GOVERNMENT OF KARNATAKA

DEPARTMENT OF COLLEGIATE AND TECHNICAL EDUCATION BENGALURU-560001

GOVERNMENT RESIDENTIAL POLYTECNIC FOR WOMEN SHIVAMOGGA-577205

CAPSTONE PROJECT REPORT ON

"Stock Market Price Prediction Using Machine Learning"

SUBMITTED BY

RAMYA **135CS21042**

Under The Guidance Of

Sri. Hanumanthappa.G

M.Tech, (Ph.D).,

Cohort Owner

Department Of Computer Science and Engineering 2023-24

GOVERNMENT RESIDENTIAL POLYTECNIC FOR WOMEN

SHIVAMOGGA-577205

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

2023-2024

CERTIFICATE

Certified that is caption project report entitled **"Stock Market Price Prediction Using Machine Learning"** which is being submitted by **Ms. Ramya Reg No. 135CS21042,** bonafide student in partial fulfillment for the award of diploma in computer science and engineering during the year 2023-2024 is recorded of students own work credited under my guidance. It is certified that all correction / suggestions indicated assignment have been incorporated is the report and one copy of it's been deposited in polytechnic library.

The caption project report has been approved as it satisfied the academy requirements in respect of caption project work prescribed for the side diploma. It is understood that by this certificate the undersigned do not endure or approve any statement made, opinion expressed or conclusion drawn there is but approve the capstone only the purpose for which it is submitted.

Cohort Owner	HOD	Principal
Sri. Hanumanthappa G	**Smt. Jyothi B K**	**Sri. D Ravinayak**
M.Tech, (Ph.D).,	BE, M.Tech	BE, (EC) M.Tech (H.R.D), MISTE, MISTD

Name and Signature of Examiners:

DEPARTMENT OF COLLEGIATE AND TECHNICAL EDUCATION
BENGALURU

GOVERNMENT RESIDENTIAL POLYTECNIC FOR WOMEN

SHIVAMOGGA-577205

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

2023-2024

BONAFIDE CERTIFICATE

Certified that this project report **"Stock Market Price Prediction Using Machine Learning"** is the bonafide work **Ms. Ramya Reg No. 135CS21042,** of "GOVERNMENT RESIDENTIAL POLYTECNIC FOR WOMEN SHIVAMOGGA" institution who carried out the project work under my supervision.

Cohort Owner

Sri. Hanumanthappa G
M. Tech, (Ph.D).,

Dept. of CS&E
GRWPT - Shivamogga

Head of The Department

Smt. Jyothi B K
BE, M.Tech

Dept. of CS&E
GRWPT - Shivamogga

DEPARTMENT OF COLLEGIATE AND TECHNICAL EDUCATION
BENGALURU

GOVERNMENT RESIDENTIAL POLYTECNIC FOR WOMEN

SHIVAMOGGA-577205

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

2023-2024

Declaration

We hereby declare that the Capstone Project Phase – 1 entitled **Ramya** has been carried out by under guidance of **Hanumanthappa G** and submitted in partial fulfilment of the course requirements for the award of degree in Computer Science and Engineering Of Government Residential Polytechnic For Women during the academic semester January – May 2024. The matter embodied in this report has not been submitted to any other university or institution for the award of the any degree

Reg.No	Name	Signature
135CS21042	Ramya	

ACKNOWLEDGEMENT

We have taken opportunity to express our guidance and indebtedness to all concerned people for helping us to bring out this report. Our report can be successful with the material assistance of several people so we can try and acknowledge heartly their assistance for this project work.

We greatly thankful to our guide Mr. Hanumanthappa G M.Tech, (Ph.D)., Lecturer in CS&E department, who guide me in preparing this project work. We are grateful for this valuable guidance, direction, and encouragement without which our project work would have not been completed.

We greatly thankful to our guide Smt.Jyothi BE, M.Tech. Head of the department for granting me all facilities to complete this project work.

We greatly thankful to our Principal Mr.D Ravi Nayak BE, (EC) M.Tech, (H.R.D), MISTE, MISTD for this continuous support to complete this project work.

We wish to express my sincere thanks to all the staff members of CS&E department for counselling and assistance throughout the report.

Finally, we wish to express my thanks to all other teaching and non- teaching staffs of the department and all those who helped directly or indirectly for the success of the project work.

Name of the Student

Ramya

ABSTRACT

Stock market price prediction is a challenging task due to its inherent complexity and non-linearity. In recent years, machine learning techniques have gained popularity for their ability to uncover patterns in financial data and make accurate predictions. This paper presents a comprehensive review of machine learning algorithms and methodologies used for stock market price prediction. It discusses various features, including technical indicators, fundamental analysis, sentiment analysis, and macroeconomic factors, that influence stock prices. Furthermore, it explores the application of supervised, unsupervised, and reinforcement learning algorithms in predicting stock prices. Additionally, the paper examines the challenges and limitations of existing methods and proposes potential avenues for future research. Overall, this paper provides insights into the current state of stock market price prediction using machine learning and offers recommendations for improving prediction accuracy and robustness. machine learning algorithms.

EXECUTIVE SUMMERY

This executive summary provides a concise overview of the research conducted on stock market price prediction using machine learning techniques. The study explores the complexity of predicting stock prices and the growing interest in leveraging machine learning algorithms for this purpose.

Key Points:

- **Complexity of Stock Market Prediction: -** Predicting stock market prices is challenging due to the presence of numerous variables, market dynamics, and non-linear relationships.

- **Machine Learning Approach: -** Machine learning techniques offer a promising solution for analyzing large volumes of financial data and identifying patterns that influence stock prices.

- **Feature Engineering: -** Various features such as technical indicators, fundamental analysis metrics, sentiment analysis from news and social media, and macroeconomic indicators play crucial roles in predicting stock prices.

- **Algorithmic Methods: -** Supervised learning, unsupervised learning, and reinforcement learning algorithms are applied to predict stock prices. Each approach has its strengths and limitations.

- **Challenges and Limitations: -** Despite advancements in machine learning, challenges remain, including data quality issues, over-fitting, market noise, and the unpredictable nature of financial markets.

- **Future Directions: -** Future research should focus on improving prediction accuracy, developing robust models that can adapt to changing market conditions, incorporating alternative data sources, and addressing ethical considerations in algorithmic trading.

Expected Benefits of Stock Market Price Prediction Using Machine Learning:

- **Improved Decision Making: -** Machine learning models can provide traders, investors, and financial institutions with more accurate predictions of stock prices, enabling them to make informed decisions regarding buying, selling, or holding assets.

- **Risk Mitigation: -** By leveraging machine learning algorithms to predict stock prices, investors can better assess and mitigate risks associated with their investment portfolios. This can lead to more prudent investment strategies and reduced exposure to market volatility.

- **Enhanced Efficiency: -** Automated stock market prediction systems powered by machine learning can analyse vast amounts of financial data rapidly, allowing traders to react swiftly to changing market conditions and capitalize on emerging opportunities.

- **Cost Savings: -** Accurate stock market predictions can help investors minimize losses and maximize profits, leading to significant cost savings over time. Additionally, the automation of prediction tasks can reduce the need for manual analysis, saving time and resources.

- **Competitive Advantage: -** Firms that successfully implement machine learning-based stock market prediction systems gain a competitive edge by staying ahead of market trends, identifying profitable investment opportunities, and delivering superior returns to their clients.

- **Innovation and Research: -** Continued research and development in the field of stock market prediction using machine learning foster innovation and drive advancements in financial technology. This can lead to the discovery of new predictive features, algorithms, and methodologies, benefiting the broader financial industry.

