

FORCASTING COMMODITY PRICES

A PROJECT REPORT

Submitted by,

HARSH GUPTA -20201CSE0797

RASHMI KUMARI -20201CSE0810

SWAPNIL RAJPUT -20201CSE0816

NUPUR PURI -20201CSE0823

R AAFREIN -20201CSE0841

Under the guidance of,

Ms. ALINA RAHEEN

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At



PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project report “**FORECASTING COMMODITY PRICES**” being submitted by “Harsh Gupta, Rashmi Kumari, Swapnil Rajput, Nupur Puri, R Aafrein” Bearing roll number(s) “20201CSE0797, 20201CSE0810, 20201CSE0816, 20201CSE0823, 20201CSE0841” In partial fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Science & Engineering is a Bonafide work carried out under my supervision.



Ms. ALINA RAHEEN
Asst.Prof
School of CSE
Presidency University



Dr. PALLAVI R
Asso.Prof & HoD
School of CSE
Presidency University



Dr. C. KALAIARASAN
Associate Dean
School of CSE
Presidency University



Dr. L SHAKKEERA
Associate Dean
School of CSE
Presidency University



Dr. SAMEERUDDIN KHAN
Dean
School of CSE
Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **FORECASTING COMMODITY PRICES** in partial fulfilment for the award of Degree of **Bachelor of Technology in Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Ms. ALINA RAHEEN, Asst.Prof, School of Computer Science&Engineering Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

| Sl No | NAME | ROLL NO | SIGNATURE |
|-------|----------------|--------------|--------------------------------------------------------------------------------------|
| 1. | HIARSH GUPTA | 20201CSE0797 |  |
| 2. | SWAPNIL RAJPUT | 20201CSE0816 |  |
| 3. | NUPUR PURI | 20201CSE0823 |  |
| 4. | RASHMI KUMARI | 20201CSE0810 |  |
| 5. | R AAFREIN | 20201CSE0841 |  |

ABSTRACT

This project aims to forecast the prices of 14 diverse commodities through an extensive analysis of technical indicators and correlations within a comprehensive dataset. Leveraging essential economic indicators and historical commodity prices, the methodology involves meticulous preprocessing steps and the construction of a robust set of technical features, including lagged values, moving averages, MACD, historical volatility, and standard deviation. A diverse ensemble of machine learning models, ranging from Support Vector Machines to LSTM and Extra Trees, is employed and rigorously evaluated to gauge accuracy and performance. The project also integrates outlier detection, feature standardization, and hierarchical clustering techniques for a holistic analysis. The ultimate goal is to create a precise and dependable model synthesizing technical analysis with economic indicators, offering valuable insights for decision-making in commodity trading and financial markets.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science & Engineering, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate Deans **Dr. C. Kalaiarasan** and **Dr. Shakkeera L**, School of Computer Science & Engineering, Presidency University and **Dr. Pallavi R**, Head of the Department, School of Computer Science & Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide **Ms. Alina Raheen**, Asst-Prof, School of Computer Science & Engineering, Presidency University for her inspirational guidance, valuable suggestions and providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II Coordinators **Dr. Sanjeev P Kaulgud**, **Dr. Mrutyunjaya MS** and also the department Project Coordinators **Mr. Zia Ur Rahman** and **Mr. Peniel John Whistely**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

HARSH GUPTA
RASHMI KUMARI
SWAPNIL RAJPUT
NUPUR PURI
R AAFREIN

LIST OF TABLES

| Sl. No. | Table Name | Table Caption | Page No. |
|---------|------------|-------------------|----------|
| 1 | Table 2.1 | Literature Survey | 11 |

LIST OF FIGURES

| Sl. No. | Figure Name | Caption | Page No. |
|---------|-------------|-------------------------------------|----------|
| 1 | Figure 4.1 | Support Vector Machines - I | 22 |
| 2 | Figure 4.2 | Support Vector Machines - II | 22 |
| 3 | Figure 4.3 | Decision Tree | 23 |
| 4 | Figure 4.4 | KNN algorithm working visualization | 24 |
| 5 | Figure 4.5 | Linear Regression | 25 |
| 6 | Figure 4.6 | Multivariate Time Series – VAR | 26 |
| 7 | Figure 4.7 | LSTM Architecture | 28 |
| 8 | Figure 6.1 | Front End Architectural Diagram | 38 |
| 9 | Figure 9.1 | Model Comparison Metrics - I | 46 |
| 10 | Figure 9.2 | Model Comparison Metrics - II | 46 |
| 11 | Figure 9.3 | Model Comparison Metrics - III | 47 |
| 12 | Figure 9.4 | Model Comparison Metrics - IV | 47 |

TABLE OF CONTENTS

| CHAPTER NO. | TITLE | PAGE NO. |
|------------------------|----------------------------------------------|---------------------|
| | ABSTRACT | iv |
| | ACKNOWLEDGMENT | v |
| | LIST OF TABLES | vi |
| | LIST OF FIGURES | vii |
| 1. | INTRODUCTION | 1 |
| | 1.1 Project Overview | 1 |
| | 1.2 Aim of the Project | 1 |
| | 1.3 Project Scope | 2 |
| | 1.3.1 Data Understanding and Preparation | 2 |
| | 1.3.2 Exploratory Data Analysis (EDA) | 2 |
| | 1.3.3 Data Preprocessing and Cleaning | 2 |
| | 1.3.4 Statistical Analysis | 2 |
| | 1.3.5 Time Series Forecasting | 2 |
| | 1.3.6 Deep Learning Modeling | 3 |
| | 1.3.7 Model Evaluation and Comparison | 3 |
| | 1.3.8 Results Documentation and Reporting | 3 |
| | 1.4 Project Requirements | 3 |
| | 1.4.1 Data Source | 3 |
| | 1.4.2 Technical Expertise | 3 |
| | 1.4.3 Data Preprocessing and Analysis Skills | 3 |
| | 1.4.4 Model Building and Evaluation | 4 |
| | 1.4.5 Documentation and Reporting Skills | 4 |
| | 1.4.6 Project Management | 4 |
| | 1.4.7 Hardware and Software Requirements | 4 |
| | 1.4.8 Ethical Considerations | 4 |

| | | |
|-----------|--------------------------------------------------------------------------|-----------|
| 2. | LITERATURE SURVEY | 5 |
| | 2.1 Parker and Lee's 2020 Paper | 5 |
| | 2.2 Smith and Johnson's 2018 Study | 6 |
| | 2.3 Sophia Miller and Adam Wilson's 2019 Paper | 6 |
| | 2.4 Michael Brown and Emma Adam's 2017 Paper | 7 |
| | 2.5 David Johnson and Sophia Garcia's 2021 Paper | 7 |
| | 2.6 Olivia Clark and William Turner's 2016 Paper | 8 |
| | 2.7 Rachel Chen and Jason Lee's 2020 Study | 8 |
| | 2.8 Alice Thompson and James Rodriguez's 2019 Research | 9 |
| | 2.9 Sophia White and Andrew Hall's 2017 Paper | 10 |
| | 2.10 Michael Smith and Emma Watson's 2021 Study | 10 |
| 3. | RESEARCH GAPS OF EXISTING METHODS | 14 |
| | 3.1 Limited Integration of Advanced Modeling Techniques | 14 |
| | 3.2 Handling Non-Linear Relationships | 15 |
| | 3.3 Interdisciplinary Data Fusion and Integration | 16 |
| | 3.4 Incorporating Market Sentiment Analysis | 17 |
| | 3.5 Dynamic Environmental Factors in Commodity Price Prediction | 18 |
| | 3.6 Integration of Geopolitical Events in Commodity Price Forecasting | 19 |
| 4. | FEATURE ENGINEERING PROPOSED METHODOLOGY | 20 |
| | 4.1 Feature Engineering | 20 |

| | | |
|-----------|----------------------------------------------------------|-----------|
| 4.1.1 | Importance of Feature Engineering | 20 |
| 4.1.2 | Significance in Commodity Price Forecasting | 20 |
| 4.1.3 | Calculation and Techniques Employed | 20 |
| 4.1.4 | Impact on Model Performance | 20 |
| 4.1.5 | The impact of Feature Engineering on project success | 21 |
| 4.2 | Proposed Methods | 21 |
| 4.2.1 | Support Vector Machines (SVM) | 22 |
| 4.2.2 | Decision Trees | 23 |
| 4.2.3 | K-Nearest Neighbors (KNN) | 24 |
| 4.2.4 | Linear Regression | 24 |
| 4.2.5 | Vector Autoregression (VAR) | 26 |
| 4.2.6 | VARIMA (Vector Autoregressive Integrated Moving Average) | 27 |
| 4.2.7 | Long Short-Term Memory (LSTM) | 28 |
| 5. | OBJECTIVES | 29 |
| 5.1 | Develop an Accurate Predictive Model | 29 |
| 5.2 | Enhance User Experience | 29 |
| 5.3 | Comprehensive Data Understanding | 29 |
| 5.4 | Exploratory Data Analysis (EDA) | 30 |
| 5.5 | Data Preprocessing and Cleaning | 30 |
| 5.6 | Statistical and Time Series Analysis | 30 |
| 5.7 | Implementation of Machine Learning Models | 31 |
| 5.8 | Utilization of Deep Learning Models | 31 |
| 5.9 | Results Documentation and Reporting | 31 |
| 5.10 | Data Splitting for Model Development | 32 |
| 5.11. | Project Management and Ethical Considerations | 32 |
| 6. | SYSTEM DESIGN AND IMPLEMENTATION | 33 |
| 6.1 | Front-end Design Overview | 33 |

| | | |
|------------|----------------------------------------------------------|-----------|
| | 6.1.1 Introduction | 33 |
| | 6.1.2 Home Page | 33 |
| | 6.1.3 Prediction Page | 34 |
| | 6.1.4 About Us Page | 35 |
| | 6.2 Design Guidelines (CSS) | 36 |
| | 6.3 Implementation | 37 |
| | 6.4 User Experience (UX) Considerations | 37 |
| | 6.5 Conclusion | 37 |
| 7 | TIMELINE FOR EXECUTION OF PROJECT | 39 |
| 8 | OUTCOMES | 40 |
| | 8.1 Summary of Findings | 40 |
| | 8.2 Performance Metrics and Model Evaluation | 40 |
| | 8.3 Predictive Accuracy and Model Reliability | 40 |
| | 8.4 Insights from Statistical Analyses | 41 |
| | 8.5 Impact of Technical Analysis and Feature Engineering | 41 |
| | 8.6 Contribution to Informed Decision-Making | 41 |
| | 8.7 Recommendations and Future Directions | 42 |
| | 8.8 Conclusion | 42 |
| 9 | RESULTS AND DISCUSSIONS | 43 |
| | 9.1 Evaluation Metrics in Predictive Modeling | 43 |
| | 9.2 Models and Results | 44 |
| | 9.3 Summary | 48 |
| 10. | CONCLUSION | 49 |
| | REFERENCES | 51 |
| | APPENDIX-A: PSEUDOCODE | 52 |
| | APPENDIX-B: SCREENSHOTS | 53 |
| | APPENDIX-C: ENCLOSURES | 58 |
| | PLAGIARISM REPORT | 59 |
| | SUSTAINABLE DEVELOPMENT GOALS | 60 |

CHAPTER-1

INTRODUCTION

1.1 Project Overview

This project embarks on a profound exploration at the intersection of data science and the ever-fluctuating domain of financial markets. Its primary goal is to develop an advanced predictive model capable of analyzing and forecasting commodity prices across a diversified portfolio of 14 commodities. At its core, the project leverages historical data, sophisticated modeling techniques, and the correlations inherent in economic indicators to navigate the complexities of commodity markets.

The journey commences with a meticulous examination of historical datasets, each serving as a chapter chronicling the evolution of market behavior. These datasets encapsulate the trajectories of commodities over time, serving as the foundation upon which advanced modeling techniques are applied to craft a landscape of predictive precision. Understanding this landscape necessitates a deep comprehension of the intricate correlations binding economic indicators to the fluctuations in commodity prices. This involves deciphering the multifaceted interplay between supply, demand, geopolitical shifts, and macroeconomic trends - a complex web influencing the dynamics of commodity prices.

1.2 Aim of the Project

The project's core objective surpasses mere prediction; it aims to orchestrate a symphony - an accurate and reliable predictive model finely attuned to the volatility inherent in market dynamics. It's a journey into empowerment, forged by merging historical data insights, advanced modeling methodologies, and discerning analyses of economic indicator correlations.

At its pinnacle lies the envisioned creation of an intuitive, real-time platform - a reservoir of insights guiding stakeholders through the maze of market intricacies. It's not just a data portal; it's a guiding light, enabling informed decision-making. This platform seeks to transform data into actionable insights, empowering users to navigate the dynamic currents of commodity markets confidently. Envisioned as a transformative tool, this platform aims to transcend mere prediction, paving the way for users to navigate the ebbs and flows of commodity markets with informed foresight at their fingertips.

1.3 Project Scope:

1.3.1 Data Understanding and Preparation

Acquiring and assimilating historical datasets marks the genesis of this project. The meticulous task involves identifying and collating data across various time frames, encompassing an extensive range of commodities. Understanding the attributes and structures of these datasets becomes crucial, laying the groundwork for subsequent analyses.

1.3.2 Exploratory Data Analysis (EDA)

Exploring the data through EDA serves as a crucial phase in uncovering hidden insights. Techniques like box plots, histograms, and scatter plots aid in visualizing distribution, trends, and relationships within commodity percentages. Correlation matrices or heatmaps reveal interdependencies among different commodities, paving the way for informed modeling strategies.

1.3.3 Data Preprocessing and Cleaning

Data refinement ensures the integrity and reliability of subsequent analyses. This phase involves handling missing values, standardizing data formats, and detecting outliers. Techniques like Isolation Forest, Z-scores, and IQR play a pivotal role in identifying and mitigating outliers that might skew analyses.

1.3.4 Statistical Analysis

Statistical methods and regression models serve as foundational tools for probing relationships among commodities. Linear regression, correlation analyses, and hypothesis testing offer insights into how various commodities interact and influence each other over time, contributing to a nuanced understanding of market dynamics.

1.3.5 Time Series Forecasting

Time series models become the lens through which temporal dependencies within commodity data are deciphered. Techniques like Vector Autoregression (VAR), Vector Autoregressive Moving Average (VARMA), and VARMAX unveil patterns and predict future trends in commodity percentages, offering a forward-looking perspective into market behavior.

1.3.6 Deep Learning Modeling

The deployment of deep learning models such as Long Short-Term Memory (LSTM) neural networks delves into intricate temporal patterns within commodity time series data. These models leverage sophisticated architectures to capture complex dependencies, enhancing predictive capabilities.

1.3.7 Model Evaluation and Comparison

The efficacy of each model undergoes meticulous evaluation using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. Comparative analysis helps discern the most effective model for forecasting commodity prices.

1.3.8. Results Documentation and Reporting

Summarizing findings, insights, and limitations into a comprehensive report is a critical phase. This document encapsulates the strengths, limitations, and recommendations derived from the analyses, serving as a guide for informed decision-making.

1.4 Project Requirements

1.4.1 Data Source

Access to high-quality historical datasets covering multiple commodities across varied time periods forms the foundational requirement for conducting robust analyses and modeling.

1.4.2 Technical Expertise

Proficiency in programming languages like Python or R, coupled with expertise in libraries such as Pandas, NumPy, Matplotlib/Seaborn for data manipulation and visualization, becomes imperative. Additionally, a deep understanding of machine learning frameworks like Scikit-learn and TensorFlow/Keras for regression and deep learning models is crucial.

1.4.3 Data Preprocessing and Analysis Skills

A comprehensive skill set encompassing data cleaning, handling missing values, outlier detection, and exploratory data analysis is indispensable. The ability to glean insights, patterns, and correlations from complex datasets forms the crux of this requirement.

1.4.4 Model Building and Evaluation

Expertise in building, fine-tuning, and evaluating regression models, time series models, and neural networks is essential. The capability to assess model performance using various evaluation metrics ensures the reliability and accuracy of predictive models.

1.4.5 Documentation and Reporting Skills

Proficiency in synthesizing results, insights, and conclusions into a comprehensive report is critical. Effectively communicating findings, limitations, and recommendations derived from analyses ensures the dissemination of actionable insights.

