments is stock trading. Stock market valuation is an attempt to predict the future value of a company's stock or other financial product being traded as part of a monetary transaction. This document explains how AI is used to forecast a stock. When preparing stock forecasts, most brokers use specialized, prime or time stock research.Python is the programming language used to implement the framework that will control the stock market. We propose an AI (ML) approach in this research. The exact extrapolation is then based on available offer data and knowledge of the advantages. This quick look uses a frame in this case. Using costs with daily and modern frequencies, the Help Vector Machine (SVM) method is used to calculate the costs of large and small cap stocks within the three specific lines of business.

**Stock Market Index Prediction Using Artificial Intelligence Methods**

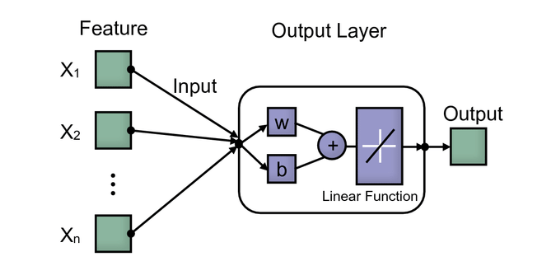
Ronald Lufuno Marwala led the study. a work submitted to the Built Climate and Design Staff at the College of the Witwatersrand in Johannesburg to meet the requirements for the Science Expert in Science certificate. The fragile form of competent market speculation (EMH) testifies to the fact that estimating the future cost of a resource is unthinkable given the information contained in the actual cost of a resource. This shows that the market behaves like a bumpy ride, making the estimate unthinkable. Furthermore, monetary determination is a difficult task due to the inherent complexity of the monetary framework.The aim of this work was to demonstrate and predict the future costs of using human-made innovation in consciousness (computational intelligence). a statement of transaction Three methods of artificial reasoning (Brain Networks (NNs), Fully Convolutional Machines and Central Nervous System Frameworks) are used to forecast the future cost of a capital market file based on past cost data. Man-made reasoning techniques are used as tools for forecasting financial time series and can take into account the complexity of financial frames.

For the benefit of investors, accurate stock market forecasting is essential since it indicates whether an investment will be profitable or not. Although many techniques have been used to predict stock prices, Artificial Neural Networks are the first technique to predict stock price trends [3]. Kim [4] used SVM to forecast stock market prices during the previous 20 years.

Historical data were examined between linear regression and support vector regression by Kavitha S. et al. (2016) [5]. They utilised the LeastMedSq function and SMOReg function, respectively, for the regression approach. LeastMedSq is a method of linear regression that minimises the median of the squares of withdrawal from the regression line. The support vector machine for regression is actualized by SMOreg. Different calculations can be used to master the boundaries. The RegOptimizer is set to select the calculation. Because of Shevade, Keerthi, and others, the most well-known calculation (RegSMOImproved) is the default RegOptimizer.

**Model**

**Linear Regression**



In the simplest terms, linear regression is a supervised machine learning model that identifies the linear relationship between the dependent and independent variables by determining the best fit linear line between them.

There are two varieties of linear regression: simple and multiple. When only one independent variable is present, simple linear regression must be used to determine its linear connection to the dependent variable.

In contrast, several independent variables are used in multiple linear regression to identify relationships.

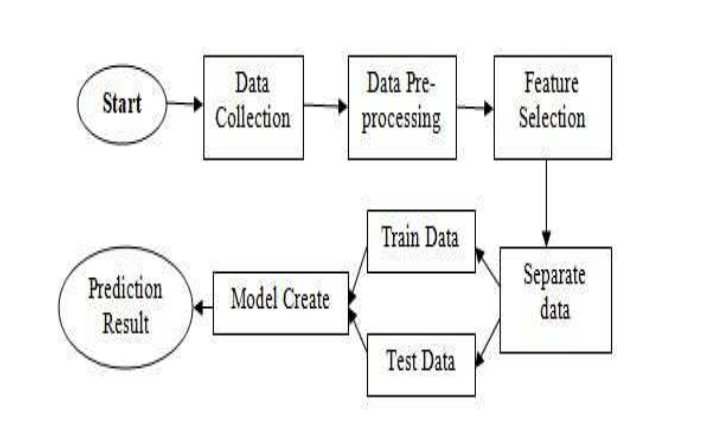
Simple linear regression equation with x as the independent variable, y as the dependent variable, b1 as the coefficient or slope, and bo as the intercept.

linear regression 1

Equation for multiple linear regression, where y is the dependent variable and b1, b2, b3, b4,..., bn are the coefficients or slopes of the independent variables x1, x2, x3,..., xn.

linear regression 2

**Workflow**



1. Dataset

At this point, we look for aggregated data from various sources before choosing a recent dataset from kaggle. This dataset is divided into three size-related parts (1 year, 6 months, and 3 months). Attributes in this dataset include Open, High, Low, Close, and Volume, but I chose Close as the label data and other attributes to extract features that help predict results.