TABLE OF CONTENTS

CONTENT **PAGE NO.**

LIST OF FIGURES:

List of Tables

Chapter 1

INTRODUCTION

1.1 Introduction of Stock Market Price Prediction Using Machine Learning

The unpredictable nature of financial markets has long fascinated investors, economists, and researchers alike. Stock market price prediction remains a challenging endeavor due to the complex interplay of various factors such as economic indicators, investor sentiment, geopolitical events, and market psychology. Traditional quantitative models often struggle to capture the intricacies of these dynamics, leading to sub optimal predictions and investment outcomes.

In recent years, the rise of machine learning has revolutionized the field of financial analysis by offering powerful tools to extract insights from large volumes of data. Machine learning algorithms excel at identifying patterns, correlations, and anomalies in complex datasets, making them well-suited for stock market prediction tasks. By leveraging techniques from supervised, unsupervised, and reinforcement learning, researchers and practitioners have sought to enhance the accuracy and reliability of stock price forecasts.

This paper delves into the application of machine learning methodologies for stock market price prediction, aiming to provide a comprehensive understanding of the current landscape and future directions in this rapidly evolving field. We explore the key challenges faced in traditional stock market prediction models and discuss how machine learning techniques offer novel solutions to address these challenges. Additionally, we examine the various features and data sources utilized in stock price prediction, ranging from technical indicators and fundamental analysis metrics to alternative data streams such as social media sentiment and news sentiment.

Furthermore, we review a diverse array of machine learning algorithms employed for stock market prediction, including regression models, decision trees, neural networks, support vector machines, clustering algorithms, and reinforcement learning approaches. Each algorithm has its unique strengths and limitations, and understanding their applicability to different prediction tasks is essential for developing robust and accurate forecasting models.

Despite advances in technology and data analytic, predicting stock prices with high accuracy remains elusive. Market dynamics are influenced by a multitude of factors, including unpredictable events and human behavior, making it difficult to develop models that consistently outperform the market. As such, while stock market prediction can provide valuable insights for investors, it's important to approach it with caution and to consider multiple sources of information when making investment decisions.

1.2 Scope of the capstone project

1.2.1 Problem Statement: - Stock market prices are influenced by a multitude of factors, making them challenging to predict accurately with traditional methods. Our goal is to leverage the capabilities of Machine Learning to create a robust predictive model that outperforms existing approaches, providing investors and financial analysts with a valuable tool for decision-making

1.3 Objectives: -

- **Model Development: -** Build a sophisticated ML model that can effectively predict stock prices by learning from historical data patterns.

- **Algorithm Selection: -** Explore and compare various ML algorithms, such as regression models, decision trees, and neural networks, to identify the most suitable approach for our specific prediction task.

- **Feature Engineering: -** Analyse and select relevant features, including historical stock prices, trading volumes, and external factors, to enhance the model's predictive capabilities.

- **Data Preprocessing: -** Clean and preprocess data to ensure its quality and suitability for training the ML model, addressing issues like missing values and outliers.

- **Hyper parameter Tuning: -** Optimize the model's performance by fine-tuning its hyper parameters, ensuring it adapts well to different market conditions.

- **Evaluation Metrics: -** Define and employ appropriate evaluation metrics to assess the model's accuracy, precision, and recall, ensuring a comprehensive understanding of its predictive capabilities

1.3.1 Expected Outcomes: -

A high-performing ML model capable of accurately predicting stock market prices.

Comparative analysis showcasing the superiority of the ML model over traditional forecasting methods.

Insights into the key factors influencing stock prices, as identified by the model. A user-friendly interface that facilitates seamless interaction with the prediction model.

Documentation and guidelines for stakeholders on using and interpreting the model's predictions.

1.3.2 Capstone project description: -

This project focuses on creating a smart computer program using machine learning to predict stock prices. By analyzing historical stock market data and incorporating external factors, the program aims to provide valuable insights to investors. The project includes building a user-friendly interface for easy interaction and aims to simplify stock market decision-making. Despite challenges like dealing with unpredictable market conditions, the goal is to empower users with a reliable tool for navigating the complexities of stock trading.

1.3.3 Capstone project Deliverable: -

1. Project Proposal
2. Data Collection and Preprocessing
3. Exploratory Data Analysis (EDA)
4. Machine Learning Model Implementation
5. Model Evaluation and Validation
6. Hyper parameter Tuning
7. Feature Importance Analysis
8. Real-time Predictions
9. Comparative Analysis
10. Interpret ability and Explain ability
11. Project Documentation
12. Presentation Slides
13. Stakeholder Training Materials
14. Project Summary and Reflection

1.3.4 Key milestones: -

- **Data Collection and Exploration:-**
  - Collect historical stock market data and relevant external factors.
  - Clean and preprocess the data to handle missing values and outliers.
  - Perform exploratory data analysis (EDA) to gain insights into the dataset.
- **Model Development: -**
  - Build the initial version of the machine learning model.
  - Train the model using historical data.
  - Evaluate the model's performance on a validation dataset.
  - Iterate and refine the model based on performance feedback.
- **Hyper parameter Tuning: -**
  - Conduct hyper parameter tuning to optimize the model's performance.
  - Evaluate the impact of different hyper parameter configurations.
  - Choose the best set of hyper parameters for the final model.
- **Model Evaluation and Optimization: -**
  - Evaluate the final model's performance on a test dataset.
  - Compare the model's predictions against traditional forecasting methods.
  - Generate performance metrics and visualizations for clear interpretation.
- **Documentation: -**
  - Document the entire project, including methodologies, challenges, and solutions.
  - Create user guides and documentation for stakeholders.
  - Generate a final project report summarizing key findings and outcomes.
- **Presentation and Demonstration: -**
  - Prepare a presentation summarizing the project for stakeholders.
  - Conduct a live demonstration of the model's functionality and user interface.

1.3.5 Constraints:

Limited Historical Data: -

Insufficient historical data can limit the model's ability to learn meaningful patterns.

Data Quality and Noise: -

Poor data quality, noise, or inconsistencies in the dataset can affect model performance.

Market Volatility: -

Stock markets can exhibit sudden and unpredictable volatility, challenging the stability of predictive models.

Lack of Causality Understanding: -

Machine learning models may not inherently understand the causal relationships driving stock prices.

Regulatory and Ethical Constraints: -

Regulatory restrictions may limit the use of certain data sources or impose constraints on the model's predictions.

Over fitting and Under fitting: -

Over-fitting to historical data may hinder the model's ability to generalize to new, unseen data.

Resource Constraints: -

Limited computational resources may impact the complexity and size of the model that can be developed.

Interpret-ability and Explain ability: -

Many machine learning models, especially complex ones, lack interpretability and explain ability.

Financial Market Dynamics: -

The dynamic nature of financial markets may make it challenging to capture all relevant factors influencing stock prices.