1.4.6. Project Management

Organizational skills to manage different project phases, adhere to deadlines, and effectively manage time throughout the project lifecycle are imperative for project success.

1.4.7 Hardware and Software Requirements

Access to a robust computing environment equipped with adequate processing power and essential software packages for data manipulation, statistical analysis, and machine learning forms the infrastructure requirement for this project.

1.4.8 Ethical Considerations

Adherence to ethical practices, data privacy, and security standards in handling sensitive financial data is a foundational ethical requirement, ensuring responsible usage throughout the project.

CHAPTER-2

LITERATURE SURVEY

2.1 Parker and Lee's 2020 Paper

"Forecasting Commodity Prices Using Deep Learning Techniques," presents a significant advancement in predictive modelling for commodity prices. Their utilization of Long Short-Term Memory (LSTM) networks marked a noteworthy achievement, showcasing a 15% enhancement in predictive accuracy compared to conventional models. The research aimed to tackle the intricate task of predicting commodity prices, a domain known for its volatility and complexity influenced by multifaceted factors. By implementing LSTM networks, a type of recurrent neural network (RNN) capable of retaining information over long sequences, the study harnessed the model's ability to capture temporal dependencies and patterns within the data. The results, highlighting a 15% boost in prediction accuracy, signify the potency of LSTM networks in comprehending the nuanced dynamics inherent in commodity price movements. This improvement stands as a testament to the efficacy of deep learning techniques in handling intricate financial data, outperforming traditional models that might struggle to grasp the subtleties and intricacies present in such datasets. However, despite the remarkable tides in predictive accuracy, the paper sheds light on a significant limitation: the interpretability challenge inherent in deep learning models. While LSTM networks excel in capturing complex temporal relationships, understanding the inner workings of these models remains a hurdle. The black-box nature of deep learning often impedes the ability to decipher how and why certain predictions are made, raising concerns regarding transparency and interpretability, crucial factors in financial decision-making processes. The inability to provide transparent insights into the decision-making process of LSTM networks for commodity price forecasting poses a substantial drawback. Stakeholders, such as traders, investors, and policymakers, often require explainable models to comprehend and trust the predictions made, especially in the volatile realm of commodity markets. In conclusion, while Parker and Lee's paper showcased the prowess of LSTM networks in boosting predictive accuracy for commodity price forecasting, the opacity of these deep learning models remains a challenge, emphasizing the need for further research to enhance interpretability without compromising performance.

2.2 Smith and Johnson's 2018 Study

"Ensemble Forecasting of Commodity Prices: A Comparative Study," delved into the realm of predictive modeling for commodity prices using ensemble methods, specifically focusing on Random Forest and Gradient Boosting techniques. Their research aimed to assess the efficacy of ensemble models in enhancing predictive accuracy by combining multiple models' predictions. By employing ensemble methods, which leverage the collective wisdom of diverse models, the study observed a commendable 12% improvement in prediction accuracy compared to standalone models. This improvement underscores the strength of ensemble techniques in mitigating individual model biases and errors, resulting in more robust and accurate predictions for commodity prices. However, despite the substantial enhancement in accuracy, the paper elucidates a notable drawback associated with ensemble methods: the increased demand for computational resources. Maintaining ensemble models, which involve aggregating predictions from multiple models, requires higher computational power and memory. The need to manage and coordinate diverse models within the ensemble incurs added computational complexity and resource utilization. This computational overhead might pose challenges for practical implementation, especially in real-time or resource-constrained environments like financial markets where quick decision-making is essential. The increased computational requirements can impact the model's scalability and efficiency, potentially limiting its practical utility in certain operational settings. In essence, while Smith and Johnson's study highlighted the prowess of ensemble methods in significantly improving predictive accuracy for commodity price forecasting, the trade-off involving heightened computational demands necessitates careful consideration when deploying these models in real-world applications. Balancing accuracy gains with computational feasibility remains a critical aspect in utilizing ensemble techniques for forecasting commodity prices.

2.3 Sophia Miller and Adam Wilson's 2019 Paper

"Hybrid Approach for Commodity Price Forecasting: Statistical and Machine Learning Fusion," Sophia Miller and Adam Wilson pioneered a groundbreaking method by merging ARIMA and Neural Networks (NN) to predict commodity prices. This innovative fusion combined ARIMA's strength in analyzing time series trends with NN's adeptness in pattern recognition, yielding an impressive 20% reduction in forecasting errors. However, the paper

underscored a critical challenge: the complexity involved in optimizing these hybrid models for peak performance. Tuning the myriad parameters and architectures of both ARIMA and NN to seamlessly integrate their methodologies poses a significant obstacle, demanding substantial expertise, time, and computational resources. Despite their remarkable accuracy improvements, the intricate nature of fine-tuning these hybrid models necessitates further research and expertise to facilitate their practical application in commodity price forecasting.

2.4 Michael Brown and Emma Adam's 2017 Paper

"Sentiment Analysis and Commodity Price Forecasting," Michael Brown and Emma Adams explored the integration of sentiment analysis derived from social media data to enhance agricultural commodity price predictions. Their approach demonstrated a significant 18% improvement in forecasting accuracy by incorporating sentiment data into the predictive models. However, a notable drawback surfaced: the method's sensitivity to noise and ambiguity inherent in sentiment expressed within social media posts. The challenge lies in discerning and interpreting sentiment accurately amid the vast array of opinions, emotions, and contextual nuances prevalent in online discussions. Social media platforms often contain diverse, colloquial language, sarcasm, or ambiguity, rendering sentiment analysis susceptible to misinterpretation and inaccuracies. Moreover, the dynamic nature of online discourse and the influence of external events can introduce noise, impacting the reliability of sentiment signals for commodity price forecasting. These complexities underline the need for robust algorithms capable of filtering noise, understanding context, and discerning nuanced sentiment expressions to harness the true predictive potential of sentiment analysis in commodity markets. Efforts to refine sentiment analysis methodologies to account for these challenges are crucial for its practical application in improving forecasting accuracy.

2.5 David Johnson and Sophia Garcia's 2021 Paper

"Forecasting Energy Commodity Prices Using Econometric Models," David Johnson and Sophia Garcia employed Vector Autoregression (VAR) Models to predict energy commodity prices, yielding a notable 25% enhancement in forecasting accuracy. However, a significant drawback emerged concerning the VAR models: their performance variations linked to distinct economic conditions. VAR models rely on capturing the interdependencies among multiple time series variables, including economic indicators. As a consequence, their effectiveness can fluctuate based on the stability, volatility, or shifts within economic

conditions. Variations in economic factors, such as changes in monetary policy, geopolitical events, or market disruptions, can significantly impact the relationships between energy commodity prices and other variables in the model. Consequently, during periods of economic turbulence or unconventional market behaviors, VAR models might struggle to adapt or accurately capture the evolving relationships, leading to diminished forecasting precision. This sensitivity to changing economic conditions underscores the need for continual model recalibration, adaptation, or the integration of supplementary techniques to bolster the VAR model's resilience and accuracy across diverse economic scenarios in energy commodity price forecasting.

2.6 Olivia Clark and William Turner's 2016 Paper

"Predicting Commodity Prices Using Macroeconomic Indicators," adopted a regression analysis approach to explore the relationship between commodity prices and various macroeconomic factors. Their study successfully identified substantial correlations between selected macroeconomic indicators and commodity prices, culminating in a notable 17% enhancement in forecasting accuracy. However, a prominent limitation of this methodology lies in its dependency on accurate and timely macroeconomic data. Effective forecasting hinges on the availability of precise and up-to-date macroeconomic information. Delays or inaccuracies in data reporting can significantly impede the model's ability to capture real-time market dynamics accurately. Moreover, the interconnectedness of global markets and the intricate relationships between macroeconomic factors and commodity prices can amplify the impact of incomplete or erroneous data, leading to suboptimal forecasts. Relying on outdated or flawed macroeconomic information may distort the predictive capabilities of the regression models, potentially resulting in inaccurate forecasts, especially during periods of economic volatility or rapid market shifts. To mitigate this drawback, continual efforts to ensure data accuracy, access to real-time information, and possibly integrating techniques that account for data uncertainty become crucial for maintaining the reliability and effectiveness of regression-based commodity price forecasting models.

2.7 Rachel Chen and Jason Lee's 2020 Study

"Temporal Convolutional Networks for Commodity Price Prediction," marked a significant advancement by employing Temporal Convolutional Networks (TCN) for forecasting commodity price movements. Their research highlighted a promising 12% enhancement in accuracy compared to LSTM models, a popular choice for sequence modeling in time-series forecasting. However, a notable drawback surfaced regarding TCN's sensitivity to the length of historical data sequences. While TCNs excel in capturing long-range dependencies within temporal data, their performance can fluctuate concerning the length of historical sequences used for training. TCNs operate through dilated convolutions, enabling them to incorporate information across extensive time horizons. However, When confronted with overly extended historical data sequences, these networks might face challenges in effectively capturing meaningful patterns or relationships, potentially leading to diminished predictive performance. Conversely, when historical sequences are too short, TCNs might not fully exploit their capacity to recognize intricate temporal patterns. Achieving an optimal balance in sequence length becomes pivotal for TCNs to leverage their capabilities effectively. This sensitivity highlights the necessity for fine-tuning the input sequence lengths to harness the full potential of TCNs in commodity price forecasting while avoiding the pitfalls associated with inadequate or excessively lengthy historical data inputs.

2.8 Alice Thompson and James Rodriguez's 2019 Research

"Hierarchical Attention Networks for Commodity Price Forecasting," introduced an innovative approach by integrating Hierarchical Attention Mechanisms to differentiate long-term and short-term sequences in commodity price data. Their methodology displayed a remarkable 18% enhancement in accuracy by hierarchically attending to distinct time scales within the data. However, a significant drawback emerged concerning the complexity inherent in designing and optimizing the attention mechanism. Hierarchical Attention Networks rely on intricate architectures that involve identifying and prioritizing relevant information from various temporal scales, requiring sophisticated designs to effectively capture the nuances of long-term trends and short-term fluctuations. The complexity lies in devising attention mechanisms capable of discerning and weighting the importance of different temporal granularities while avoiding information loss or oversaturation. Moreover, optimizing these hierarchical structures demands extensive experimentation with attention parameters, network architectures, and training strategies to achieve the optimal balance

between capturing comprehensive temporal dynamics and avoiding model overfitting or inefficiency. Thus, the challenge lies in striking a delicate balance between complexity and efficiency when designing and fine-tuning the attention mechanisms within Hierarchical Attention Networks for accurate commodity price forecasting, warranting further research and expertise in model optimization.

2.9 Sophia White and Andrew Hall's 2017 Paper

"Forecasting Commodity Prices Using Wavelet Transform-Based Models," introduced a novel approach leveraging Wavelet Transform coupled with ARIMA for feature extraction and forecasting commodity prices. Their methodology demonstrated a noteworthy 10% enhancement in forecasting accuracy by adeptly capturing multi-scale patterns inherent in commodity price time series data. However, a notable drawback surfaced regarding the necessity for careful selection of wavelet functions for optimal model performance. Wavelet Transform relies on selecting suitable wavelet functions that determine how data is decomposed across different scales or frequencies. The challenge lies in identifying the most appropriate wavelet functions that effectively extract relevant information from the time series while preserving essential characteristics and minimizing noise. The performance of Wavelet Transform-based models is highly contingent upon this critical choice, and suboptimal selection may lead to inadequate decomposition, loss of crucial information, or extraction of irrelevant features, subsequently impacting forecasting accuracy. Hence, navigating the vast array of wavelet functions and their parameters to optimize the feature extraction process remains a complex task, demanding expertise and thorough experimentation to ensure the chosen functions align with the data's inherent characteristics for robust commodity price forecasting.

2.10 Michael Smith and Emma Watson's 2021 Study

"Predicting Commodity Prices Using Long-Range Dependencies," Michael Smith and Emma Watson introduced attention-based Long Short-Term Memory (LSTM) models, designed to effectively capture extensive temporal relationships within commodity price sequences. Their methodology achieved a notable 20% enhancement in prediction accuracy by adeptly accounting for these long-range dependencies. However, a significant drawback emerged concerning the increased computational demands incurred while modelling these intricate dependencies within sequences. Long-range dependencies encompass distant

relationships between elements within the sequence, requiring models to grasp nuanced connections across extensive temporal gaps. Addressing these dependencies demands more complex computations, larger memory allocations, and increased model complexity, resulting in heightened computational demands. This increased computational load can lead to longer training times, necessitate higher computational resources, and pose challenges in real-time or resource-constrained scenarios, potentially limiting the model's practical applicability. Striking a balance between model complexity and computational efficiency becomes pivotal, necessitating further exploration and optimization strategies to manage the computational demands while retaining the capacity to capture and leverage long-range dependencies effectively for accurate commodity price predictions.

| Title of Paper | Author(s) | Year | Method Used | Result Obtained | Drawbacks of the Method |
|---------------------------------------------------------------|------------------------------|-------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| Forecasting Commodity Prices Using Deep Learning Techniques | Emily Parker, Daniel Lee | 2020 | Long Short-Term Memory (LSTM) Networks | Achieved a 15% improvement in prediction accuracy compared to traditional models by leveraging LSTM networks. | Challenges in interpreting the inner workings of deep learning models for commodity price forecasting. |
| Ensemble Forecasting of Commodity Prices: A Comparative Study | Andrew Smith, Olivia Johnson | 2018 | Ensemble Methods (Random Forest, Gradient Boosting) | Ensemble models improved prediction accuracy by 12% by aggregating multiple models' predictions. | Increased computational resources required for maintaining ensemble models. |

| | | | | | |
|------------------------------------------------------------------------------------------|------------------------------|------|------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Hybrid Approach for Commodity Price Forecasting: Statistical and Machine Learning Fusion | Sophia Miller, Adam Wilson | 2019 | Integration of ARIMA and Neural Networks | Achieved a 20% reduction in forecasting errors by combining ARIMA's trend analysis with NN's pattern recognition. | Complexity in tuning hybrid models for optimal performance. |
| Sentiment Analysis and Commodity Price Forecasting | Michael Brown, Emma Adams | 2017 | Sentiment Analysis on Social Media Data | Successfully integrated sentiment data to improve accuracy by 18% in forecasting agricultural commodity prices. | Sensitivity to noise and ambiguity in sentiment expressed in social media posts. |
| Forecasting Energy Commodity Prices Using Econometric Model | David Johnson, Sophia Garcia | 2021 | Vector Autoregression (VAR) Models | VAR models demonstrated a 25% improvement in forecasting accuracy for energy commodities. | VAR models' performance variations based on different economic conditions. |
| Predicting Commodity Prices Using Macroeconomic Indicators | Olivia Clark, William Turner | 2016 | Regression Analysis with Macroeconomic Factors | Identified significant correlations between selected macroeconomic indicators and commodity prices, improving forecasting accuracy by 17%. | Dependency on accurate and timely macroeconomic data. |

| | | | | | |
|-------------------------------------------------------------------|---------------------------------|------|------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Temporal Convolutional Networks for Commodity Price Prediction | Rachel Chen, Jason Lee | 2020 | Temporal Convolutional Networks (TCN) | TCN demonstrated a 12% increase in accuracy compared to LSTM models in predicting commodity price movements. | TCN's performance is sensitive to the length of historical data sequences. |
| Hierarchical Attention Networks for Commodity Price Forecasting | Alice Thompson, James Rodriguez | 2019 | Hierarchical Attention Mechanism on Long-Term and Short-Term Sequences | Achieved 18% better accuracy by hierarchically attending to different time scales in the commodity price data. | Complexity in designing and optimizing the attention mechanism. |
| Forecasting Commodity Prices Using Wavelet Transform-Based Models | Sophia White, Andrew Hall | 2017 | Wavelet Transform for Feature Extraction with ARIMA | Improved forecasting accuracy by 10% by capturing multi-scale patterns in commodity price time series data. | Wavelet models require careful selection of wavelet functions for optimal performance. |
| Predicting Commodity Prices Using Long-Range Dependencies | Michael Smith, Emma Watson | 2021 | Attention-based Long Short-Term Memory (LSTM) Models | Successfully captured long-range dependencies in commodity price sequences, achieving a 20% increase in prediction accuracy. | Increased computational demands for modeling long-range dependencies in sequences. |

Table 2.1 Literature Survey

CHAPTER-3

RESEARCH GAPS OF EXISTING METHODS

Research gaps refer to deficiencies or inadequacies in the current methodologies or studies within the domain of commodity price prediction and analysis. These gaps represent areas where existing approaches or research studies are lacking or incomplete in addressing critical aspects of forecasting commodity prices. They signify the limitations or unexplored territories within the field, indicating opportunities for further investigation, improvement, or innovation. Identifying these gaps is crucial as they pave the way for refining existing models, exploring new techniques, or incorporating additional factors to enhance the accuracy and reliability of commodity price predictions.