B. Data preprocessing and cleansing

During this time, we used statistical techniques to look for null values ​​in cleaned and populated data and unidentified data types. For example, when I found a null value in a mapped attribute, I was looking for a discrete or categorical value. If it is a classification value, the median will be calculated; otherwise, the mean value will be calculated and placed on the null places.

C. Divide the data into a test and train dataset

prepared information divided into the components instruct and verify 20% and 80% respectively in the ratio. The percentage of exams and trains may have an impact on how accurately the outcome is predicted. I used the same ratio for the educate and test for a dataset with 3 sizes. It's up to you to decide at this point what ratio of educated to checked data you want to use, however accuracy might be improved if you use more educated datasets. For the check and teach datasets, popularity ratios are 80% and 20%, respectively.

D. Model Fit of Train Data

To improve predictions, the appropriate method must be chosen based on the dataset. I selected Linear Regression and Decision Tree Regression for this dataset in order to compute the prediction of the dataset's varying size and train the model using the training dataset.

E. Review the Data

The test dataset, which makes up 20% of the total data set, was used in this learning approach to produce results, which were then assessed against the actual output.

**Dataset**

**Dior Stock Information Xls file**

Parameters Like Open, Close, High, Low,Volume

**Opening cost**

The value of each share at the start of trading on the S&P 500 stock exchange is known as the starting price. The opening price provides a reliable forecast of the stock's daily movement. The beginning price need not match the closing price from the previous day because the stock market is comparable to an auction when buyers and sellers meet to negotiate with the highest bidder.

**Day's highest/lowest pricing**

The day's highest and lowest prices are calculated from the previous day and provide an indication of how much the shares typically fluctuate throughout the day and how this will ultimately effect the closing price. Additionally, it displays each share's overall cyclical movement.

**Closing price**

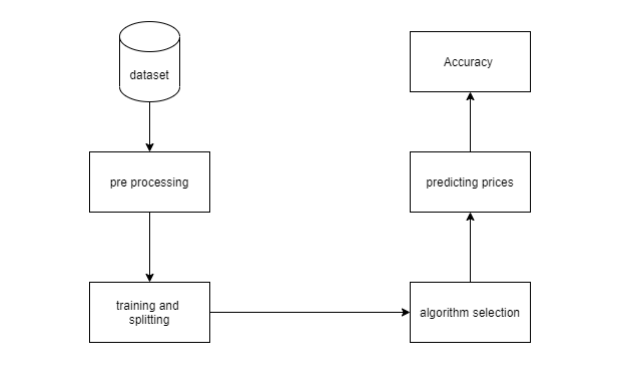
The closing price of a stock on any given trading day is revised to reflect any distributions and corporate actions that occurred at any point prior to the opening of the following day. The adjusted closing price is typically used for analysing historical returns in-depth or when examining prior returns.

**Volume**

Volume is one of the most fundamental and practical concepts to understand when trading stocks. Volume is defined as "the number of shares or contracts traded in a security or in the entire market during a specified period of time."

**Detail design of Features**

**Component Of System**



In UML, a part graph is a type of chart. The goal is also distinct from the previous charts discussed. It does not represent the framework's utility, rather it depicts the components that enable such capabilities.

Part graphs are used to visualize, depict, and archive part-based frameworks, as well as to create executable frameworks by forwarding and picking apart. Part graphs are essentially class charts that focus on a framework's components and are frequently used to depict a framework's static execution viewpoint.