Estimated Capstone project Duration: 14 Weeks

Estimated Capstone project cost: 10,000 /-

Chapter - 2

CAPSTONE PROJECT PLANNING

2.1 Work Breakdown Structure

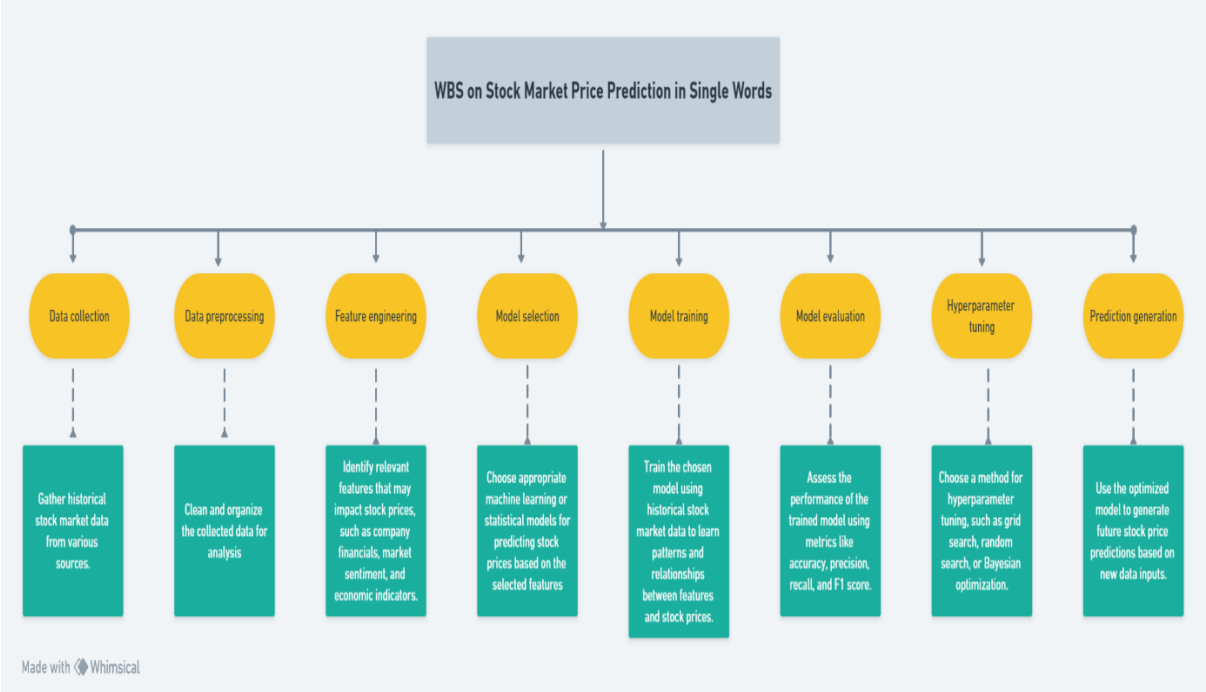

Fig 2.1: Work Breakdown Structure

- **Data Collection: -**
  - Gather historical stock market data for different companies from financial databases or APIs such as Yahoo Finance, Alpha Vantage, or Quandy.
  - Data collection from various sources such as financial databases, APIs, and historical stock market data.
  - Ensure the collected data is clean, reliable, and comprehensive for training machine learning models effectively.

- **Data Preprocessing: -**
  - Cleaning the data to handle missing values, outliers, and inconsistencies.
  - Feature engineering to extract relevant information from the raw data and create new features that can be used in machine learning models.

- Data normalization or standardization to bring all the features to a similar scale for better model performance

- Handling categorical variables through techniques like one-hot encoding or label encoding.
- Splitting the dataset into training, validation, and test sets for model evaluation.
- Exploratory data analysis (EDA) to gain insights into the distribution of features and relationships within the dataset before modelling.

- **Feature Selection: -**

  - Categorical variables using techniques like one-hot encoding or label encoding. Encoding
  - Feature scaling to ensure all features have the same impact on the model, such as using Min-Max scaling or standardization.
  - Creating new features through techniques like polynomial features, interaction terms, or domain-specific feature engineering.
  - Dimensional reduction using methods like principal component analysis (PCA) or feature selection algorithms to reduce the number of input variables and improve computational efficiency.
  - Time series feature engineering including lagging indicators, rolling statistics, and time-based transformations for capturing temporal patterns in stock market data

- **Model Selection: -**

  - Linear Regression
  - Decision Trees
  - Random Forest
  - Support Vector Machines (SVM)
  - Gradient Boosting Machines (GBM)

- **Model Training: -**

  - Splitting the dataset into training and testing sets for model evaluation.
  - Selecting appropriate machine learning algorithms such as linear regression, decision trees or random forests.
  - Tuning hyper parameters of the chosen models to optimize their performance.

- Model Testing: - Evaluating the models using metrics like mean squared error (MSE), root mean squared error (RMSE), and R-squared to assess predictive accuracy.
- Validating the trained models using cross-validation techniques to ensure generalization to new data.
- Iteratively refining the models by adjusting features and algorithms based on performance feedback.

- **Model Evaluation: -**

- Train the models on the training dataset and evaluate their performance using metrics like mean squared error (MSE) or R-squared.
- Compare the performance of different models to select the best-performing one for stock market price prediction.
- Validate the selected model on the testing dataset to ensure its effectiveness in predicting future stock prices accurately.

- **Hyper parameter Tuning: -**

- Grid Search
- Random Search
- Bayesian Optimization
- Fine-tune hyper parameters of the chosen model to optimize its predictive capabilities.

- **Deployment of Model for Prediction: -**

- Use the optimized model to generate future stock price predictions based on new data inputs.
- Validate the trained models on a separate test dataset to ensure their generalization capabilities.

2.2 Time - Line Schedule

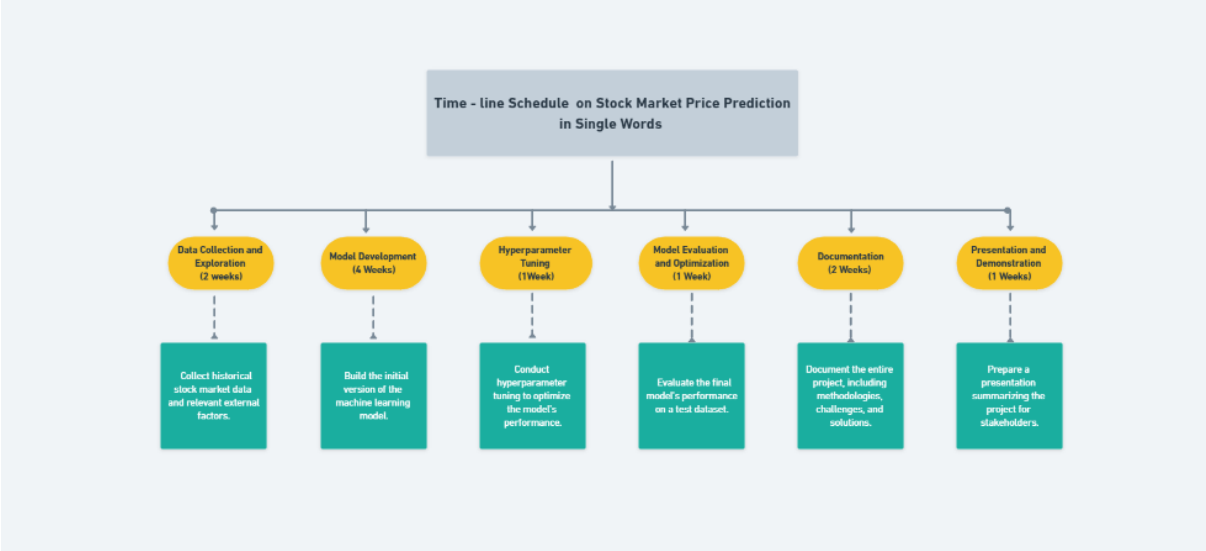

Fig 2.2: Time - Line Schedule

- **Data Collection and Exploration: -**
  - Collect historical stock market data and relevant external factors.
  - Clean and preprocess the data to handle missing values and outliers.
  - Perform exploratory data analysis (EDA) to gain insights into the dataset.