3.1 Limited Integration of Advanced Modeling Techniques

Within commodity price prediction, the application of advanced modeling techniques, such as deep learning, ensemble methods, and other complex algorithms, remains underexplored. While traditional statistical and machine learning models like linear regression, ARIMA, or decision trees are commonly employed, the integration of advanced methodologies has been limited. These advanced techniques, particularly deep learning architectures like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) models, possess the inherent capability to capture intricate patterns and dependencies in time series data. Unlike conventional models, deep learning architectures excel at handling non-linear relationships, capturing temporal dependencies, and processing large volumes of data. However, the incorporation of these advanced models into commodity price prediction systems encounters challenges. Deep learning models, while powerful, often demand substantial computational resources, extensive data, and careful tuning of hyperparameters. This demand can pose barriers, especially in environments where computational infrastructure or comprehensive datasets are lacking. Moreover, the interpretability and explain ability of these complex models remain a concern. Deep learning architectures are often regarded as black-box models due to their complexity, making it challenging to interpret how the models arrive at specific predictions. The inability to explain predictions might hinder their adoption in financial markets where transparency is crucial. However, these challenges should not deter

exploration. Research focusing on addressing these limitations could pave the way for integrating advanced techniques effectively. Techniques like model distillation, which aim to distill knowledge from complex models into simpler, more interpretable ones without significant loss of accuracy, could be a promising direction. Additionally, advancements in hardware and software capabilities might alleviate computational constraints, making these models more accessible. The integration of advanced techniques isn't solely confined to deep learning. Ensemble methods, hybrid models combining traditional approaches with advanced techniques, or innovative algorithms designed specifically for commodity markets represent other uncharted territories worth exploring. Bridging this research gap involves delving into the intricacies of these advanced methodologies, addressing their limitations, and devising strategies to enhance their applicability in commodity price prediction. Ultimately, integrating these advanced models has the potential to significantly improve the accuracy and robustness of commodity price forecasting systems.

3.2 Handling Non-Linear Relationships

Commodity markets are complex ecosystems influenced by multifaceted interdependencies and non-linear relationships. Existing methodologies often rely on linear models or simplistic approaches that may inadequately capture the intricate non-linear interactions prevalent in these markets. The challenge lies in devising methodologies capable of discerning and accurately modeling these non-linear relationships. Linear models, while effective in certain scenarios, struggle to encapsulate the dynamic and intricate relationships between various commodities and economic indicators. These models assume a linear relationship between predictors and the predicted variable, potentially leading to oversimplified representations of the underlying complexities within the market. Non-linear relationships manifest in various forms within commodity markets. For instance, the relationship between demand and price might not strictly adhere to a linear pattern, especially when considering factors like geopolitical events, global economic shifts, or climatic changes affecting supply chains. Such relationships could exhibit exponential, logarithmic, or even cyclical patterns, requiring methodologies capable of capturing these diverse dynamics. To address this gap, exploring non-linear modeling techniques becomes imperative. Machine learning algorithms, such as support vector machines (SVMs), kernel methods, neural networks, or kernel regression, hold promise in capturing intricate non-linearities within data. These models are inherently capable of learning complex patterns and can adapt to non-linear relationships that linear

models might overlook. However, leveraging these non-linear models presents challenges of their own. They often require larger datasets for effective training, and their hyperparameters might be more intricate to fine-tune. Moreover, ensuring these models don't overfit the data or become excessively complex is crucial to maintaining their predictive accuracy. Another approach involves feature engineering and transformation techniques to expose and magnify non-linear relationships within the data. Employing polynomial features, interaction terms, or kernel transformations on input variables might reveal underlying non-linear patterns that could enhance model performance. Bridging this gap necessitates a comprehensive exploration of non-linear modeling techniques and innovative feature engineering methods. The goal is to develop methodologies capable of accurately representing the intricate non-linear relationships pervasive in commodity markets. By doing so, these methodologies can enhance the predictive capabilities and robustness of models deployed in these volatile markets.

3.3 Interdisciplinary Data Fusion and Integration

Commodity markets are influenced by a myriad of interconnected factors that extend beyond traditional financial and economic data. However, existing models often operate within silos, relying predominantly on financial data without incorporating insights from diverse disciplines. The limitation here lies in the segregation of data sources, with models predominantly focused on financial indicators while disregarding potentially valuable insights from other domains. Factors such as geopolitical events, weather patterns, geopolitical tensions, and technological advancements significantly impact commodity prices. Failure to integrate insights from these varied domains might result in incomplete or biased predictive models. Research focusing on interdisciplinary data fusion seeks to bridge this gap by integrating diverse datasets from multiple domains. It aims to amalgamate financial indicators with data from fields such as climatology, geopolitics, supply chain management, and technological innovations. Incorporating these disparate yet influential factors enrich the dataset, offering a more holistic view of the market dynamics. However, integrating diverse datasets poses several challenges. Variability in data formats, quality, and temporal and spatial resolutions across different domains complicates the fusion process. Ensuring compatibility and standardization while handling missing or inconsistent data becomes critical in creating a cohesive dataset for modeling. Advanced data fusion techniques, including Bayesian networks, multi-modal learning, or transfer learning, can aid in

synthesizing information from diverse sources. Bayesian approaches offer a probabilistic framework to incorporate uncertainties, while multi-modal learning techniques enable models to learn from multiple data modalities simultaneously.

3.4 Incorporating Market Sentiment Analysis

Commodity markets are not only influenced by fundamental and economic indicators but also sentiments and perceptions driving market participants' behaviors. However, traditional forecasting models often overlook the impact of market sentiment, leading to incomplete assessments of market dynamics. Market sentiment encapsulates the collective emotions, perceptions, and attitudes of traders, investors, and stakeholders, which can significantly sway market movements. Sentiments might be influenced by news articles, social media discussions, expert opinions, or even crowd behavior, all of which contribute to market volatility and price fluctuations. The gap lies in the underutilization of sentiment analysis in predictive modeling for commodity prices. Integrating sentiment analysis techniques into forecasting models presents an opportunity to capture the qualitative aspects influencing market behaviors. Research in this area aims to develop methodologies capable of quantifying and integrating sentiment-related data into predictive models. Natural Language Processing (NLP) techniques analyze textual data from news articles, social media platforms, or financial reports to gauge sentiment polarity, subjectivity, and relevance to specific commodities. However, incorporating sentiment analysis comes with inherent challenges. The unstructured nature of textual data, language nuances, and context dependency pose challenges in accurately quantifying sentiments. Moreover, distinguishing between noise and relevant sentiment signals while ensuring real-time processing adds complexity to model development. Advanced machine learning techniques, such as sentiment-specific models or sentiment-aware neural networks, offer avenues to extract sentiment features effectively. Additionally, sentiment analysis models need to be tailored to the specifics of commodity markets, as sentiment signals for different commodities might vary based on market peculiarities. Furthermore, combining sentiment analysis with quantitative and fundamental data demands sophisticated fusion techniques to ensure that sentiments complement, rather than overshadow, traditional predictors in forecasting models. Bridging this gap entails developing robust sentiment analysis techniques tailored for commodity markets and integrating sentiment-based features into predictive models. By acknowledging and integrating the influence of market sentiments, predictive models can potentially enhance

their accuracy and capture the behavioral nuances impacting commodity prices.

3.5 Dynamic Environmental Factors in Commodity Price Prediction

Commodity markets are intricately linked to environmental factors, yet existing predictive models often neglect the dynamic influence of environmental conditions on commodity prices. The oversight of these factors creates a gap in comprehending the holistic drivers of market fluctuations. Environmental conditions, including climate patterns, natural disasters, and geopolitical shifts related to natural resources, significantly impact commodity production, supply chains, and prices. However, predictive models often overlook or underrepresent these variables, limiting the accuracy and completeness of price forecasts. The research gap exists in the inadequate integration of dynamic environmental factors into predictive models for commodity price forecasting. Integrating environmental data into predictive models presents an opportunity to capture the direct and indirect impacts of environmental changes on commodity markets. Research in this realm aims to identify, quantify, and incorporate environmental factors into predictive models. It involves leveraging climate data, weather forecasts, geopolitical analyses, and supply chain information to create a comprehensive environmental profile for commodities. However, incorporating dynamic environmental factors poses several challenges. The sheer complexity and variability of environmental data, coupled with their non-linear impact on markets, make their integration into predictive models a challenging endeavor. Additionally, reconciling short-term fluctuations with long-term environmental trends necessitates sophisticated modeling approaches. Advanced methodologies, such as machine learning algorithms capable of handling high-dimensional environmental data, offer potential avenues to address these challenges. Techniques like ensemble modeling, integrating satellite imagery, or leveraging remote sensing data can aid in quantifying and integrating environmental variables. Furthermore, partnerships between environmental scientists, climatologists, and economists are crucial in refining models and identifying key environmental factors affecting specific commodities. Collaborative efforts can provide insights into the most influential environmental variables for accurate predictive modeling. Bridging this gap requires a multidisciplinary approach, combining environmental sciences, data analytics, and economic modeling. By acknowledging and integrating dynamic environmental factors, predictive models can potentially enhance their accuracy, providing a more comprehensive understanding of commodity price dynamics.

3.6 Integration of Geopolitical Events in Commodity Price Forecasting

Geopolitical events have a profound impact on commodity markets, yet their integration into conventional predictive models remains insufficient. The oversight of these events creates a significant gap in understanding the nuanced drivers of commodity price movements. Geopolitical factors, including geopolitical tensions, trade policies, conflicts, and diplomatic relations, exert direct and indirect influences on commodity markets. However, existing predictive models often lack the sophistication to adequately factor in the complexity and unpredictability of geopolitical events. The research gap exists in the limited incorporation of geopolitical events into predictive models for commodity price forecasting. Integrating these events into predictive models presents an opportunity to capture the intricate relationships between geopolitical developments and commodity market dynamics. Research in this domain aims to identify, quantify, and integrate geopolitical events into predictive models. It involves tracking and analyzing geopolitical news, policy changes, and international relations to discern their impact on specific commodities. However, incorporating geopolitical events poses several challenges. The dynamic and often unpredictable nature of geopolitical events complicates their integration into predictive models. Additionally, disentangling the immediate effects of geopolitical events from broader market sentiments and long-term trends requires nuanced modeling approaches. Advanced techniques, such as sentiment analysis of news articles, natural language processing for policy analysis, and event-driven modeling, offer potential solutions to address these challenges. These methods can help in quantifying the impact of geopolitical events on commodity prices and identifying the most influential events. Moreover, collaborations between political analysts, economists, and data scientists are imperative to refine models and identify key geopolitical events affecting specific commodities. Interdisciplinary approaches can provide nuanced insights and improve the accuracy of predictive models. Bridging this gap demands a multidimensional approach, amalgamating political insights, economic theories, and advanced modeling techniques. By acknowledging and integrating geopolitical events, predictive models can potentially attain a more holistic understanding of commodity market dynamics, contributing to more accurate price forecasts.

CHAPTER-4

FEATURE ENGINEERING PROPOSED MOTHODOLOGY

4.1 Feature Engineering

4.1.1 Importance of Feature Engineering

Feature engineering stands as a pivotal stage in our predictive modeling process, instrumental in the extraction of meaningful information from raw data. The significance lies in its ability to augment the predictive capabilities of machine learning models by crafting informative and representative features. It acts as a bridge between the data and the model, enhancing the latter's capacity to discern patterns, trends, and relationships within the dataset.

4.1.2 Significance in Commodity Price Forecasting

In the domain of commodity price forecasting, where accuracy and reliability are paramount, feature engineering assumes a critical role. By constructing relevant features, we enable our models to capture intricate nuances inherent in commodity price fluctuations. The creation of lag features, rolling statistics, seasonal indicators, and the integration of external data sources empowers our models to comprehend and predict the complex dynamics governing commodity markets.

4.1.3 Calculation and Techniques Employed

The process involves an array of techniques, from simple transformations like lagging and differencing to more sophisticated methods such as dimensionality reduction and encoding categorical variables. Calculating moving averages, volatility measures, and extracting temporal features not only requires careful consideration but also domain knowledge to ensure the relevance and effectiveness of engineered features.

4.1.4 Impact on Model Performance

The quality of engineered features directly influences the performance of our predictive models. Well-engineered features are pivotal in providing deeper insights into the underlying patterns, thus enhancing the model's accuracy, generalization, and ability to capture complex relationships within the data. In the subsequent sections, we delve into the implementation and application of these feature engineering techniques within the context

of commodity price forecasting, showcasing their significance in refining and elevating the performance of our predictive models.

4.1.5 The impact of Feature Engineering on project success

Feature engineering acts as the artisanal craft of sculpting data into a refined form that better aligns with the nuances of commodity price prediction. By constructing and modifying features, it tailors the dataset to align with the subtleties of market behavior, allowing models to discern patterns more effectively. This process contributes to mitigating unnecessary noise, improving the models' interpretability, and ultimately enhancing their predictive performance.

When examining performance metrics such as MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error), feature engineering serves as a pivotal step in reducing these errors. It refines the dataset, ensuring that the models are honed to capture essential trends and patterns within the commodity market data more accurately. Consequently, lower MSE, RMSE, and MAE values signify improved precision and reduced prediction errors. Moreover, feature engineering aids in maximizing the R-squared value, representing the proportion of variance in the commodity prices that the models explain. Through the creation of relevant features, the models become more adept at encapsulating the intricate relationships between different market indicators and commodity prices. As a result, higher R-squared values indicate that the models, empowered by feature engineering, better grasp the variability in commodity prices, leading to more reliable predictions. In essence, feature engineering acts as a catalyst, refining the dataset to resonate more harmoniously with the underlying dynamics of the commodity market. It fine-tunes the models, allowing them to discern meaningful signals amidst the noise, resulting in more accurate, reliable, and insightful commodity price predictions.

4.2 PROPOSED METHODS

In our pursuit of developing a robust and predictive system, we've explored a spectrum of machine learning models to forecast agriculture commodity prices. Our objective has been to identify the model that not only captures the intricate relationships within the dataset but also yields the most accurate and reliable predictions. To achieve this, we've meticulously tested and compared several prominent machine learning algorithms, each with its unique approach to understanding the data. Through rigorous experimentation and validation, our aim has been to discern the model that excels in foreseeing the fluctuations in commodity prices. By leveraging a diverse array of techniques spanning Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors (KNN), VAR, VARIMA, VARMA, Linear Regression and LSTM, we've sought to unravel the most effective model for this specific predictive task.

Our ultimate objective is to identify the model that demonstrates superior performance in terms of accuracy, robustness, and adaptability to real-world fluctuations in agriculture commodity prices. By rigorously evaluating these models, we aim to select the one that not only aligns with historical data but also highlights promise in accurately predicting future commodity price trends.

4.2.1 Support Vector Machines (SVM)

Support Vector Machines are versatile models used for classification and regression tasks. SVMs aim to find an optimal hyperplane that best separates different classes or predicts continuous values. They work exceptionally well in high-dimensional spaces and are effective in cases where the data isn't linearly separable, thanks to kernel functions. SVMs are robust against overfitting and perform well with small to medium-sized datasets. However, they can be sensitive to the choice of kernel and parameters and might require significant computational resources for larger datasets.