**Analysis of data**

The money market is a dynamic and complex framework in which anyone can trade forms of currency, stocks, offers and subsidiaries through virtual platforms that traders work with. These stock exchanges allow lenders to buy shares in public companies through trades or over-the-counter exchanges. Contrasted with the risk of starting a new business or pursuing a lucrative job, this market offers lenders the potential opportunity to make money and live rich lives by putting small sums aside at an acceptable interest rate. Many variables affect stock markets, causing market vulnerability and extreme instability. Robotic trading systems (ATS) use computer programming, although humans can take orders and send them to the market. Projects can process orders faster and more accurately than individuals. However, to assess and manage HTM filing, board risk frameworks and human-based wellness measures are undeniably necessary. Many variables are integrated and analyzed when promoting an ATS, for example the trading methodology to be used, complex numerical skills reflecting the state of a particular stock, AI calculations relating the stock value expectancy futures and specific news on the share tax is decomposed. .

**System configuration**

This task might be completed with normal equipment. The entire operation was performed using an Intel I5 CPU with 8 GB Slam and a 2 GB Nvidia Realistic Processor. It also has two processors that run at 1.7 GHz and 2.1 GHz. The primary part of the interaction is the preparation stage, which takes around 10-15 minutes, and the next part is the trying stage, which takes only a few moments to set expectations and calculate exactness.

**Hardware Requirements:**

• Slam: 4 GB

• Capacity: 500 GB

• Computer processor: 2 GHz or quicker

• Engineering: 32-cycle or 64-bit

Programming prerequisites

• Python 3.5 in Google Collab is utilized for information pre-handling, model preparation and expectation.

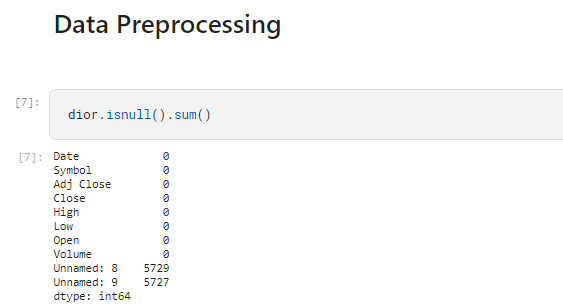
**Software requirements**

Python 3.5 in Google Collab is utilized for information pre-handling, model preparation and expectation.

• Working Framework: windows 7 or more or Linux based operating system or Macintosh operating system.

**Data pre-processing** is a data mining technique that involves transforming unstructured data into a format that can be used. There are issues with our dataset, including missing records, erroneous values, and null values. To make sure that our data could be predicted accurately, we applied the pre-processing techniques indicated below.

1. **Cleaning up data** Data cleaning aims to resolve conflicts in the data and fill in missing values. The index, date, and time closing prices from the NSE dataset are used as input. Public holidays caused some values to be missing. Incorrect indexes and null values were removed. A few extraneous columns in the dataset that weren't used as input were present. We eliminated those columns to simplify our prediction model.



**Implementation**

Information perception is the discipline of attempting to interpret information by visualizing it in order to highlight examples, patterns, and correlations that may otherwise go undiscovered.

Python offers a plethora of fantastic charting packages that are jam-packed with essential features. Python provides an excellent framework for creating dynamic or highly adaptive diagrams.

To give you an idea, here are a number of well-known plotting libraries:

• Matplotlib is simple to use and provides a lot of versatility.

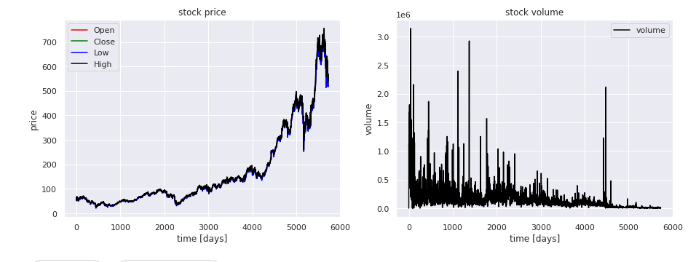
• Seaborn offers an amazing preset style library and a high-level UI.

• Pandas Visualization is built on Matplotlib and features an easy-to-use interface.

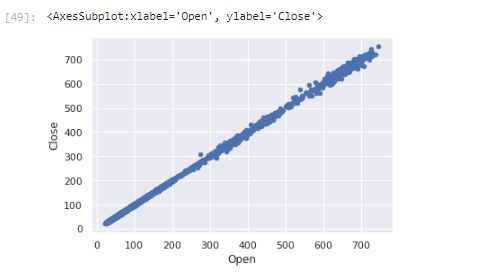
• Plotnine is an R-based plotting tool that uses Grammar of Graphics.