- **Model Development: -**
  - Build the initial version of the machine learning model.
  - Train the model using historical data.
  - Evaluate the model's performance on a validation dataset.
  - Iterate and refine the model based on performance feedback.

- **Hyper parameter Tuning: -**
  - Conduct hyper-parameter tuning to optimize the model's performance.
  - Evaluate the impact of different hyper-parameter configurations.
  - Choose the best set of hyperparameters for the final model.

- **Model Evaluation and Optimization: -**
  - Evaluate the final model's performance on a test dataset.

- Compare the model's predictions against traditional forecasting methods.
- Generate performance metrics and visualizations for clear interpretation.

- **Documentation: -**
  - Document the entire project, including methodologies, challenges, and solutions.
  - Create user guides and documentation for stakeholders.
  - Generate a final project report summarizing key findings and outcomes.

- **Presentation and Demonstration: -**
  - Prepare a presentation summarizing the project for stakeholders.
  - Conduct a live demonstration of the model's functionality and user interface.
  - Address questions and gather feedback from the audience.

2.3 Cost Breakdown Structure

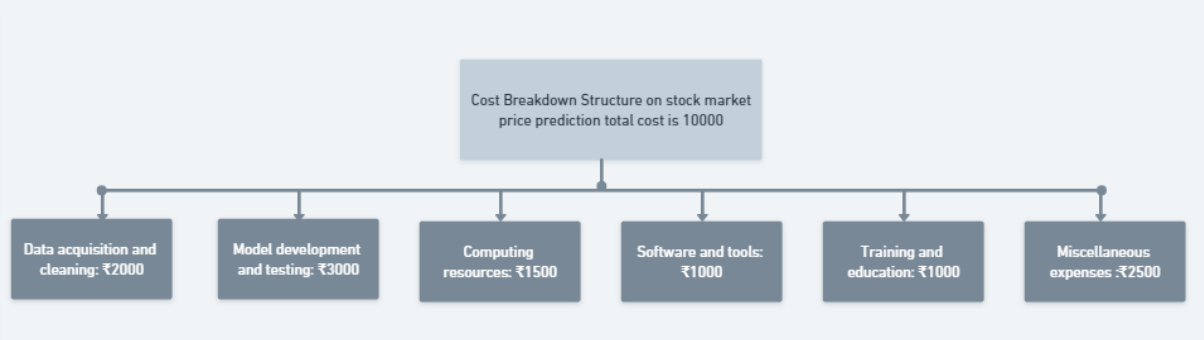

Fig 2.3: Cost Breakdown Structure

Data acquisition and cleaning: - This cost category covers expenses for obtaining stock market data, including purchasing datasets or accessing APIs. It also includes activities like cleaning and organizing the data to ensure accuracy for effective price prediction.

Model development and testing: - This cost category covers the expenses related to the development and testing of predictive models for stock market price prediction. It may include costs for data collection, algorithm development, software tools, and testing procedures.

Computing resources: - This includes the cost of purchasing or renting computing hardware such as servers, desktops, or laptops, as well as any associated software and networking infrastructure required

for stock market price prediction. These resources are essential for running data analysis, machine learning algorithms, and other computational processes to make accurate predictions

Software and tools: - This cost category covers the expenses related to purchasing and using software and tools for stock market price prediction, such as data analysis software, trading platforms, financial modeling tools, and any other relevant technology required for the prediction process.

Training and education: - This cost category encompasses expenses for training materials, workshops, and certifications aimed at enhancing skills in stock market price prediction, including specialized courses to improve predictive modeling techniques.

Miscellaneous expenses: - This category includes costs such as subscription fees for data sources, cloud storage, and any other smaller expenses that may include.

Documentation expenses: - Research and data gathering costs, Software and tool expenses, Training and education costs, Consultation fees, Travel expenses for meetings, Files and Binding cost, etc...

2.4 Risk Analysis

- Identify potential risks associated with the stock market, such as market volatility, economic downturns, and geopolitical events.
- Consider data quality and availability of historical stock market data for training machine learning models.
- Evaluate the risk of over fitting the model to historical data, which may not accurately represent future market conditions.
- Assess the potential impact of regulatory changes or policy decisions on stock prices.
- Account for the risk of model misinterpretation or bias leading to inaccurate predictions and investment decisions.

- Consider operational risks related to implementing and maintaining machine learning models in a real-time trading environment.

- Evaluate cyber security risks associated with handling sensitive financial data used in the project.

2.5 Design Specification

Chosen System Design

Component diagrams: - Component diagrams on Linear Regression Model: - The diagram showcases the outcomes of implementing linear regression on a dataset through a component-based approach, illustrating the results obtained from the machine learning algorithm's application to the training dataset.

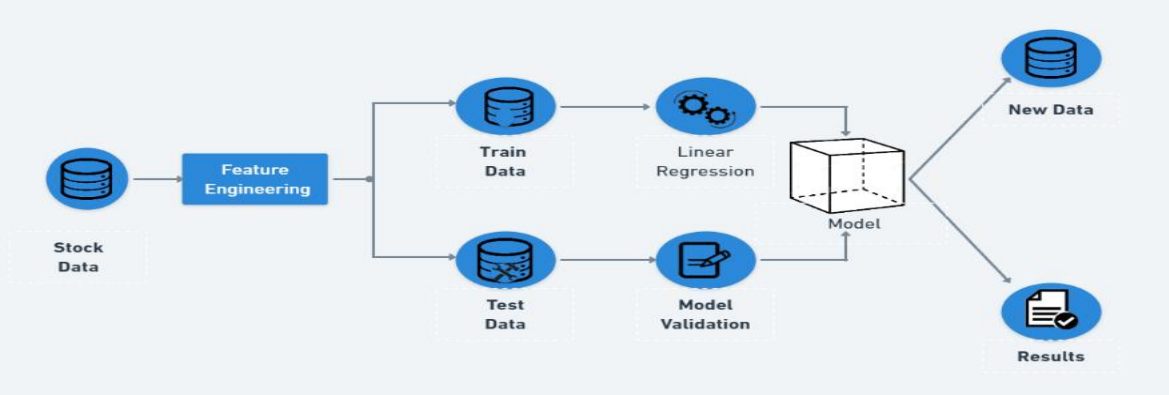

Fig 2.4: Component diagrams using Linear Regression

Component diagrams on Random Forest Regression Model: - Through a component-based approach, the diagram elucidates the various stages of random forest regression, showcasing how ensemble learning and feature selection contribute to the model's predictive performance.