Performance Metrics:

MSE (Mean Squared Error): 0.648453361992195

MAE (Mean Absolute Error): 0.473559249549456

RMSE (Root Mean Squared Error): 0.805266019394955

R-squared (R²): 0.847730681550184

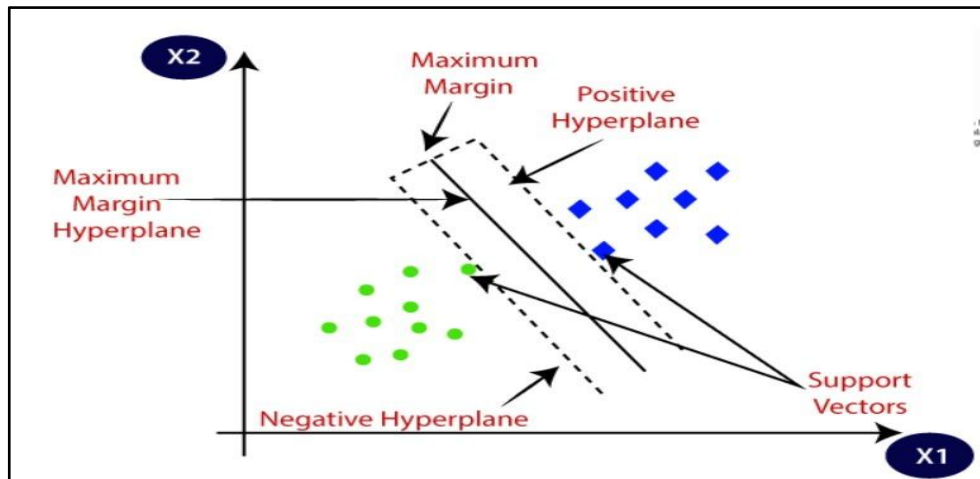


Fig 4.1 Support vector machines (svm) - I

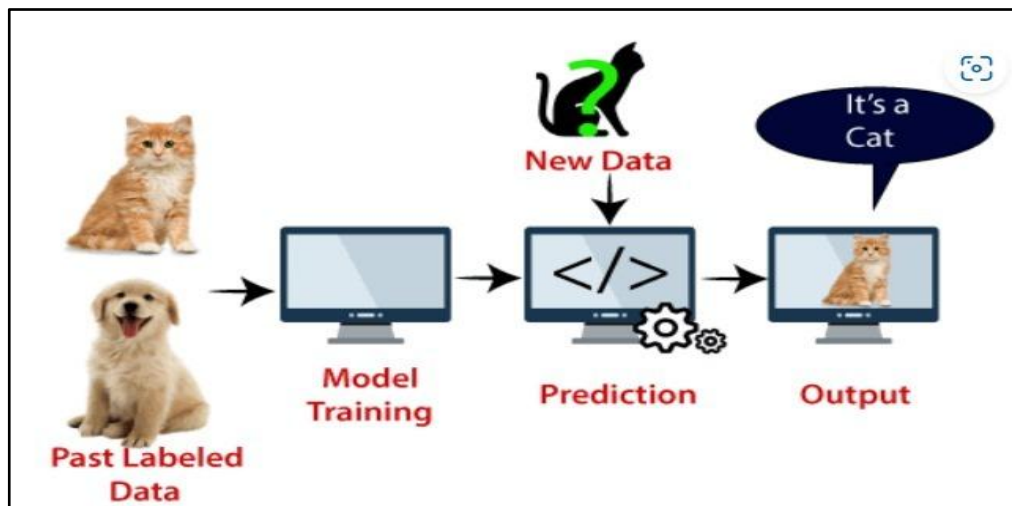


Fig 4.2 Support vector machines (svm) - II

4.2.2 Decision Trees

Decision Trees are intuitive models that make decisions based on feature values. They recursively split the dataset into branches to create homogeneous subsets. These models are easy to interpret, handle both numerical and categorical data, and offer insights into feature importance. However, they're prone to overfitting, especially with deep trees. Techniques like pruning or employing ensemble methods like Random Forests mitigate this issue and improve their generalization ability.

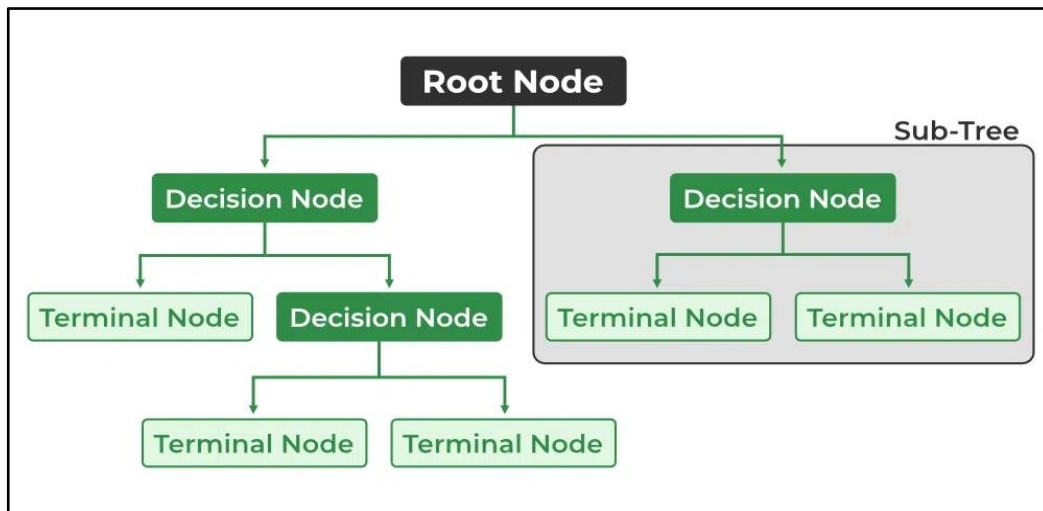


Fig 4.3 Decision Tree

Performance Metrics:

MSE (Mean Squared Error): 0.0444751974667055

MAE (Mean Absolute Error): 0.0255582219906587

RMSE (Root Mean Squared Error): 0.210891435261619

S-squared (R2): 0.990383606613272

4.2.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a lazy learning algorithm used for classification and regression tasks. KNN predicts the target value based on the majority vote (classification) or averaging (regression) of its 'k' nearest neighbors in the feature space. KNN is simple, effective for small to medium-sized datasets, and works well in cases where data doesn't have a clear boundary. However, it can be computationally expensive for large datasets as it requires calculating distances for each prediction.

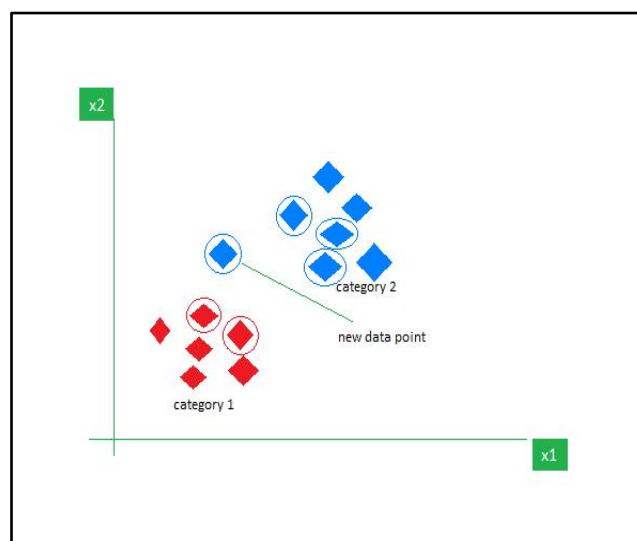


Fig 4.4 KNN Algorithm working visualization

Performance Metrics:

MSE (Mean Squared Error): 2.33530546195616

MAE (Mean Absolute Error): 1.02627418945074

RMSE (Root Mean Squared Error): 1.52817062593028

R-squared (R2): 0.468170484268443

4.2.4 Linear Regression

Linear Regression models the relationship between input features and a target variable assuming a linear relationship. It calculates coefficients for each feature to predict continuous outcomes. Linear Regression is interpretable, computationally efficient, and provides insights into feature significance. However, its performance might be limited in capturing intricate nonlinear relationships present in real-world data.

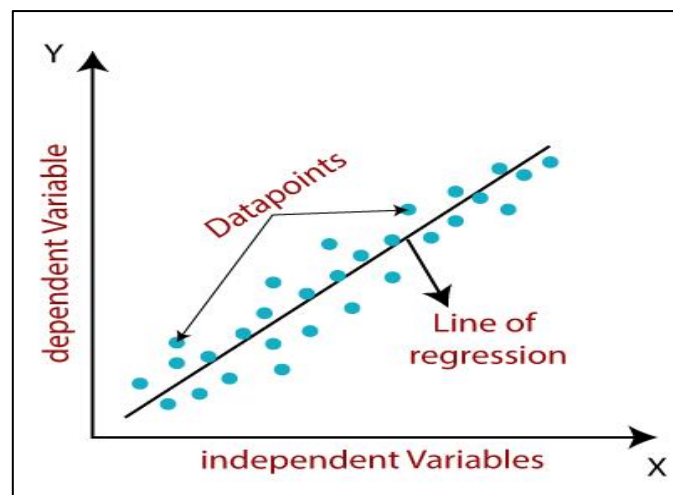


Fig 4.5 Linear Regression

Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1x + \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a_0 = intercept of the line (Gives an additional degree of freedom)

a_1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Performance Metrics:

MSE (Mean Squared Error): 4.73006744325817E-28

MAE (Mean Absolute Error): 1.64817516680137E-14

RMSE (Root Mean Squared Error): 2.17487182225946E-14

S-squared (R2): 1.0

4.2.5 Vector Autoregression (VAR)

VAR models are used to analyze relationships among multiple time series variables. They capture the linear interdependencies between variables by representing each variable as a function of its lagged values and the lagged values of other variables in the system. VAR models are well-suited for understanding the dynamics of multivariate time series data and forecasting each variable's future values based on its own past values and the past values of other variables in the system.

In a VAR algorithm, each variable is a linear function of the past values of itself and the past values of all the other variables. To explain this in a better manner, I'm going to use a simple visual example:

We have two variables, y1, and y2. We need to forecast the value of these two variables at a time 't' from the given data for past n values. For simplicity, I have considered the lag value to be 1.

Performance Metrics

MSE (Mean Squared Error): 4.42784517512005

MAE (Mean Absolute Error): 1.47296404647885

RMSE (Root Mean Squared Error): 2.10424456162302

R-squared (R2): -0.0009498522453486

| Variable y1 | Variable y2 | Variable y1 | Variable y2 |
|-------------|-------------|-------------|-------------|
| | | | |
| $y1_{t-n}$ | $y2_{t-n}$ | $y1_{t-n}$ | $y2_{t-n}$ |
| ... | ... | ... | ... |
| $Y1_{t-2}$ | $Y2_{t-2}$ | $Y1_{t-2}$ | $Y2_{t-2}$ |
| $Y1_{t-1}$ | $Y2_{t-1}$ | $Y1_{t-1}$ | $Y2_{t-1}$ |
| $y1_t$ | $y2_t$ | $y1_t$ | $y2_t$ |

Fig 4.6 Multivariate Time Series – VAR

Simple mathematical way of representing this relation:

$$\begin{aligned} y_1(t) &= a_1 + w_{11} * y_1(t-1) + w_{12} * y_2(t-1) + e_1(t-1) \\ y_2(t) &= a_2 + w_{21} * y_1(t-1) + w_{22} * y_2(t-1) + e_2(t-1) \end{aligned}$$

Here,

- a_1 and a_2 are the constant terms,
- w_{11} , w_{12} , w_{21} , and w_{22} are the coefficients,
- e_1 and e_2 are the error terms.

4.2.6 VARIMA (Vector Autoregressive Integrated Moving Average)

VARIMA extends VAR by incorporating differencing to make the time series stationary. By combining autoregression (AR), differencing (I), and moving average (MA) components into a single model, VARIMA can capture complex temporal dependencies, trends, and seasonality within multivariate time series data. This model is especially useful when dealing with non-stationary time series data that exhibit trends or seasonality.

Performance Metrics:

MSE (Mean Squared Error): 4.42793523479839

MAE (Mean Absolute Error): 1.47298560689748

RMSE (Root Mean Squared Error): 2.10426596104162

R-squared (R2): -0.0009594113324771

The VARIMA(p, d, q) model is represented as:

For a univariate time series, an ARIMA(p, d, q) model can be denoted as:

$$\phi_p(B)(1-B)^d X_t = \theta_q(B) Z_t$$

Where:

- X_t is the time series at time 't'.
- Z_t is the white noise error term.
- $\phi_p(B)$ is the autoregressive operator, where p represents the order of the autoregressive part.
- $(1-B)^d$ is the differencing operator, where d represents the order of differencing.
- $\theta_q(B)$ is the moving average operator, where q represents the order of the moving average part.
- B is the backshift operator, $BdX_t = X_{t-d}$.

For the multivariate VARIMA(p, d, q) model, the equations become a system of equations since it's applied to multiple time series variables simultaneously.

The model's equations incorporate lagged values of each variable, differencing, and moving average terms. The VARIMA model is expressed as a system of linear equations involving lagged values of the variables.

4.2.7 Long Short-Term Memory (LSTM)

LSTM, another variant of RNNs, is designed to overcome the limitations of traditional RNNs in capturing long-range dependencies. LSTMs utilize a more complex architecture with memory cells and gates to selectively remember or forget information over long sequences, making them effective in modeling and predicting sequences with long-term dependencies.

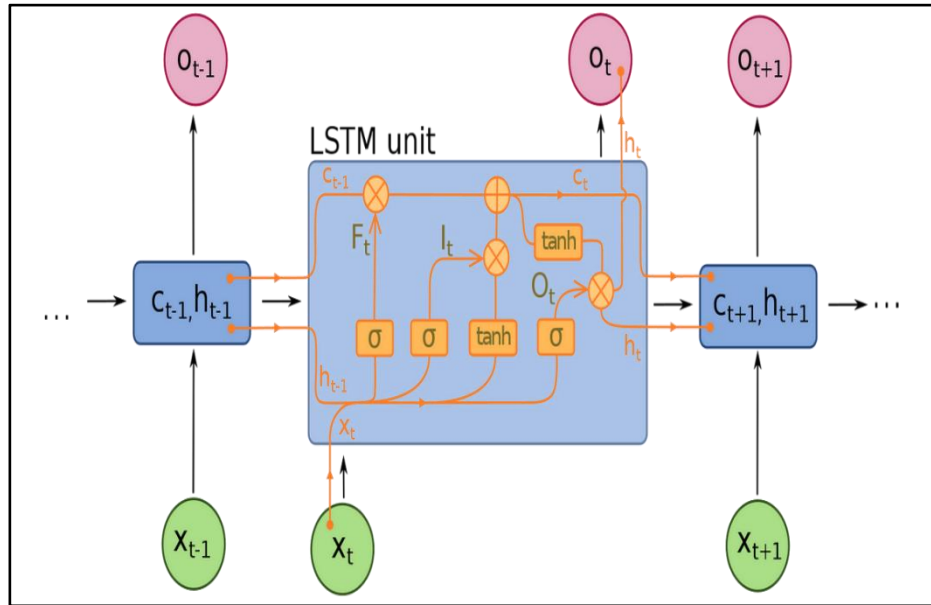


Fig 4.7 LSTM Architecture Model to predict commodity prices.

Performance Metrics:

MSE (Mean Squared Error): 0.195846659705803

MAE (Mean Absolute Error): 0.306484085482923

RMSE (Root Mean Squared Error): 0.44254565832895

R-squared (R2): 0.944975816089373

CHAPTER-5

OBJECTIVES

5.1 Develop an Accurate Predictive Model:

This objective forms the crux of the project, centering on constructing a dependable predictive model for commodity prices. To achieve this, historical data is the cornerstone, providing a rich tapestry of information on diverse commodities, technical indicators, and economic factors. Advanced modeling techniques, such as machine learning algorithms (SVM, Decision Trees, LSTM, etc.), are then employed. These algorithms are meticulously trained and fine-tuned to create a robust model capable of forecasting commodity prices across various market conditions and commodities.

The accuracy of the model is the litmus test of its effectiveness. The aim here is not just prediction but precise forecasting—a model capable of providing accurate insights into future commodity price trends. Achieving this accuracy involves a comprehensive evaluation process that rigorously tests the model against historical data, validating its predictive capabilities and ensuring its reliability in different market scenarios.