• Plotly allows you to create interactive graphs.

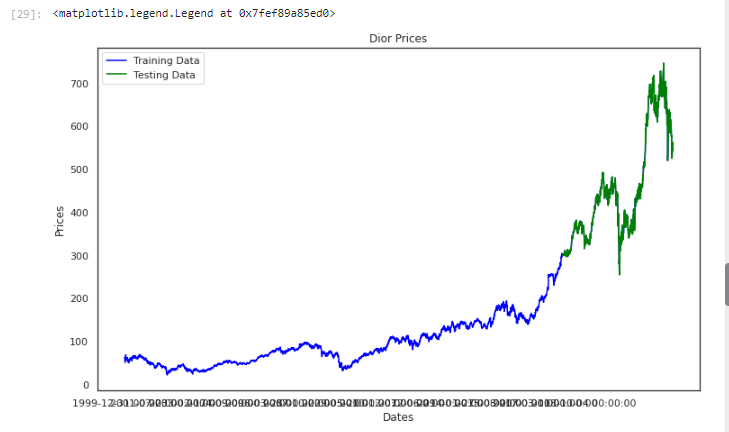
**Graph with data flow based on number of days**



**The first way we can plot things is using the .plot extension from Pandas dataframes We'll use this to make a scatterplot of the dior stock features.**



**Time Series Prediction**



**Scatterplot**

The dissipation strategy in Matplotlib can be used to create a scatterplot. We will also use plt. subplots to create a shape and pivot to give our chart a title and markers.

**Histogram**

By including the hist strategy in matplotlib, we can create a histogram. For example, if we provide absolutes, the focus portion of the wine survey dataset determines how often each class appears.

**Chart with bars**

The strategy of bars can be used to create a bar chart. Since bar contouring doesn't naturally work with repeating a classification, we use Pandas' estimate counting technique to achieve this. The bar outline is useful for unmitigated data with fewer than 30 different classifications, but it can be quite confusing in any case.Box and whisker plots of time series by interval of The histograms and the thickness charts reveal the transmission of all the information, but we could be more interested in the appropriation of the values ​​over time. The case and stubble diagram is another type of diagram that can be used to summarize the acquisition of perceptions. This realist structures a bin around the 25th and 75th percentiles of information, capturing the middle half of perceptions. To account for the broad levels of information, a line is drawn at the 50th percentile (in the middle) and hairs are drawn above and below the bin. The exceptions outside the hairs or degrees of information are separated by periods. For each section, box and stubble charts can be created and differentiated. over a number of periods, e.g. years, months or days.

**Python Heatmaps**

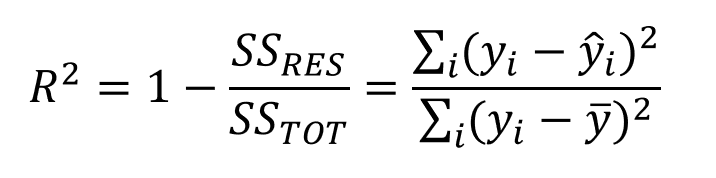
Practical Python Data Science For each value to be plotted, a heatmap contains values ​​showing some nuance of a similar variety. The fainter shades of the chart generally relate to higher qualities than the lighter shades. Something completely different can also be used for an essentially unique value.

**Violin plot**

A fiddle plot works the same as a stubble and fall plot. It represents the -distribution of quantitative information over many degrees of (at least one) class variable, allowing these means to be caught up.

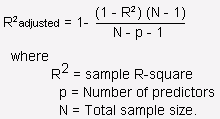
## Evaluation Metrics for Regression Analysis

1. **R squared or Coefficient of Determination:** R squared is the most often employed metric for model evaluation in regression analysis. It is a ratio of variation to total variance, according to one definition. R squared ranges from 0 to 1, with a value closer to 1 indicating a better model.



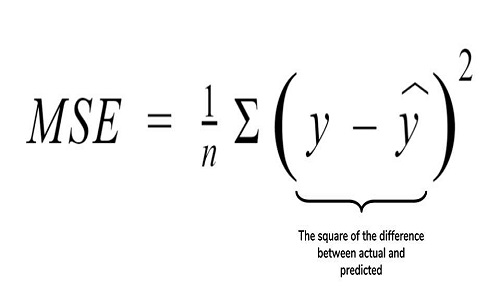
where SSRES is the Residual Sum of squares and SSTOT is the Total Sum of squares

2. **Adjusted R squared:** It is a R squared improvement. The issue or negative with R2 is that as the number of features increases, so does the value of R2, giving the impression that the model is sound. As a result, the Adjusted R2 addresses R2's flaw. It solely takes into account the attributes that are crucial to the model and demonstrates how much the model has actually improved. R2 is always higher than adjusted R2.