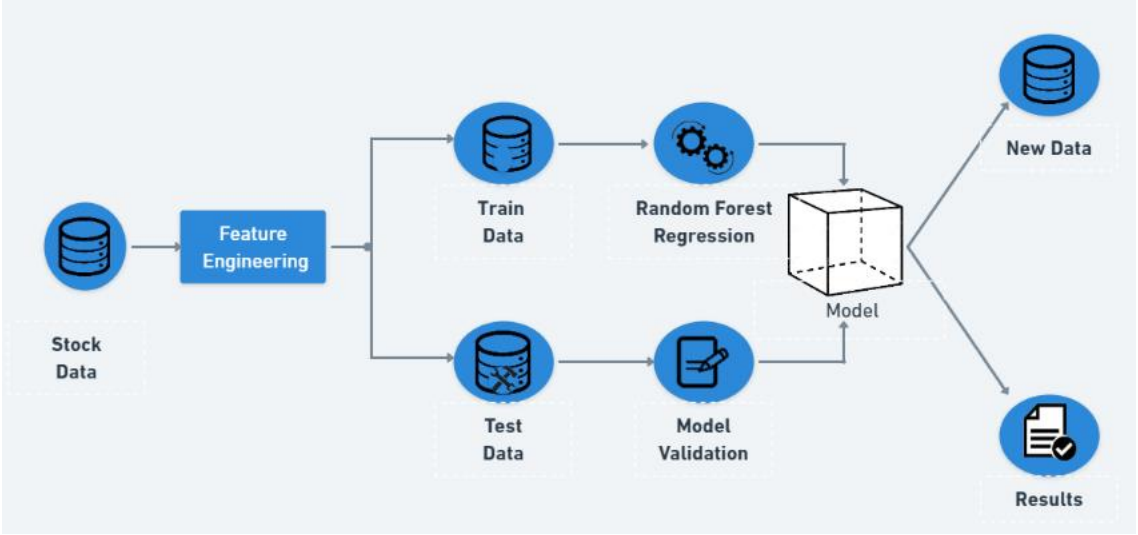

Fig 2.4: Component diagram using Random Forest Regression

Component diagrams on Decision Tree Model: - The component diagram provides a comprehensive overview of decision tree construction, showcasing how the algorithm recursively selects features and thresholds to optimize predictive performance.

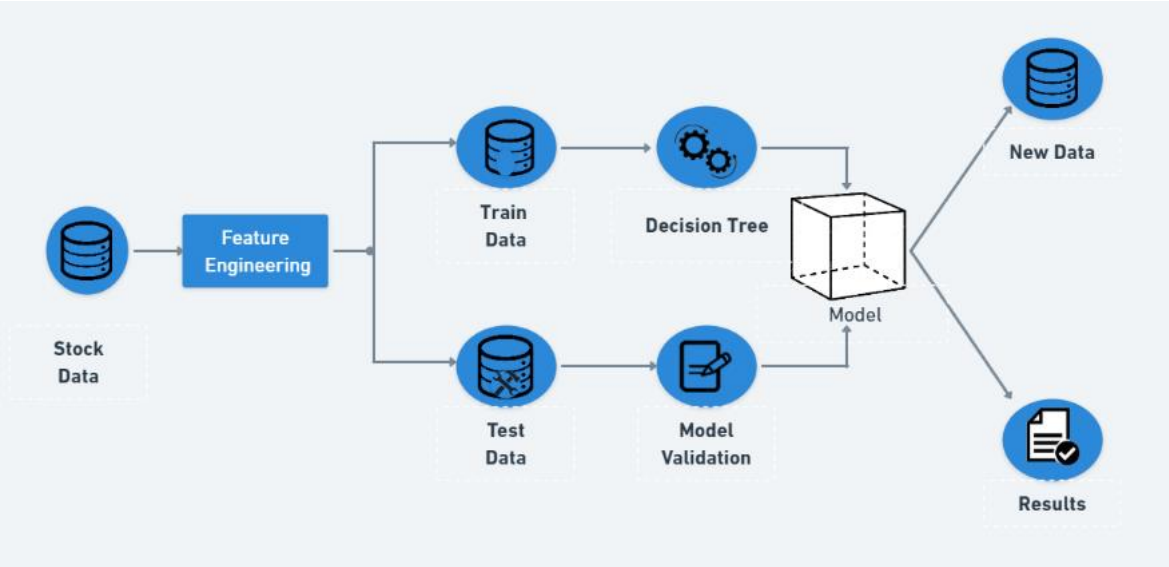

Fig 2.4: Component diagram using Decision Tree

Component diagrams on XGBoost Regression Model: - The diagram showcases the outcomes of implementing XGBoost Regression on a dataset through a component-based approach, illustrating the results obtained from the machine learning algorithm's application to the training dataset.

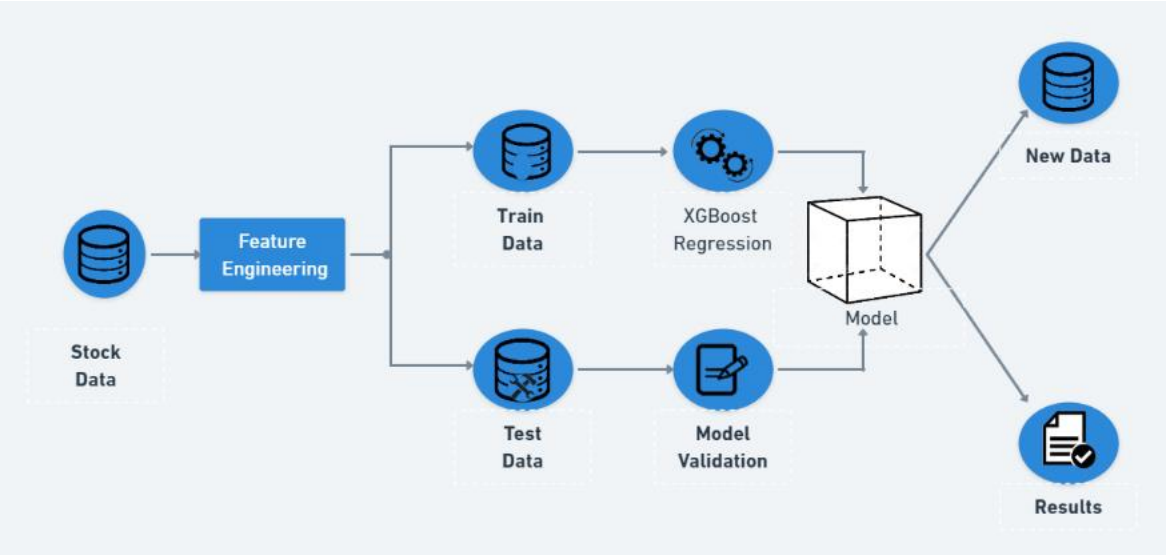

Fig 2.4: Component diagram using XGBoost Regression

Chapter - 3

METHODOLOGY

When working with a stock market dataset to predict continuous values like stock prices or returns, it's common to consider algorithms such as Linear Regression, Random Forest, Decision Tree, Support Vector Regression (SVM), XGBoost Regression, and Ridge Regression.