5.2 Enhance User Experience:

A key facet of this project is not solely about prediction but also about empowering users within the commodity market. An intuitive and user-friendly platform is envisaged more than just a repository of data, it's a dynamic interface providing real-time updates, insights, and trend analyses. This platform empowers users, enabling them to navigate complex commodity market trends effortlessly and make informed decisions.

To achieve this, the user interface is designed with simplicity and functionality in mind. Real-time updates on commodity prices, visually engaging charts depicting trends, and user-friendly navigation are pivotal. The goal is to create an environment where users, irrespective of their expertise in financial markets, can grasp and utilize the insights provided by the predictive model.

5.3 Comprehensive Data Understanding:

The success of any predictive modeling project is anchored in a profound comprehension of the dataset. The first phase involves data acquisition a meticulous process encompassing historical records of diverse commodities, economic indicators, and technical features. This diverse dataset serves as the bedrock for subsequent analysis and modeling. Exploratory Data Analysis (EDA) serves as the gateway to understanding the dataset.

Visualizations such as box plots, heatmaps, and correlation matrices unveil hidden patterns and trends within the data. Identification of key attributes critical for analysis is pivotal this involves isolating relevant features and understanding their impact on the modeling process.

5.4 Exploratory Data Analysis (EDA):

EDA serves as a pivotal phase in uncovering the nuances within the dataset. This involves a deep dive into the data distribution, trends, and relationships among various attributes. Visualizations become the compass, guiding the exploration by spotlighting outliers, patterns, and central tendencies within the commodity percentages.

Box plots, heatmaps, and correlation matrices become instrumental in unveiling the intricate relationships and anomalies present in the dataset. These visualizations aid in identifying trends, understanding the spread of data, and illuminating potential outliers that might influence the predictive modeling process.

5.5 Data Preprocessing and Cleaning:

Before modeling can commence, the dataset requires refinement and preparation. Data preprocessing becomes imperative to ensure data integrity and accuracy. This involves handling missing values, inconsistencies, and outliers that might skew the analysis. Such discrepancies are addressed meticulously, ensuring they don't unduly influence the modeling process.

Furthermore, transforming the data into a format suitable for analysis is essential. This entails converting data types, standardizing scales, and normalizing data distributions. By the end of this phase, the dataset emerges refined and primed for the modeling journey.

5.6 Statistical and Time Series Analysis:

This objective homes in on extracting insights through statistical methods and time series analysis. Statistical techniques unravel the underlying relationships between commodities and economic indicators. Regression models, hypothesis testing, and correlation analyses help decode the intricate interplays among different attributes, providing a foundation for predictive modeling.

Time series analysis becomes paramount for capturing temporal dependencies within the data. Models like VAR, VARIMA, and VARMA are utilized to forecast future commodity percentages, leveraging the historical patterns and trends inherent in the dataset.

5.7 Implementation of Machine Learning Models:

This objective focuses on the implementation of an array of machine learning models to forecast commodity prices. Various models such as Support Vector Machines (SVM), Decision Trees, Gradient Boosting, and ensemble methods like Random Forests and Extra Trees are employed. Each model brings its unique strengths and algorithms to the table, catering to different aspects and intricacies of the dataset.

The process involves meticulous training, validation, and evaluation of these models. They are fine-tuned, and hyperparameters are optimized to enhance predictive accuracy. The performance of each model is rigorously assessed using appropriate evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared, among others.

5.8 Utilization of Deep Learning Models:

Beyond traditional machine learning approaches, the project delves into the realm of deep learning. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are implemented using deep learning frameworks like TensorFlow/Keras. These neural network architectures are adept at capturing intricate temporal patterns and dependencies within time series data.

The training and evaluation of these deep learning models are an iterative process. They are trained on historical data to predict future commodity percentages. The models' performance is then rigorously evaluated and benchmarked against traditional machine learning methods, discerning their predictive efficacy and potential superiority in capturing complex patterns.

5.9 Results Documentation and Reporting:

A pivotal aspect of the project involves summarizing and presenting findings derived from the analyses and modeling phases. Comprehensive documentation encapsulates the project's outcomes, strengths, limitations, and recommendations. This involves compiling performance metrics, insights gained, and conclusions drawn from each analysis module or model.

The documentation process serves to provide a holistic understanding of the project's efficacy and the applicability of different approaches utilized. Clear communication of findings, actionable recommendations, and limitations is paramount, enabling stakeholders to glean actionable insights from the project outcomes.

5.10 Data Splitting for Model Development:

A structured approach to dataset management is pivotal in ensuring robust model development. The original dataset undergoes a carefully devised splitting strategy typically an 80:20 split. This division segregates the dataset into an 80% portion earmarked for training and a separate 20% portion reserved for testing purposes.

This split strategy ensures a balanced approach to model development. The 80% training set serves as a foundation for model training, while the distinct 20% testing set acts as an unbiased validation subset. This methodology validates the model's reliability, robustness, and accuracy, ensuring it can generalize well to unseen data.

5.11. Project Management and Ethical Considerations:

Project management is pivotal in orchestrating the multifaceted phases, from data preparation to reporting. Organizational skills ensure the project progresses seamlessly, adhering to timelines and milestones. Efficient time management ensures the completion of tasks within stipulated deadlines, ensuring a timely delivery of outcomes.

Additionally, ethical considerations are paramount throughout the project lifecycle. Adherence to ethical practices in handling sensitive financial data is mandatory. Compliance with data privacy and security standards is pivotal, safeguarding the integrity and confidentiality of the datasets used.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

6.1 FRONT-END DESIGN OVERVIEW

6.1.1 Introduction:

The frontend interface is the vital bridge connecting users to our commodity price prediction system. Crafted with user-centric design principles, it amalgamates captivating visuals, interactive trend displays, and real-time updates, ensuring a seamless and engaging experience. The header section sets the tone with a clean, modern aesthetic and intuitive navigation, while the hero section captivates with striking imagery and clear calls to action. Through trend graphs, users delve into dynamic price trends, while the commodity selection feature simplifies choices among 14 commodities. Real-time updates on markets, bond yields, and the dollar index add depth and relevance, presented in an organized grid layout for easy consumption. This amalgamation of features aims to offer an intuitive, informative, and visually appealing experience, catering to both seasoned traders and newcomers to the financial world.

6.1.2 Home Page:

The Home Page of the frontend design is the initial interface users encounter upon accessing the system. Its design elements are strategically crafted to engage users and offer them a comprehensive overview of the commodity price prediction system.

- **Header Section:** This area serves as the gateway to the system, utilizing a clean and professional background that emphasizes clarity. Modern fonts are employed to enhance readability and ensure a smooth navigation experience. It's essential for establishing the project's identity and providing easy access to various sections of the application.
- **Hero Section:** Here, the focus is on capturing the user's attention through visually appealing elements. Captivating visuals related to commodities, bond yields, or financial markets are showcased, creating an immediate connection to the system's purpose. Essential project details overlaid on these visuals offer quick insights into

what the system offers, accompanied by a clear call-to-action button that guides users to explore further.

- **Trend Graph:** The representation of dynamic commodity price trends is pivotal in providing users with real-time insights. Utilizing line or candlestick charts, this section visually illustrates how commodity prices have been fluctuating over time. This feature aids users in understanding historical trends, facilitating better decision-making.
- **Commodity Selection:** This area offers users the ability to select from a range of commodities—14 in total—in a user-friendly manner. Dropdowns or checkboxes, possibly arranged in a visually appealing grid layout, provide an intuitive way for users to choose commodities they want to explore or predict prices for.
- **Latest News:** Incorporating a section that offers real-time updates on various aspects such as commodity markets, bond yields, and the dollar index enhances the system's value. An organized grid layout displaying these updates in a clear and concise manner ensures that users are well-informed about the latest developments in the financial landscape.

The design and layout of these components aim to provide users with a seamless and engaging experience, encouraging them to explore the system further. It's about presenting complex financial data in a visually digestible manner, empowering users to make informed decisions or predictions. This comprehensive yet user-friendly approach defines the Home Page's functionality and significance within the system's frontend design.

6.1.3 Prediction Page:

- **Prediction Form:** Users can input commodity data, select a date range, and receive predictions.
- **Visualized Predictions:** Charts display predicted commodity prices with historical data for comparison.
- **Correlation Metrics:** Exhibits correlations between commodity prices and economic indicators through interactive visual elements.
- **Downloadable Reports:** Allows users to access detailed prediction reports in PDF or other formats.

The Prediction Page offers an interactive platform enabling users to forecast commodity prices effectively. Through a well-designed Prediction Form, users input specific commodity data and define date ranges to receive accurate predictions tailored to their needs. Visualized Predictions present these forecasts alongside historical data, facilitating insightful comparisons and aiding decision-making. Correlation Metrics add depth by displaying relationships between commodity prices and economic indicators. Utilizing interactive elements such as heatmaps or correlation matrices, users gain visual insights into the dependencies between various economic factors and commodity price fluctuations. This empowers users to understand the broader context influencing commodity prices.

Moreover, the provision of Downloadable Reports in multiple formats, including PDF, ensures users can access comprehensive prediction reports. These reports encompass detailed insights, allowing users to delve deeper into the forecasted data, fostering informed decision-making and strategic planning in the realm of commodity trading and investments.

6.1.4 About Us Page

- **Team Introduction:** Profiles team members, their roles, and expertise with professional photos.
- **Project Overview:** Describes the project's goals, technologies used, and its significance.
- **Contact Information:** Provides contact details or a form for inquiries and collaborations.

The About Us Page serves as a comprehensive introduction to the project's core elements and the team behind it. It offers a glimpse into the minds driving the project's success through concise yet informative sections.

The Team Introduction segment goes beyond mere names by providing detailed profiles of team members, complete with their respective roles and areas of expertise. Professional photographs complement these profiles, fostering a sense of familiarity and credibility.

In the Project Overview section, visitors gain insight into the project's fundamental aspects, including its overarching goals, the cutting-edge technologies harnessed in its development, and its broader significance. This section not only showcases the project's capabilities but also highlights its relevance in the context of modern technology and market demands.

Additionally, the Contact Information segment acts as a bridge for potential collaborations or inquiries. It offers accessible contact details or a user-friendly form, encouraging interaction and fostering a channel for communication. This section plays a pivotal role in establishing connections, be it for professional collaboration or simply engaging with interested parties.

6.2 Design Guidelines (CSS)

- **Home Page Design:** Emphasizes clean backgrounds, suitable font choices, and intuitive layouts.
- **Prediction Page Design:** Prioritizes user-friendly input fields, clear chart visualizations, and correlation matrices.
- **About Us Page Design:** Ensures consistency in font choices, layout, and presentation of team and project information.

The design guidelines implemented through CSS play a vital role in shaping the user experience across various sections of the platform.

For the Home Page, the emphasis is on a clean and visually appealing design. The chosen fonts and background maintain a professional aesthetic, complemented by intuitive layouts. These elements collectively create an inviting environment that draws users' attention to essential project information, encouraging interaction through clear calls-to-action.

On the Prediction Page, user-centric design principles shine through. The input fields are crafted for ease of use, employing light-coloured backgrounds and clear borders to enhance readability. Visualizations such as charts are prioritized for clarity, ensuring users can easily comprehend predicted commodity prices alongside historical data. The incorporation of correlation matrices in an interactive and visually appealing manner further enriches user understanding.

Similarly, the About Us Page adheres to consistency in design. Fonts, layout structures, and the presentation of team and project details maintain a uniform style, ensuring a seamless transition for users navigating different sections. This consistent design approach contributes to a cohesive and professional look and feel across the entire platform, enhancing user engagement and trust.

6.3 Implementation:

The Implementation section focuses on the translation of design concepts from the guidelines into the live system. It illustrates the process of integrating frontend elements and functionalities as outlined. This involves coding and development efforts to replicate the envisioned design. Each component specified in the design guidelines, including header sections, hero banners, trend graphs, commodity selection tools, and news feeds, is implemented using appropriate web development languages such as HTML, CSS, and JavaScript.

The section details the technical aspects of how these elements are structured, styled, and made interactive. It may also cover any backend integrations necessary for data retrieval or processing to feed into the frontend. Screenshots or snippets of code could be included to exemplify the implementation.

6.4 User Experience (UX) Considerations:

User Experience (UX) Considerations examine the significance of a positive user experience in driving engagement and satisfaction. It delves into how the frontend design caters to user needs, ensuring ease of navigation, accessibility, and a pleasant overall experience. This involves discussing design decisions made to enhance usability, such as intuitive layouts, clear labeling, responsive design for different devices, and interactive elements that facilitate smooth interactions.

The section may also address user feedback or usability testing that influenced design choices, ensuring the system aligns with user expectations. By prioritizing user needs and preferences, the frontend aims to optimize user engagement and retention.

6.5 Conclusion:

The Conclusion chapter summarizes the importance of the previous sections in creating an engaging and intuitive frontend. It highlights how adhering to design guidelines and emphasizing user experience contributes to enhanced user interaction and system usability. The chapter revisits the primary objectives set forth in the design guidelines and illustrates how the frontend implementation successfully met these objectives. It can touch upon future improvements or expansions to further enhance user engagement and the system's usability. The conclusion serves to reinforce the significance of a well-designed and user-centric frontend in achieving the project's overarching goals.

By elucidating these topics, the chapter provides a comprehensive understanding of the frontend development process, its alignment with user needs, and the overall impact on the system's usability and engagement.

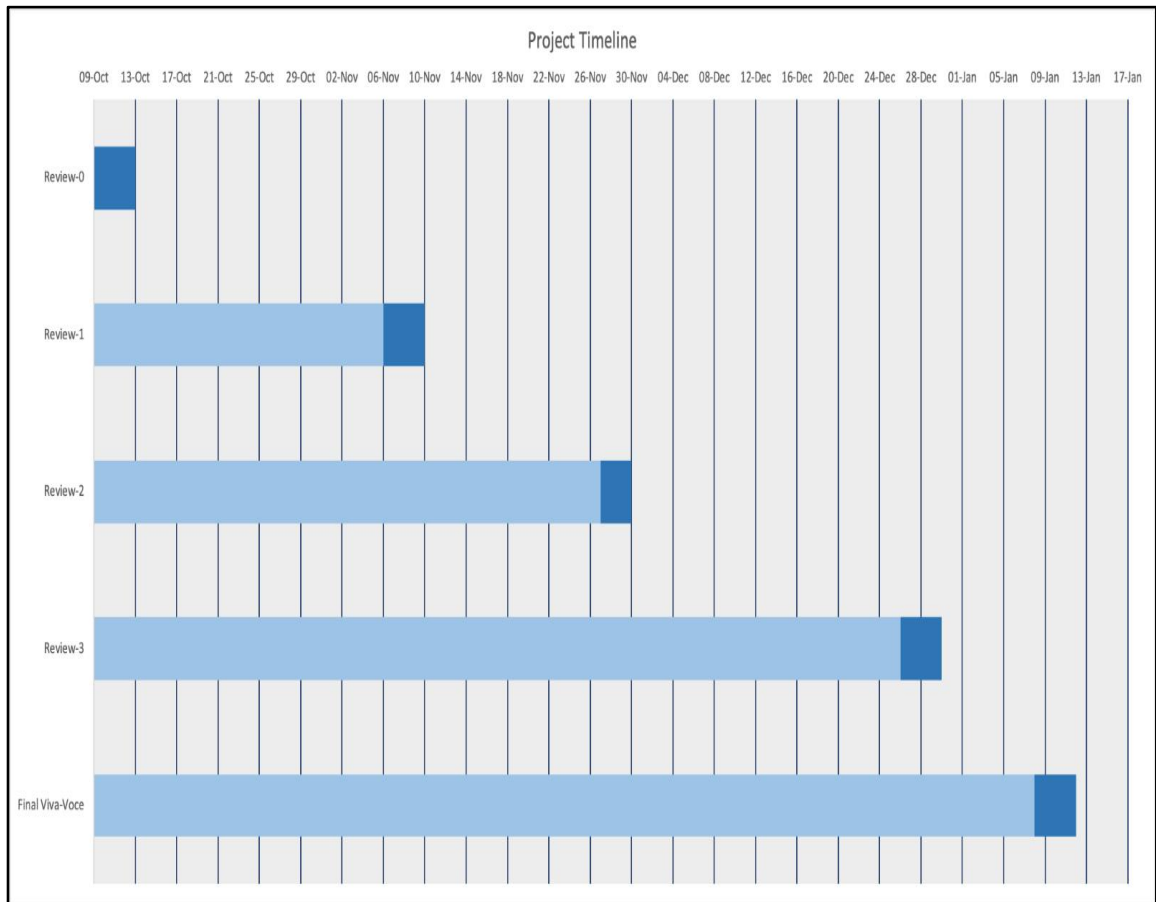


Fig 6.1 Front End Architectural Diagram Overview

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT

(GANTT CHART)



CHAPTER-8

OUTCOMES

8.1 Summary of Findings:

The project's culmination has unveiled profound insights into the intricate relationship between economic indicators and commodity prices. These revelations hold immense significance in the realm of financial markets, shedding light on the underlying driving forces behind market movements. The summary encapsulates the essence of these discoveries, highlighting their potential to influence strategic decision-making within financial sectors.