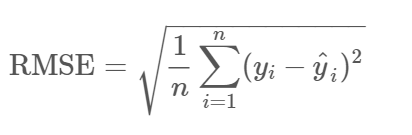
Source: stats.stackexchange.com

3. **Mean Squared Error (MSE)**: Mean squared error, which is the average of the squared difference between actual and anticipated values, is another popular statistic for evaluation.



Source: cppsecrets.com

4. **Root Mean Squared Error (RMSE)**: It is the mean difference between actual and predicted values, or the root of the MSE. MSE does not punish significant errors, although RMSE does.



**Linear Regression Implementation**

**Algorithms / Pseudocode**

1. Start
2. Read Number of Data (n)
3. For i=1 to n:

Read Xi and Yi

Next i

1. Initialize:

sumX = 0

sumX2 = 0

sumY = 0

sumXY = 0

1. Calculate Required Sum

For i=1 to n:

sumX = sumX + Xi

sumX2 = sumX2 + Xi \* Xi

sumY = sumY + Yi

sumXY = sumXY + Xi \* Yi

Next i

1. Calculate Required Constant a and b of y = a + bx:

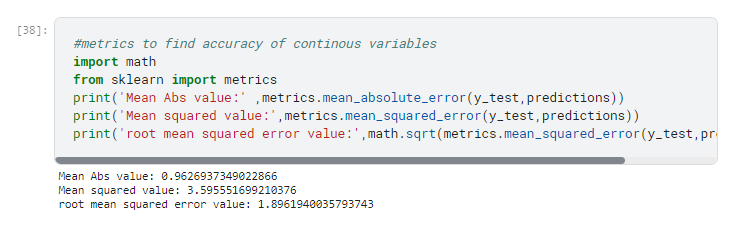
b = (n \* sumXY - sumX \* sumY)/(n\*sumX2 - sumX \* sumX)

a = (sumY - b\*sumX)/n

1. Display value of a and b
2. Stop

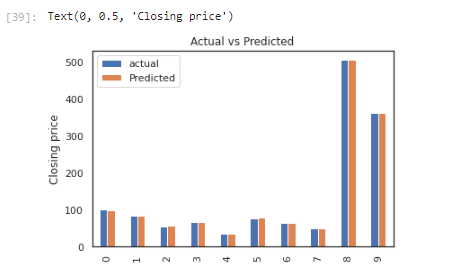
**Results**

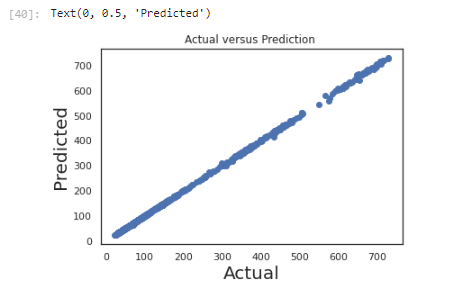
**Final Results**



Closing Graph

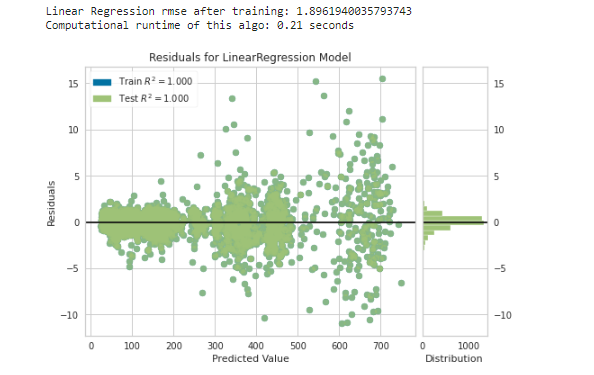
The barplot is one of the most widely used graphic types. It is shown how a numerical and a categorical variable relate to one another. The entities of the categorical variable are each represented by a bar. The length of the bar is proportional to its numerical value.