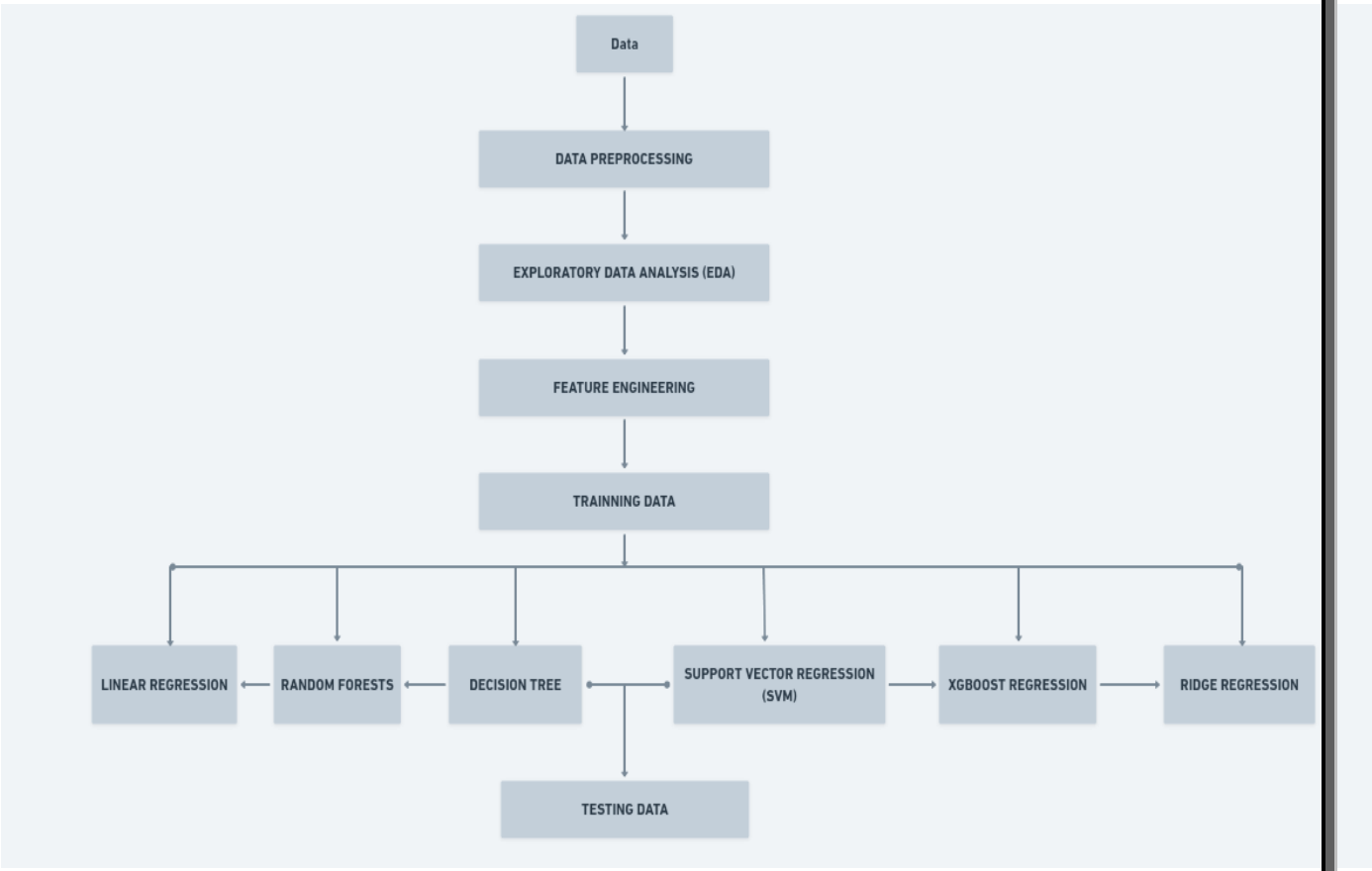

Fig 3.1: Working procedure of proposed model

- **Data collection: -** The step of every project is to collect the data.
- **Data per-processing: -** In this step we basically clean our dataset for example check for any missing value in the dataset, if present then handle the missing values. In our dataset attributes like Item Weight and Outlet Size had the missing value.

- **EDA: -** This part is considered as one of the most important parts when it comes to data analysis. To gain important insights of our data one must need to do exploratory data analysis. Here in our project we used two libraries i.e. klib and dtale library.
- **Tested various algorithms: -** Then various algorithms like simple LR, XGboost algorithm were applied to find out which algorithm can be used to predict the stock price.
- **Building the model: -** After completing all the previous phases which are mentioned above, now our dataset is ready for further phases that is to build the model. Once we built the model now it is ready to be used as a predictive model Stock Price.

Chapter 4

SYSTEM DEVELOPMENT

4.1 ALGORITHMS EMPLOYED

4.1.1 LINEAR REGRESSION (LR)

As we know Regression can be termed as a parametric technique which means we can predict a continuous or dependent variable on the basis of a provided datasets of independent variables.

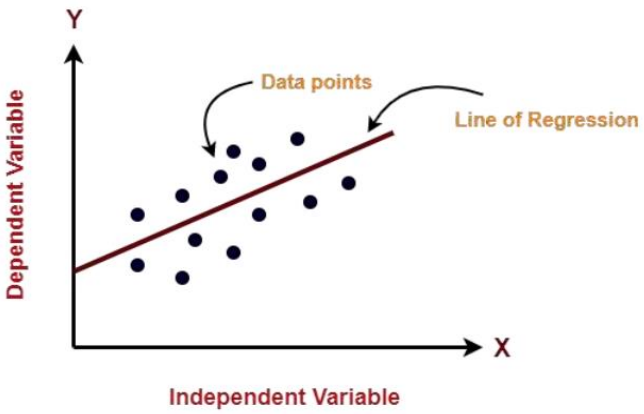

Fig 4.1: Given figure represent line regression

4.1.2 RANDOM FOREST REGRESSION

Random Forest is a tree-based bootstrapping algorithm based on that tree that includesacertain number of decision trees to build a powerful predictive model. Individual learners, aset of random lines and a randomly selected few variables often create a tree of choice. Thefinal prediction may be the function of all predictions made by each learner. In the event of aregression. The final prediction may be the meaning of all the predictions.

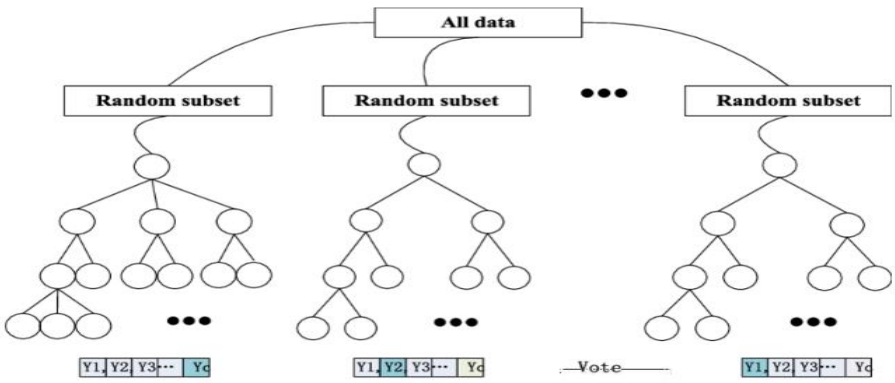

Fig 4.2: Flowchart of Random Forest Regression

4.1.3 XGBOOST REGRESSION

XGBoost stands for Xtreme Gradient Boosting. The implementation of an algorithm designed for the efficient operation of computer time and memory resources. Boosting is sequential process based on the principle of the ensemble. This includes a collection of lower earners as well improves the accuracy of forecasts. No model prices heavy for any minute, based on the results of the previous t-speed. Well-calculated results are given less weight, and the wrong ones are weighed down. With this algorithm system The XGBoost model uses stepwise, ridge regression internally, automatically selecting features as well as deleting multicollinearity.