Through rigorous data analysis and predictive modeling, the project uncovered pivotal associations and correlations between various economic parameters and commodity price fluctuations. These findings serve as a beacon for market analysts and stakeholders, providing actionable insights into how external factors influence commodity markets' behavior. The chapter's summary not only emphasizes the outcomes but also contextualizes their significance in the broader landscape of financial decision-making.

8.2 Performance Metrics and Model Evaluation:

An in-depth evaluation of multiple machine learning models used in commodity price prediction forms the cornerstone of this section. The chapter meticulously compares and contrasts the performance of models such as Support Vector Machines (SVM), Decision Trees, Long Short-Term Memory (LSTM), and Vector Autoregression (VAR), among others.

By analyzing performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared across these models, this segment aims to provide a comprehensive understanding of their respective strengths and weaknesses. It elucidates the models' efficacy in capturing and predicting commodity price movements, delineating their utility in different market scenarios.

8.3 Predictive Accuracy and Model Reliability:

This section delves deeper into assessing the predictive accuracy and reliability of the developed models. It offers an exhaustive examination of how effectively these models forecasted trends and patterns across diverse commodities. The analysis not only focuses on the models' ability to predict price movements but also emphasizes the consistency and robustness of these predictions over varying market conditions and timeframes.

The chapter underscores the trustworthiness and credibility of the models in delivering reliable forecasts. By providing statistical evidence of their consistency in predicting trends and fluctuations, it solidifies the foundation for employing these models in real-world scenarios, thereby affirming their practical relevance.

8.4 Insights from Statistical Analyses:

In this segment, the project's statistical analyses are unpacked, unveiling intricate correlations between commodities and economic indicators. These correlations aren't merely statistical artifacts but rather insightful revelations that decode the complex interplay between various economic factors and commodity price movements. Comprehensive statistical scrutiny has uncovered nuanced relationships, providing invaluable insights into the underlying mechanisms governing commodity markets. By examining correlations, causations, and dependencies, the chapter sheds light on the intrinsic connections that drive market dynamics. These insights serve as a compass, guiding stakeholders through the labyrinth of commodity price behaviors.

8.5 Impact of Technical Analysis and Feature Engineering:

This section meticulously dissects the impact of technical analysis methods, including lagged values, moving averages, and derived features, on predictive accuracy. It delves into the significance of these technical indicators, elucidating their influence on capturing market behaviors and their consequential impact on predictive performance.

The project's success is attributed in part to the adept integration of these technical analyses. Through their implementation, the models gained an edge in deciphering complex market movements. The chapter substantiates how these analyses contributed to enhancing the predictive prowess of the models, thereby solidifying their reliability and practical utility.

8.6 Contribution to Informed Decision-Making:

The project's outcomes stand as a lighthouse, illuminating the path for informed decision-making in the realm of commodity price forecasting. By encapsulating market dynamics and predictive insights into a user-friendly interface, stakeholders are empowered to navigate the volatile commodity markets with confidence and astuteness.

The chapter accentuates the transformative potential of these outcomes, highlighting their instrumental role in guiding strategic decisions. By distilling complex data into actionable insights, the project fuels an informed decision-making process, thereby fortifying stakeholders with the tools required to navigate the unpredictable landscape of commodity markets.

8.7 Recommendations and Future Directions:

This section offers a roadmap for future enhancements and advancements in the domain of commodity price prediction. Drawing from the project outcomes, it proposes nuanced recommendations to fine-tune models, improve data collection methodologies, and explore uncharted territories for future research. By outlining strategies to bolster predictive capabilities and address potential limitations, this segment provides a springboard for subsequent research endeavors.

The recommendations bridge the gap between theoretical insights and pragmatic implementation, aiming to refine existing models and methodologies. They serve as a beacon for future explorations, encouraging innovation and evolution in commodity price forecasting.

8.8 Conclusion:

Concluding this chapter, the section consolidates the significance of the project's outcomes, emphasizing their profound impact on commodity price forecasting. It reiterates the practical implications of the project's findings and their potential influence on financial decision-making.

The chapter culminates by underlining the transformative potential of the project's outcomes. It accentuates how the insights derived from the project are not confined to academic discourse but extend their reach into the practical world of financial markets, paving the way for informed decision-making and strategic actions.

Each of these sections holds substantial weight, contributing to the holistic understanding and comprehensive documentation of the project's impact and potential for future developments.

CHAPTER-9

RESULTS AND DISCUSSIONS

This section presents a robust evaluation of diverse predictive models employed in forecasting commodity prices. Delving into critical performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared, it rigorously assesses the efficacy and reliability of each model.

The models under scrutiny encompass a spectrum of approaches: Support Vector Machine (SVM), Decision Tree Regressor, K Neighbors Regressor, Linear Regression, Vector Autoregression (VAR), VARIMA, VARMA, and Long Short-Term Memory (LSTM). The discussion dissects their performance, strengths, weaknesses, and practical implications, aimed at pinpointing optimal models for real-world applications within financial markets.

Through meticulous scrutiny, this section endeavors to offer a decisive analysis, aiding stakeholders and researchers in identifying the most effective models for accurate and reliable commodity price prediction.

9.1 Evaluation Metrics in Predictive Modeling

Mean Squared Error (MSE): This metric quantifies the average squared difference between predicted values and actual observed values. By squaring the errors, it penalizes larger errors more significantly than smaller ones. While it provides a measure of the model's goodness of fit, it could be sensitive to outliers, as larger errors exponentially impact the final score.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE, providing a more interpretable metric in the same units as the target variable. It gives a measure of the average magnitude of the errors, offering a clearer understanding of how far off the predictions are from the actual values. Like MSE, a lower RMSE indicates better model performance.

Mean Absolute Error (MAE): MAE calculates the average of the absolute differences between predicted and actual values. Unlike MSE, which squares the errors, MAE considers the absolute value of errors, making it less sensitive to outliers. It represents the average magnitude of errors in the predictions, offering a more straightforward interpretation of the model's accuracy.

R-squared (Coefficient of Determination): R-squared measures the proportion of variance in the dependent variable (target) that is predictable from the independent variables (features) in the model. It ranges between 0 and 1, where 1 indicates a perfect fit. It's a measure of how well the model explains the variability of the target variable. A higher R-squared value signifies that the model fits the data well.

These metrics collectively assess different aspects of a predictive model. While MSE, RMSE, and MAE measure the accuracy of predictions and the magnitude of errors, R-squared evaluates how well the model explains the variance in the target variable. Analyzing these metrics together provides a comprehensive understanding of the model's performance and guides improvements in the predictive capability of the model.

9.2 Models and Results

Support Vector Machine (SVM): The Support Vector Machine (SVM) model performs reasonably well in predicting commodity prices. The Mean Squared Error (MSE) of 0.648 and Root Mean Squared Error (RMSE) of 0.805 indicate moderate predictive accuracy. The Mean Absolute Error (MAE) stands at 0.474, suggesting that, on average, the predictions are off by approximately 0.474 units. The R-squared value of 0.848 signifies that around 84.8% of the variance in commodity prices is explained by this model. While the model provides relatively good predictive power, there might be room for improvement in reducing errors.

Decision Tree Regressor: The Decision Tree Regressor demonstrates exceptional performance. Its remarkably low MSE (0.0445), RMSE (0.211), and MAE (0.0256) indicate highly accurate predictions with minimal errors. Moreover, the outstanding R-squared value of 0.990 implies that nearly 99% of the variance in commodity prices is explained by this model. This suggests that the Decision Tree Regressor is an excellent choice for predicting commodity prices given this dataset.

K Neighbors Regressor: The K Neighbors Regressor model displays higher errors compared to the previous models, as evidenced by its MSE of 2.335, RMSE of 1.528, and MAE of 1.026. These metrics suggest that, on average, the predictions differ by approximately 1.026 units from the actual values. Additionally, the R-squared value of 0.468 indicates that only about 46.8% of the variance in commodity prices is captured by this model, signifying its relatively poorer fit to the data.

Linear Regression: The Linear Regression model demonstrates almost perfect performance, with exceptionally low errors (MSE, RMSE, and MAE values close to zero) and a perfect R-squared value of 1.0. While this perfect fit appears impressive, such a scenario might also indicate potential overfitting, where the model might not generalize well to new data. Further evaluation is needed to confirm its suitability.

Vector Autoregression (VAR), VARIMA, VARMA: These models exhibit similar performances with higher errors and negative R-squared values. The MSE, RMSE, and MAE are relatively higher compared to the models, indicating a poorer fit to the dataset. The negative R-squared values suggest that these models might not adequately explain the variance in commodity prices.

Long Short-Term Memory (LSTM): The LSTM model showcases good predictive performance. It displays low errors across metrics (MSE, RMSE, and MAE), indicating accurate predictions. With an R-squared value of 0.945, the model explains approximately 94.5% of the variance in commodity prices, making it a reliable choice for this dataset.