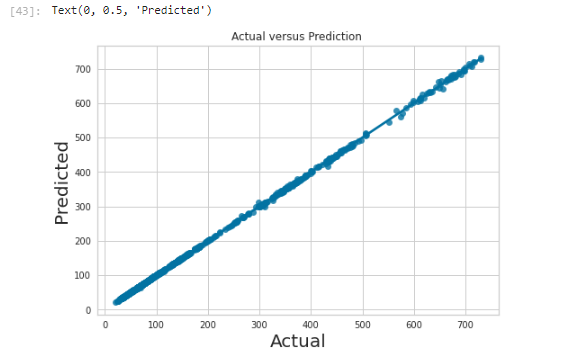




Residual Plot

The discrepancy between the values of the fitted response and the observed response is displayed on a residual plot. The null residual plot, which represents the ideal residual plot, exhibits a random distribution of points that form a band around the identity line that is roughly constant in width.





**Project Management**

**Implementation status report**

**Work completed:**

|  |  |  |
| --- | --- | --- |
| **Description/Task** | **Responsibility - Person** | **Contributions - percentage** |
| Data Read and Preprocessing | Cleaning of the data and data processing - - Naga Tulasi Parvathi Veerla | 30% |
| Data Analysis | Implementation of various data visualizations –  Gowthami Mulpuri | 35% |
| Data Split and Data model Creation (Linear Regression) | Splitting of data, linear regression creation - Harshitha Kolli | 35% |
| Prediction and Result Analysis | Model prediction using train set and comparison of acutual and predicted values and Analysis of the Result  -Harshitha Kolli |  |

**Github Link**

[**https://github.com/HarshithaKolli98/DIOR**](https://github.com/HarshithaKolli98/DIOR)

**References/Bibliography**

1. <https://scholar.google.com/citations?view_op=view_citation&hl=en&user=jj33szgAAAAJ&citation_for_view=jj33szgAAAAJ:d1gkVwhDpl0C>
2. Forecasting with artificial neural networks: the state of the art, International Journal of Forecasting 14 (1998) 35–62, by G. Zhang, B.E. Patuwo, and M.Y. Hu <https://www.researchgate.net/publication/222489987_Forecasting_With_Artificial_Neural_Networks_The_State_of_the_Art>
3. G. Zhang, B.E. Patuwo, and M.Y. Hu, "Forecasting with artificial neural networks: The State of the Art," International Journal of Forecasting, 14, no. 4, (1998), pp. 35–62.

<https://scirp.org/reference/referencespapers.aspx?referenceid=1452987>

1. K. Kim, Support vector machine-based financial time series forecasting. 2003, Neurocomputing 55: 307–319

<https://www.sciencedirect.com/science/article/abs/pii/S0925231203003722>

1. A comparison of linear regression and support vector regression is presented in the online proceedings of the 2016 International Conference on Green Engineering and Technologies (IC-GET). S. Kavitha, S. Varuna, and R. Ramya.

<https://www.researchgate.net/publication/316727392_A_comparative_analysis_on_linear_regression_and_support_vector_regression>

1. "Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining," Queensland University of Technology, Wanzhong Yang.  July 2009 technology

<https://eprints.qut.edu.au/30398/1/Wanzhong_Yang_Thesis.pdf>

1. R.V. Argiddi and S. S. Apte, "A Study of Association Rule Mining in Fragmented Item-Sets for Transaction Outcome Prediction in Stock Trading Systems"  IJCET-2012.

<https://www.semanticscholar.org/paper/A-STUDY-OF-ASSOCIATION-RULE-MINING-IN-FRAGMENTED-OF-Argiddi-Apte/3b3218c6687bae19cf5a7901469e4858304f9bf7>

1. H. Kunhuang and T. H. K. Yu (2006). the use of neural networks to predict ambiguous time data. Physical A, 363(2), 481-491, Statistical Mechanics and Its Applications.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4887655/>

1. Wikipedia Developer, "About Stock Market, Stock Brokerages, and Stock Exchange," available at https://en.wikipedia.org/ wiki/Stock market

<https://www.researchgate.net/publication/354085600_Stock_Market_Analysis_Using_Linear_Regression_and_Decision_Tree_Regression>

1. Amazon, "Collected data from Amazon," available at finance.yahoo.com/quote/AMZN/history?p=AMZN.

<https://www.researchgate.net/publication/354085600_Stock_Market_Analysis_Using_Linear_Regression_and_Decision_Tree_Regression>