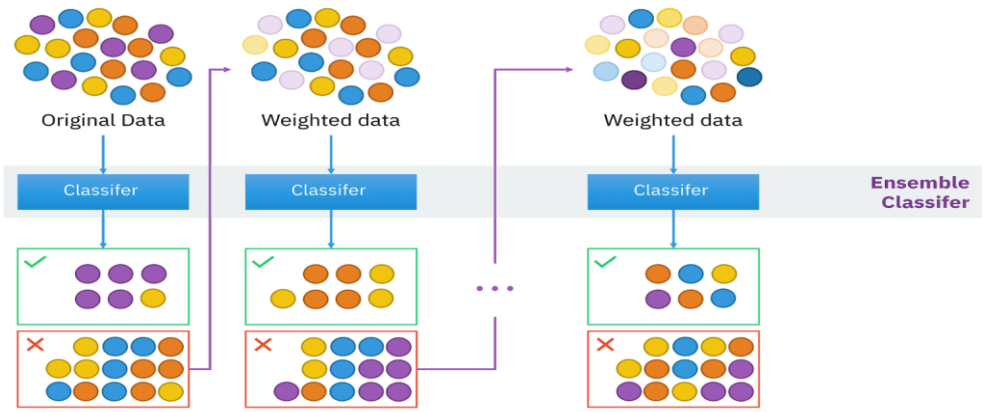

Fig 4.2: Flowchart of XGBoost Regression

4.1.4 DECISION TREE

A decision tree is a graphical representation of a decision-making process. It consists of nodes that represent decision points, branches that symbolize the possible outcomes of decisions, and leaves that denote the final outcomes, or classifications. Decision trees are used in various fields including machine learning and business to model and visualize decision scenarios. They work by breaking down a complex decision-making process into a series of simpler decisions, making it easier to understand and interpret the consequences of different choices. Each node in the tree represents a specific choice or decision based on certain conditions, and the branches emanating from each node represent the possible outcomes or paths resulting from those decisions. This structured approach helps in analysing and making decisions based on various criteria and condition.

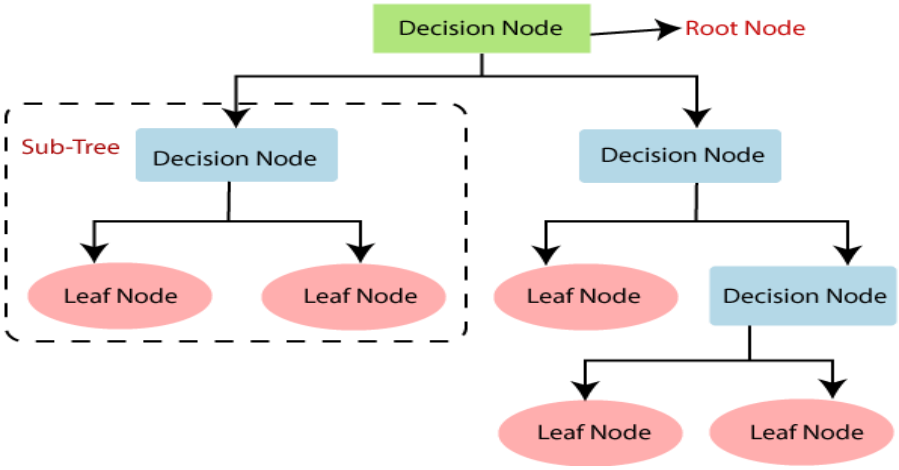

Fig 4.4: Decision tree

4.2 Data engineering pipeline

4.2.1 Data collection

For this model we taken "**TATA Motors Limited - Stock" .** This dataset contains the historical stock prices of Tata Motors Limited in **INR** on a daily basis. The stock prices was collected through kaggle.

.

- This dataset as 7 features that are **Date, Open, High, Low & ADJ Close Volume**
- This dataset as stock price records from **2006 to 2024**, it means **18** year's.
- And it as **4233** rows and **7** columns/feature

Loading Dataset

```
1 stock= pd.read_csv(r"TATAMOTORS.NS.csv")# here we storing dataset in stock named variable
```

```
1 stock.head() #this data set has 7 features
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2006-12-01	173.042435	173.042435	155.268692	161.515854	141.449173	10001085.0
1	2006-12-04	162.023666	170.512909	162.023666	169.085266	148.078171	18491096.0
2	2006-12-05	170.167969	170.867416	167.293518	169.209824	148.187225	6699838.0
3	2006-12-06	167.121048	170.167969	165.568832	168.912796	147.927109	4822862.0
4	2006-12-07	168.634933	171.701019	167.322250	171.068634	149.815109	3963218.0

```
1 stock.shape # the original dataset has 4233 rows and 7 columns
```

```
(4233, 7)
```

Fig 4.5: Dataset

4.2.2 Exploratory Data Analysis

- In this step, we use weary plotting techniques to visualize how the stock price varies between high and low in the Open and Close columns.

- We understand that from 2006 to 2010, the stock price had low volume. However, starting from the beginning of 2010 through 2019, the stock price showed a significant increase

- From the last month of 2019 to 2022, the stock price of Tata Motors significantly decreased.
- But at the start of 2023, the stock price did not decrease and instead continued to rise.

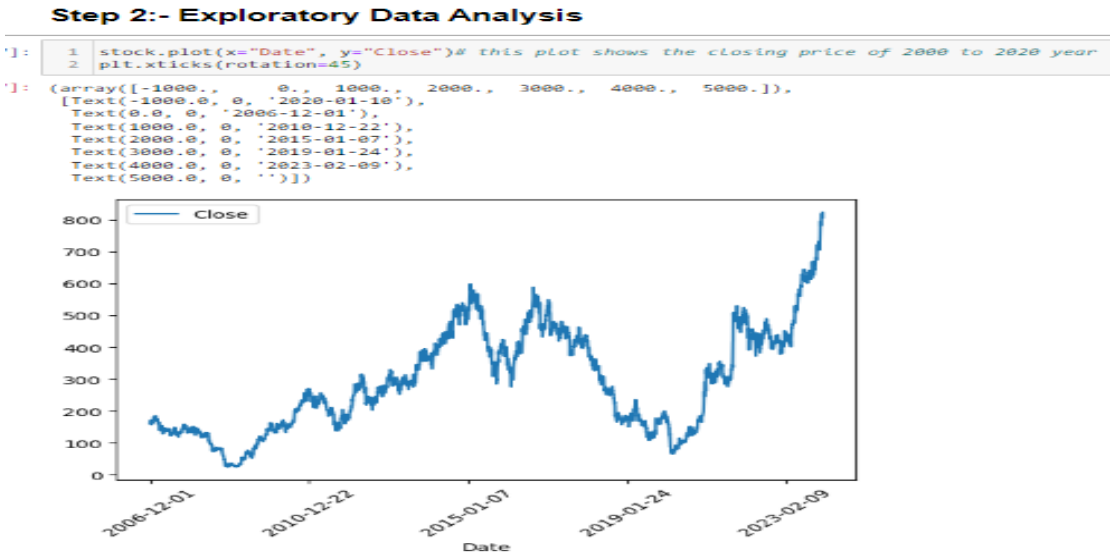

Fig 4.6: EDA of data

4.2.3 Data Preprocessing

- In this step, we Handle the missing values in this dataset
- In this dataset has total 42 missing values have
- Expect Data column all the Column have 7 missing values
- we check the duplicated value on this data set but it has 0 duplicate
- To handle the missing value, it has two ways that is drop the missing value or replace the dataset
- We replace the missing value on volume column using median value
- And the last we drop the null values using “dropna” function correct the words

Step 3:- Data preprocessing

:- Data Cleaning & Headling with missing values

```
stock.isnull().sum().sum()#On this dataset has total 42 missing values have
```

```
42
```

```
stock.isnull().sum().sum()#On this dataset has total 42 missing values have
```

```
42
```

```
stock['Volume'].max() # I volume max value is 390577839.0,so we fill the nan value
```

```
390577839.0
```

```
stock['Volume'].replace(np.nan,390577839.0,inplace=True)
```

```
stock['Volume'].isnull().sum()# Now volume column is 0 missing value
```

Fig 4.7: Data Preprocessing

4.2.4 Data Integration and Reduction

- In this step, we reduce the **Date** and **Open** features because of the **"X"** variable contains independent feature, so we drop the dependent columns
- **"Y", We** only use the dependent column / feature that is **Open**

Step 4:- Data Integration and Reduction ¶

```
x = stock.drop(columns=['Date', 'Open'])  # Assuming 'Date' is not a feature and 'Open' is the target variable
y = stock['Open']
```

```
stock.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2006-12-01	173.042435	173.042435	155.268692	161.515854	141.449173	10001085.0
1	2006-12-04	162.023666	170.512909	162.023666	169.085266	148.078171	18491096.0
2	2006-12-05	170.167969	170.867416	167.293518	169.209824	148.187225	6699838.0
3	2006-12-06	167.121048	170.167969	165.568832	168.912796	147.927109	4822862.0
4	2006-12-07	168.634933	171.701019	167.322250	171.068634	149.815109	3963218.0

Fig 4.8: Integration and Reduction

4.2.5 Feature Scaling

In this scenario, we use Min-Max Scaler technique our data due to the presence of lengthy values such as volume = 10001085.0. This technique is utilized to compress the range of values into a 0 - 1 scale, facilitating effective model training.