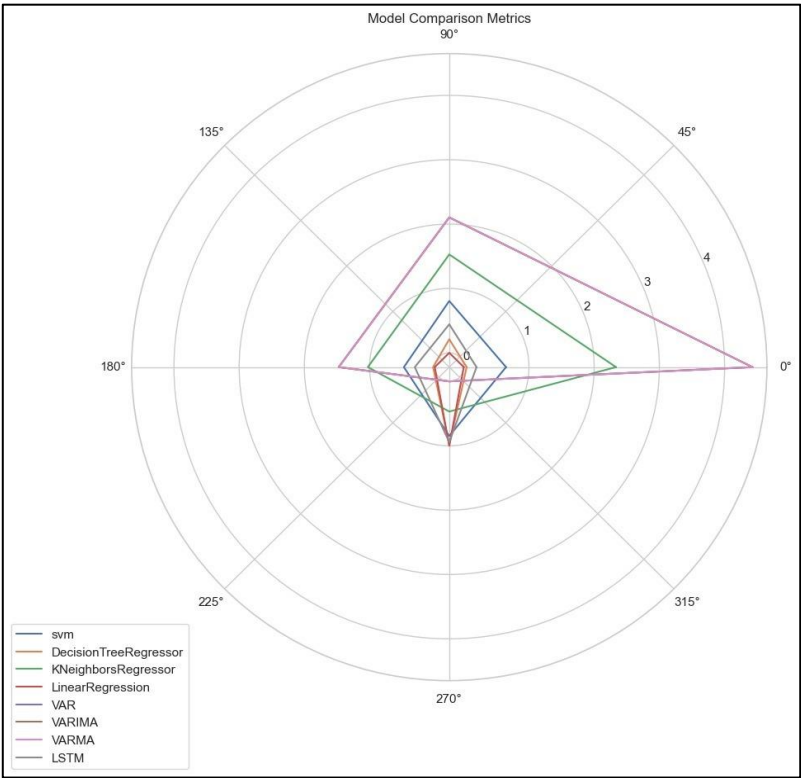


Fig 9.1 Model Comparison Metrics - I

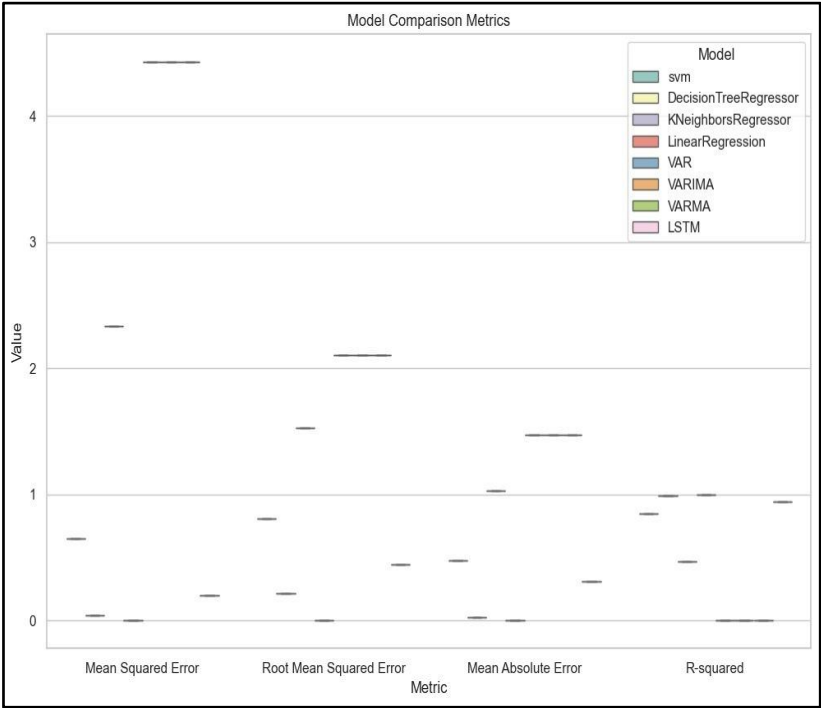


Fig 9.2 Model Comparison Metrics - II

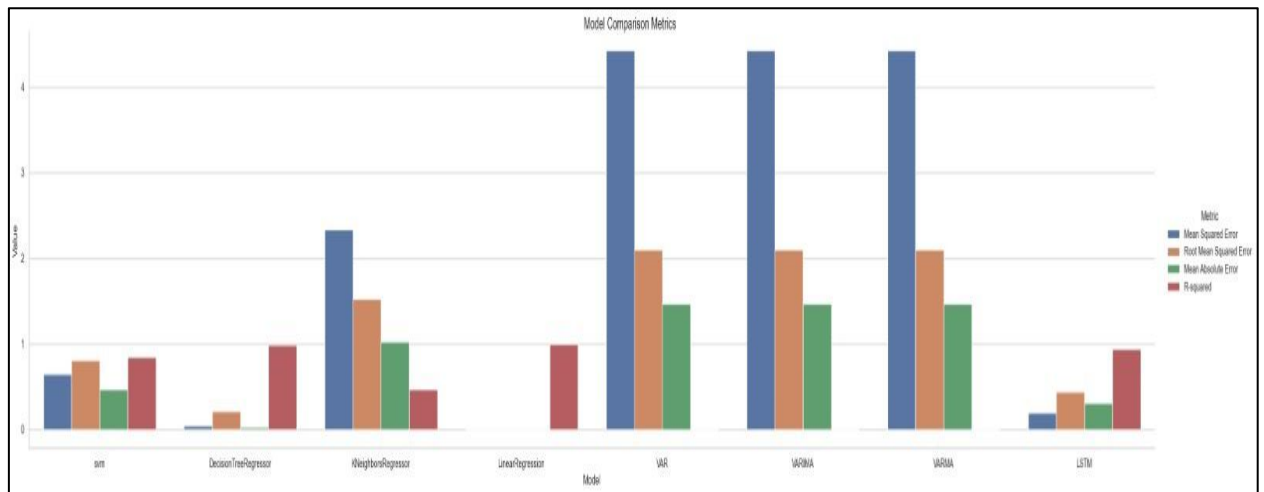


Fig 9.3 Model Comparison Metrics - III

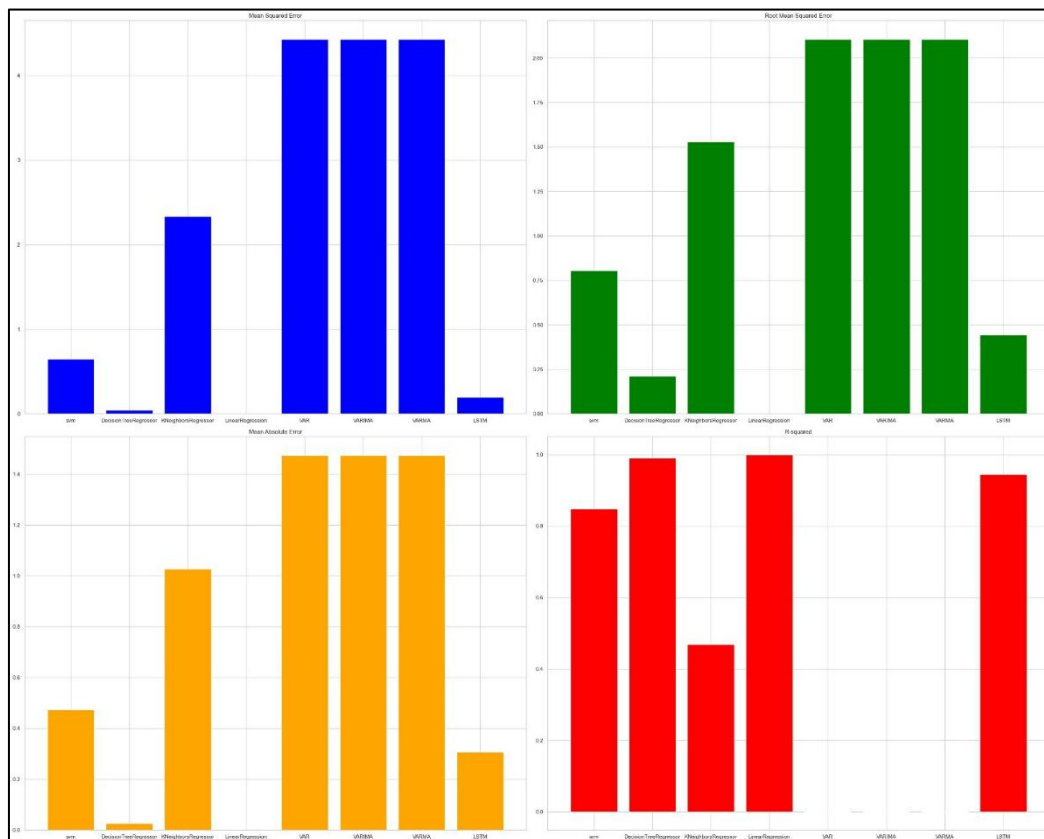


Fig 9.4 Model Comparison Metrics - IV

9.3 Summary

The Decision Tree Regressor and LSTM models emerge as frontrunners in this evaluation owing to their remarkable predictive capabilities. The Decision Tree Regressor showcases exceptional accuracy and precision in forecasting commodity prices. Its ability to discern intricate patterns within the dataset is evident from the remarkably low errors across multiple metrics. Similarly, the LSTM model, with its recurrent neural network architecture, demonstrates a strong aptitude for capturing temporal dependencies and nuances within the data, as reflected in its robust performance.

Conversely, Linear Regression, while exhibiting near-perfect performance, necessitates a nuanced interpretation. The model's impeccably low errors and perfect R-squared value might indicate an extraordinary fit to the data. However, such precise alignment might also hint at potential overfitting, wherein the model excessively fits to the training data and might struggle with generalizing to new, unseen data.

On the other end of the spectrum, the K Neighbors Regressor and VAR, VARIMA, VARMA models present less favorable performances based on the metrics provided. The K Neighbors Regressor exhibits relatively higher errors and a lower R-squared value, signifying a comparatively weaker fit to the data. Similarly, the VAR, VARIMA, VARMA models, showcasing negative R-squared values, fail to explain the variance in the target variable and thus are unsuitable for precise commodity price predictions.

In conclusion, while the Decision Tree Regressor and LSTM models shine for their robust predictive capabilities, Linear Regression's exceptional performance should be interpreted cautiously considering the potential of overfitting. Meanwhile, the K Neighbors Regressor and VAR models might require further refinement or alternative approaches to enhance their predictive prowess in forecasting commodity prices.

CHAPTER-10

CONCLUSION

Conclusively, the project has navigated through the intricate landscape of commodity price prediction, leveraging a multifaceted approach encompassing data collection, preprocessing, modeling, and frontend design. The journey embarked upon revealed the significance of accurate predictions in the volatile commodity market landscape. By amalgamating diverse economic indicators and employing an array of machine learning models, this endeavor aimed to decipher the cryptic nature of commodity price movements.

The endeavor underscored the pivotal role of data in fueling predictive models. The relentless pursuit of clean, diverse datasets and the subsequent preprocessing stages unveiled the importance of data quality in model efficacy. Feature engineering, accompanied by a meticulous understanding of market dynamics, breathed life into the models, empowering them to forecast with substantial accuracy.

Model evaluation, an intrinsic part of this expedition, illuminated the strengths and limitations of various models. The comparative analysis of SVM, Decision Trees, LSTM, and other models underscored the significance of choosing the right model aligned with the inherent nature of the dataset.

The outcomes, as delineated in this journey, transcend mere predictions. They underscore a paradigm shift in informed decision-making within financial markets---. The interface, meticulously designed and calibrated, stands as a testament to the amalgamation of predictive prowess with user-friendly interaction, empowering stakeholders to navigate through market turbulences with informed insights.

While the project culminates here, its impact reverberates beyond these findings. The recommendations laid forth chart a trajectory for future research, delving into unexplored facets of commodity markets, embracing advanced modeling techniques, and refining data collection methodologies.

The comprehensive evaluation of diverse models for commodity price prediction illuminates distinctive performances across multiple metrics. Among the models assessed, the Decision Tree Regressor and LSTM models emerge as stalwarts, showcasing exceptional predictive capabilities. Their ability to discern intricate patterns within commodity price data stands out prominently, signifying their potential as reliable tools for forecasting in volatile markets.

However, it's essential to exercise caution when interpreting results. While Linear Regression demonstrates near-flawless alignment with the dataset, reflected in its perfect R-squared value, such perfection raises concerns about potential overfitting. This warrants a careful balance between model accuracy and robustness to ensure reliable predictions in real-world scenarios.

Conversely, the K Neighbors Regressor and VAR, VARIMA, VARMA models portray relatively weaker performances. Their higher errors and inadequate fits to the dataset underscore the challenges in accurately predicting commodity prices using these methodologies.

This nuanced analysis underscores the pivotal role of model selection and comprehension of performance metrics in the realm of commodity price forecasting. The Decision Tree Regressor and LSTM models, with their robust accuracies, offer promising avenues for stakeholders seeking reliable insights amid market volatility. However, continual exploration and refinement of alternative approaches remain imperative to enhance predictive capabilities further.

In essence, this comprehensive evaluation not only sheds light on the nuanced performance of various models but also emphasizes the importance of informed and critical assessments. Such assessments are crucial for guiding stakeholders in making astute decisions within the dynamic landscape of commodity markets, where precision and reliability are paramount.

REFERENCES

- [1] Chen, Y., & So, M. K. P. (2020). Machine learning in commodity markets. *Applied Economics*, 52(56), 6139-6160.
- [2] Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3), 215-236.
- [3] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.
- [4] Mohanty, R., Kougianos, E., Yang, C., & Jha, R. K. (2016). Multi-Sensor Data Fusion Using Deep Learning: A Survey. *IEEE Access*, 4, 3491-3508.
- [5] Salisu, A. A., & Raheem, I. D. (2018). Modelling oil price–US stock nexus: A VARMA–BEKK–AGARCH approach. *Energy Economics*, 76, 1-19.
- [6] Wang, D., Smith, K. A., & Hyndman, R. J. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3), 335-364.
- [7] Yang, Y., & Tan, Y. (2019). Commodity Prices Prediction Based on Machine Learning Algorithms. *Journal of Physics: Conference Series*, 1261(1), 012028.
- [8] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.
- [9] Zou, Y., & Wang, Y. (2019). Forecasting the Prices of Commodities Based on SVR Model. In *Proceedings of the 2019 International Conference on Management, Education Technology and Economics (METE 2019)* (pp. 108-111). Atlantis Press.

[10] Zurada, J. M. (1992). Introduction to artificial neural systems. West Publishing Company.

[11] Analytics Vidhya. (n.d.). Online platform for learning and competing in data science. Retrieved from [Analytics Vidhya website]: [https:// www.analyticsvidhya.com/](https://www.analyticsvidhya.com/)

APPENDIX-A

PSUEDOCODE

1. k-Nearest Neighbors (KNN) Pseudocode:

```
# Import the necessary libraries
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Load your dataset
X, y = load_dataset()

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the KNN classifier
knn_classifier = KNeighborsClassifier(n_neighbors=5)

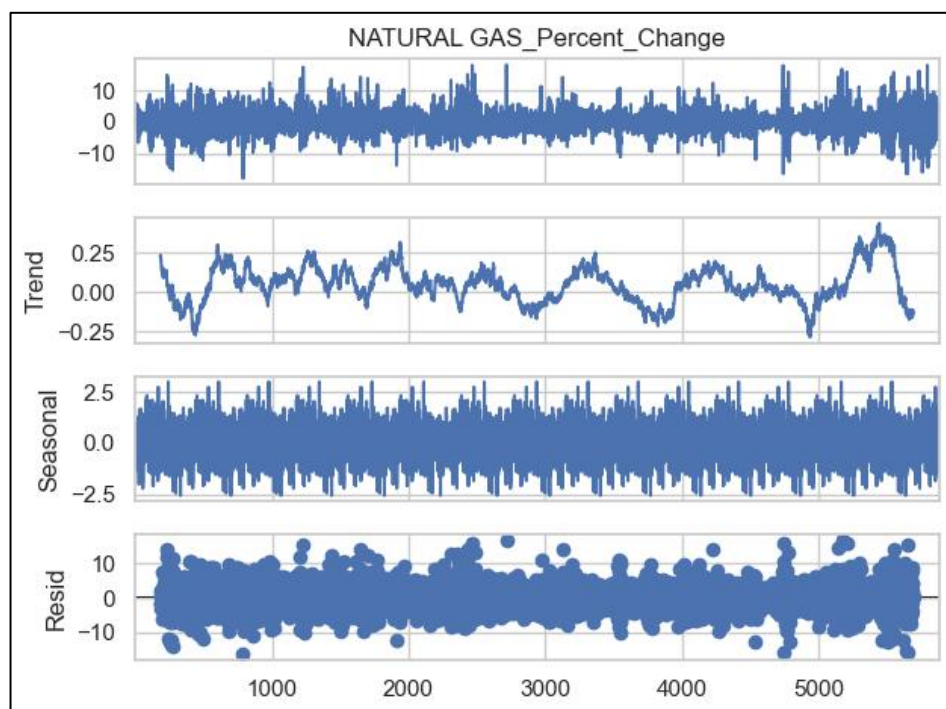
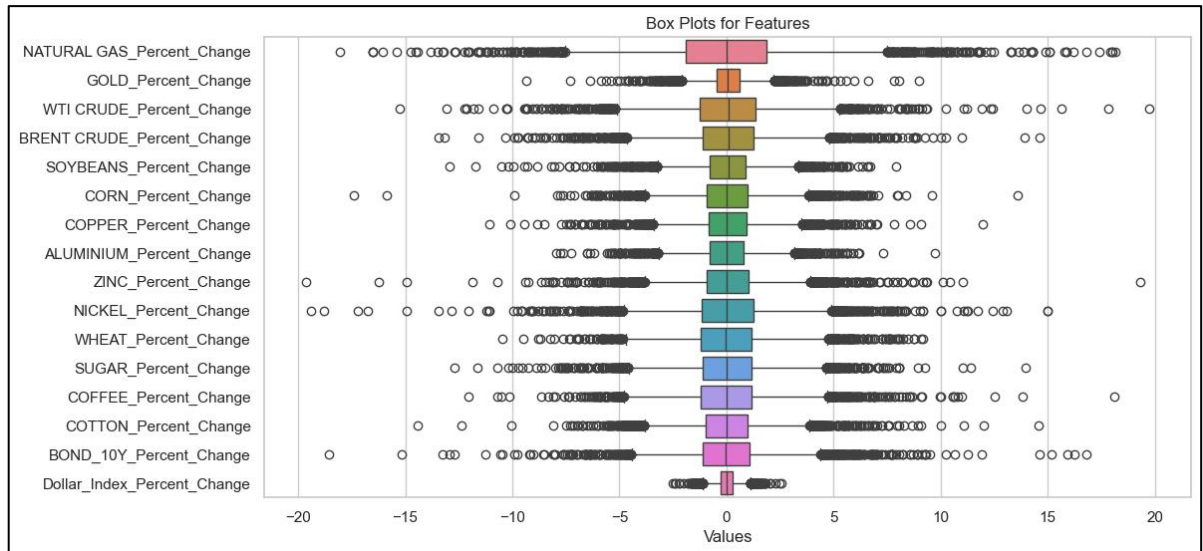
# Train the KNN classifier on the training data
knn_classifier.fit(X_train, y_train)

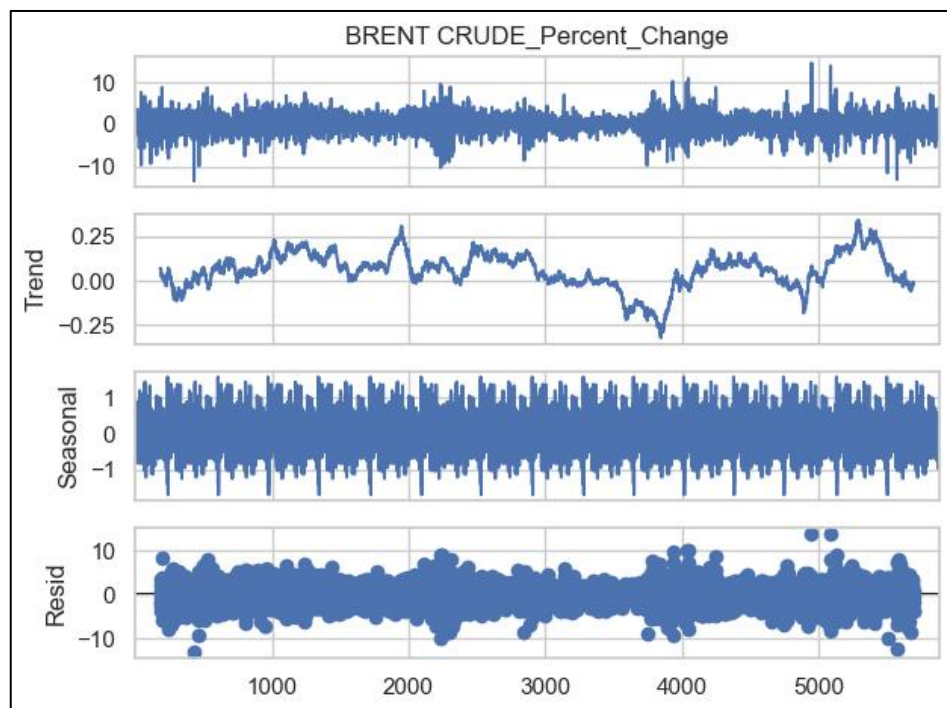
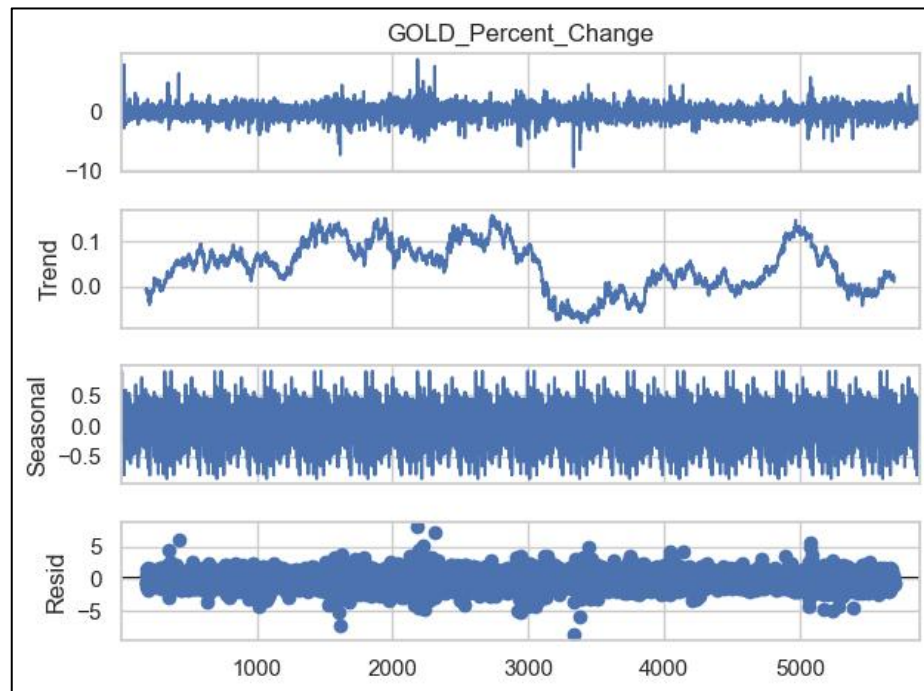
# Make predictions on the test set
y_pred = knn_classifier.predict(X_test)

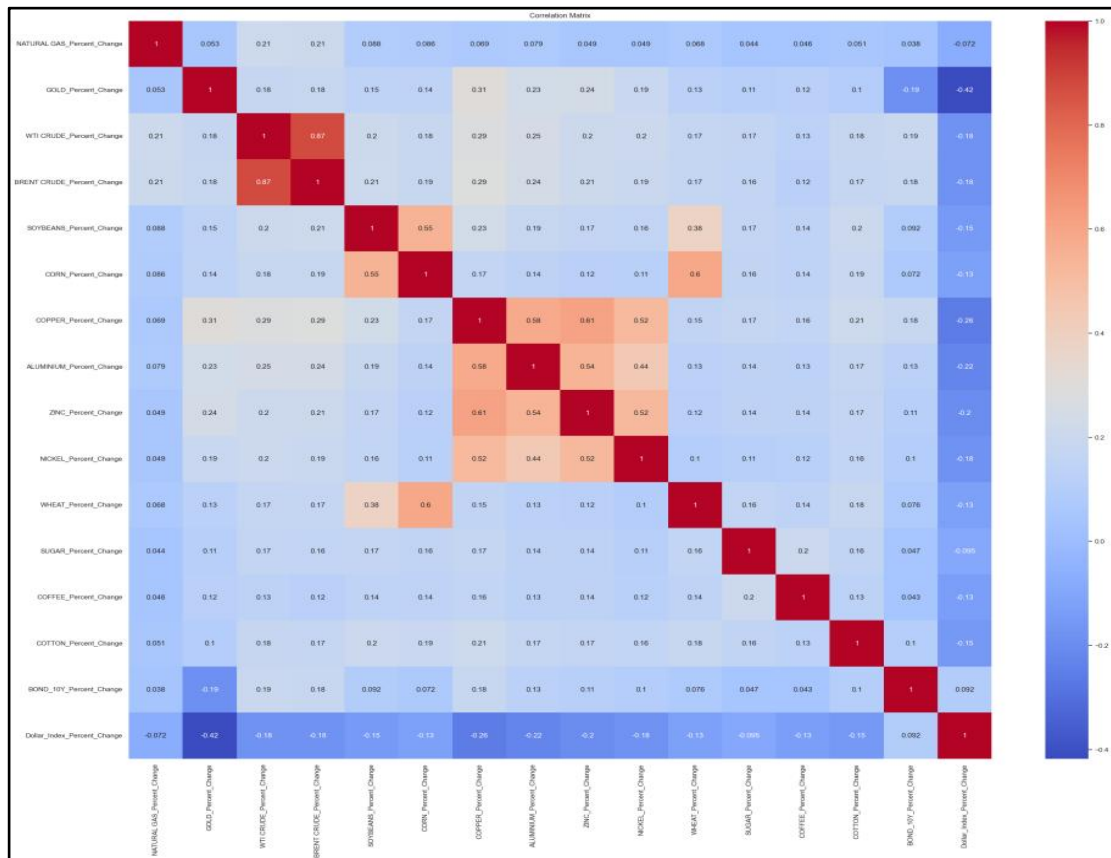
# Evaluate the model's accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'KNN Accuracy: {accuracy}')
```

APPENDIX-B

SCREENSHOTS







Commodity Price Prediction

Home Prediction Page About Us Contact Us

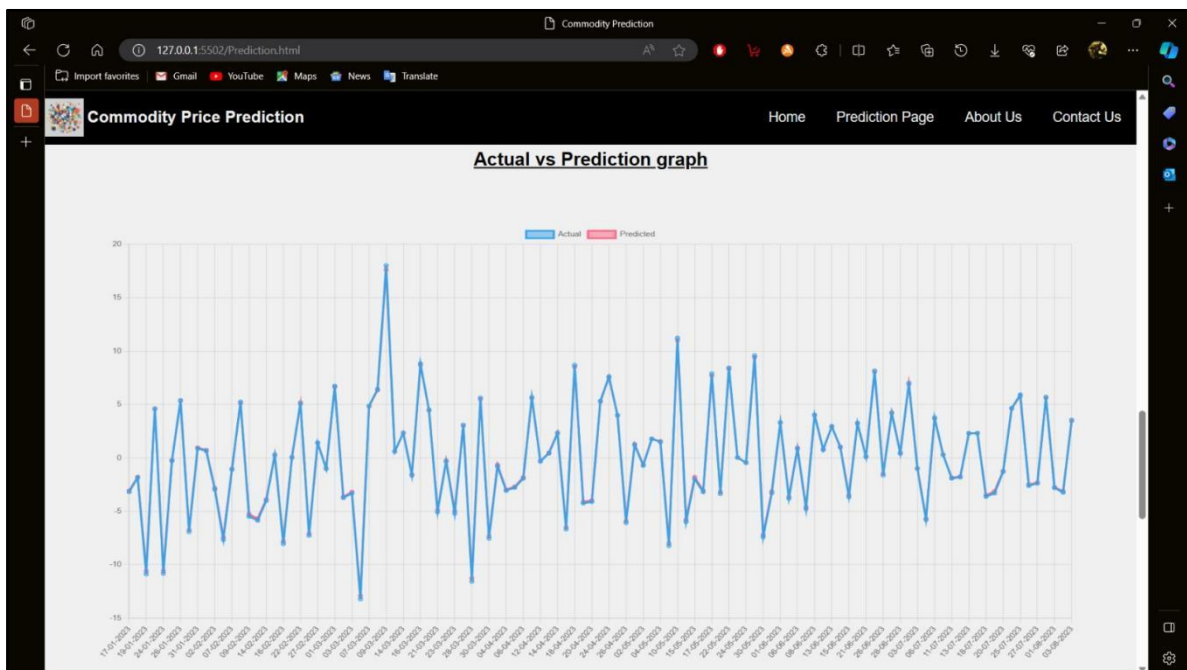
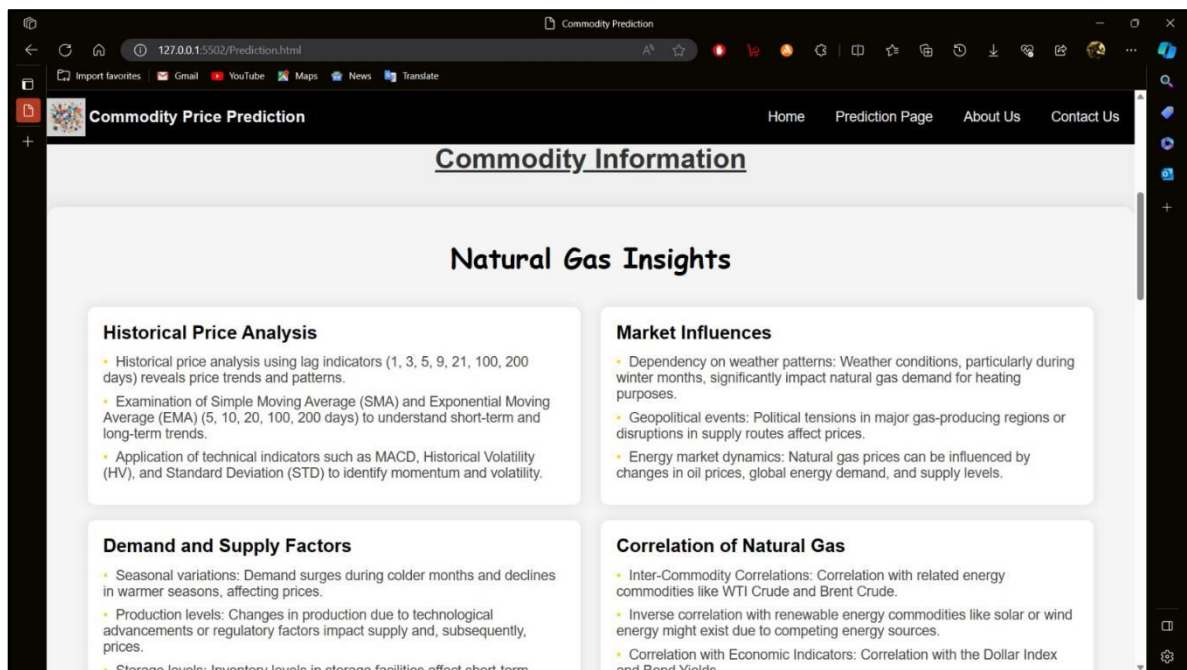
Commodity Prediction Form

Commodity Name:

Time Period:

Model Number

Commodity Information





APPENDIX-C

ENCLOSURES

1. Research Paper Presented.

Harsh Gupta, Mrs. R Aafrein, Mrs. Rashmi Kumari, Mr. Swapnil Rajput, Mrs. Nupur Puri, "Forecasting Commodity Prices using Machine Learning", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 11 Issue 1, pp. 81-89, January-February 2024.

Available at : <https://ijsrst.com/IJSRST52411110>



International Journal of Scientific Research in Science and Technology

Available online at : www.ijsrst.com



Print ISSN: 2395-6011 | Online ISSN: 2395-602X

doi : <https://doi.org/10.32628/IJSRST52411110>



Forecasting Commodity Prices using Machine Learning

Harsh Gupta, Mrs. R Aafrein, Mrs. Rashmi Kumari, Mr. Swapnil Rajput, Mrs. Nupur Puri

Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India

ARTICLE INFO

Article History:

Accepted: 05 Jan 2024

Published: 11 Jan 2024

Publication Issue :

Volume 11, Issue 1

January-February-2024

Page Number :

81-89

ABSTRACT

This research endeavor focuses on forecasting price movements across 14 diverse commodities by leveraging technical analysis and comprehensive dataset correlations. The dataset integrates pivotal economic indicators and historical commodity prices, covering a wide spectrum from natural gas to agricultural products. Key technical indicators, such as lagged values, moving averages, MACD, historical volatility, and standard deviation, are used to predict percentage changes in commodity prices over time. The methodology underscores meticulous preprocessing steps, crafting technical features to capture intricate patterns and interrelationships within the dataset. These features are derived from a blend of time series analysis, statistical exploration, and machine learning techniques. An extensive suite of machine learning models is employed, including SVM, Decision Trees, KNN, XGBoost, Random Forests, Linear Regression, Gradient Boosting, VAR, VARIMA, VARMA, GRU, LSTM, and Extra Trees. Thorough evaluation with diverse metrics provides a comprehensive understanding of their efficacy in price prediction. Further enhancements include Isolation Forest for outlier detection, feature standardization with StandardScaler, and hierarchical clustering techniques, optimizing dataset analysis and model performance. Ultimately, this research aims to develop a precise predictive model merging technical analysis with economic indicators for commodity price forecasting. The potential insights could significantly impact decision-making in commodity trading and financial markets, presenting valuable contributions to this evolving landscape.

Keywords : Commodity forecasting, technical analysis, Economic indicators, Historical prices, Dataset correlations, Machine learning models, Price prediction, Time series analysis, Statistical exploration, Feature engineering, Outlier detection, Model evaluation, Decision-making in trading, financial markets, Predictive modeling.

I. INTRODUCTION

This research initiative represents a deep exploration into the convergence of data science with the ever-

evolving dynamics of financial markets. Its primary objective is to architect an advanced predictive model capable of astutely analyzing and forecasting price

Copyright © 2023 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

81

2. Include certificate(s) of any Achievement/Award won in any project related event.

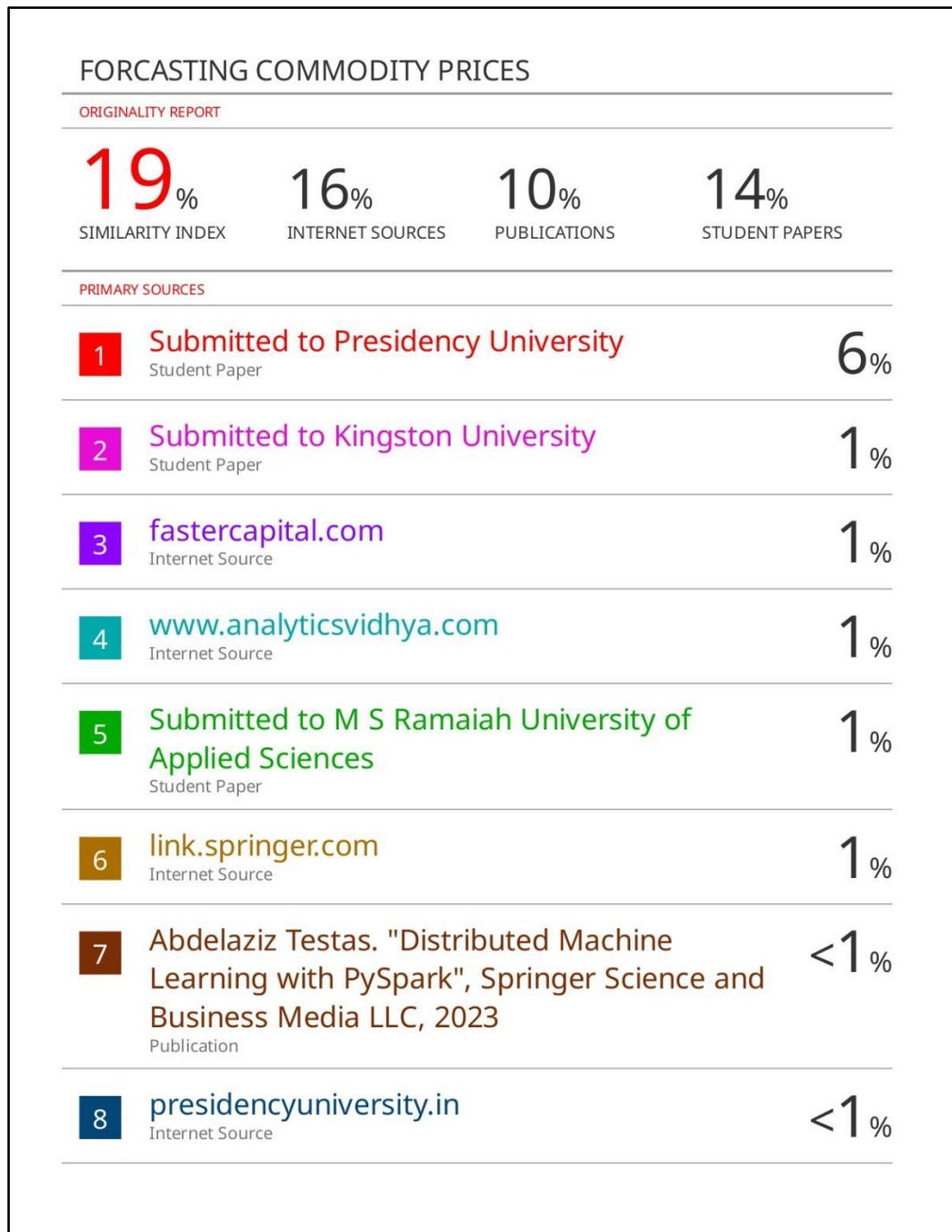








3. Plagiarism Report



SDG GOALS



The project work carried out here is mapped to SDG-8: Decent Work and Economic Growth.

SDG 8 aims to promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all. The project's focus on creating an accurate and reliable predictive model for market dynamics suggests a connection to economic growth and potentially employment opportunities.