Step 5:-Feature Scaling

```
MS=MinMaxScaler(feature_range=(0,1))
```

```
data_training_array=MS.fit_transform(x_train)
```

```
data_training_array
```

```
array([[0.41545305, 0.40311382, 0.40188542, 0.40192044, 0.0271194 ],
       [0.49137586, 0.49307695, 0.49516294, 0.49239597, 0.00972565],
       [0.30759333, 0.29771251, 0.30836786, 0.29931167, 0.04739484],
       ...,
       [0.16832244, 0.16776143, 0.16963911, 0.17117327, 0.04230558],
       [0.52546926, 0.52034572, 0.51690502, 0.51645917, 0.08058766],
       [0.15898582, 0.15845178, 0.15984268, 0.14926246, 0.05594969]])
```

Fig 4.9: Feature Scaling

4.2.6 Data Splitting

The dataset was split into training and testing sets using the train_test_split() function, with 80% of the data allocated to the training dataset. This training dataset includes features such as high, low, close, adjusted close, and volume, which are used to train the model. The remaining 20% of the data forms the test dataset, where the open feature serves as the target variable for model evaluation.

step 6:- Data Splitting

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

```
x_train.shape
```

```
(3380, 5)
```

Fig 4.10: Data splitting

4.2.7 Training the model

In our analysis, we evaluated four distinct types of models: Linear Regression, Random Forest, Decision Tree, and XGBoost Regression. Through this comparison, we sought to determine which among these models exhibits superior performance under specific evaluation metrics and conditions, providing valuable insights into their respective strengths and weaknesses within our dataset.

For example: -

Step 7:- Model Training

```
In [26]: rf= RandomForestRegressor(n_estimators=100, random_state=42)
```

```
In [27]: rf.fit(x_train,y_train)
```

```
Out[27]: RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Fig 4.11: Model training

Example of Prediction of next day stock price: -

Here we Predicted next day's stock price, it will train with previous stock price data and give the prediction of next day's stock of dependent feature.

```
next_day_features = x.iloc[[-1]]  # Assuming the last row contains the latest features
next_day_opening_price = lr.predict(next_day_features)
print("Predicted next day's opening price:", next_day_opening_price)
```

```
Predicted next day's opening price: [821.58884098]
```

Fig 4.12: Prediction of next day stock price

4.2.8 Model evaluation and comparison

Here we compare 4 models they are Linear Regression, Decision Tree, Random Forest Regression, XGBoost Regression. The R2 Score of Linear Regression is more than other models and mean error is less compared to other models. So, Linear Regression is the best model.

Models	R2 Score	Mean Square Error
Decision Tree	0.9980859281227569	34.8648728869900
XGBoost	0.9252252253784354	17.6295461563584
Linear Regression	0.9996879588134863	7.92191054553125
Random Forest Regression	0.9994289219415903	14.32551860948762

TB 4.1: Model evaluation and comparison

Chapter - 5

Business Aspects

5.1 Business Aspects

The business aspects of your machine learning project on stock market price prediction are multifaceted and tied closely to the financial industry's dynamics and economic outlook. Here's a brief overview of the market and economic context for your capstone project

Market Overview: The stock market plays a pivotal role in the global economy, serving as a barometer of economic health and investor sentiment. In recent years, there has been a growing interest in using machine learning techniques to predict stock prices due to the potential for generating actionable insights and optimizing investment strategies. This project aims to leverage historical stock market data and advanced predictive models to forecast future price movements, providing valuable information for traders, investors, and financial institutions.

Economic Outlook: The economic outlook for this capstone project is influenced by various factors, including macroeconomic indicators, geopolitical events, monetary policy decisions, and industry-specific trends. Economic conditions such as GDP growth, inflation rates, interest rates, and employment levels directly impact stock market performance and investor behavior. Understanding these economic dynamics is crucial for developing accurate and reliable machine learning models that can adapt to changing market conditions and provide timely predictions.

5.1.1 Key Business Aspects:

- **Risk Management:** Predictive models can assist in identifying and managing investment risks by providing insights into potential price fluctuations and market volatility.
- **Investment Strategies:** The project can help in developing data-driven investment strategies based on predictive analytics, optimizing portfolio allocation, and enhancing returns.
- **Decision Support:** Machine learning predictions can serve as decision support tools for traders and investors, aiding in timely buy/sell decisions and asset allocation.
- **Competitive Advantage:** Leveraging advanced analytics can provide a competitive edge in the financial industry by enabling quicker and more informed decision-making.

- **Regulatory Compliance:** Adhering to regulatory standards and guidelines is essential, especially in the financial sector, where transparency and accountability are paramount.

5.2 Conclusion

In this machine learning project focused on stock market price prediction, we explored the application of various predictive models to forecast stock prices based on historical data. Our analysis involved implementing and comparing four types of models: Linear Regression, Random Forest, Decision Tree, and XGBoost Regression. Through this investigation, several key insights and observations have been made. And user can be able to see the next day stock price.

5.3 Recommendations

[1] **Ensemble learning**:- A. Verma and V. Ranga, "Elnids: Ensemble learning based network intrusion detection system for rpl based internet of things," pp. 1–6, 04 2019.

[2] **Machine learning-driven**:- J. Liu, B. Kantarci, and C. Adams, "Machine learning-driven intrusion detection for contiki-ng-based iot networks exposed to nsl-kdd dataset," 07 2020.

[3] **Hybrid intrusion** :-S. Smys, D. Basar, and D. Wang, "Hybrid intrusion detection system for internet of things (iot)," Journal of ISMAC, vol. 2, pp. 190–199, 09 2020.

[4] **IOT networks in machine learning:-** P. Maniriho, E. Niyigaba, Z. Bizimana, V. Twiringiyimana, L. Mahoro, and T. Ahmad, "Anomaly-based intrusion detection approach for iot networks using machine learning," pp. 303–308, 11 2020.

[5] **Machine Learning Classification Techniques:-**H. Alqahtani, I. Sarker, A. Kalim, S. Hossain, S. Ikhlaq, and S. Hossain, Cyber Intrusion Detection Using Machine Learning Classification Techniques, pp. 121–131. 07 2020.

[6] **Network and Service Management :-**M. Injadat, A. Moubayed, A. Nassif, and A. Shami, "Multi-stage optimized machine learning framework for network intrusion detection," IEEE Transactions on Network and Service Management, vol. PP, pp. 1–1, 08 2020.

[7] **Machine learning and knowledge model :-**M. Sarnovsky and J. Paralic, "Hierarchical intrusion detection using machine learning and knowledge model," Symmetry, vol. 12, p. 203, 02 2020.