Empowering Commodity Price Prediction Using XGBoost Algorithm

Contents

Abstract

Commodity price forecasting is a critical endeavor in today's dynamic economic landscape. This research delves into the accurate prediction of two vital commodities within Andhra Pradesh, India: petrol and electricity. These commodities are essential to our daily life and industry, subject to multifaceted influences. Leveraging the XGBoost algorithm's powerful predictive capabilities, we meticulously collect and analyze data, including factors like crude oil prices, exchange rates, tariffs, and coal prices, which impact these commodities. The project aims to enhance understanding of price dynamics and provide practical implications for energy management, economic planning, and investment decisions. These insights gained have practical implications for different industrial sectors, thereby offering tangible benefits to stakeholders in Andhra Pradesh's commodities market. The results show that using price influencing features can improve XGBoost model prediction accuracy than base models built on historical price data.

Introduction

- Commodities are often the building blocks of the global economy and are categorized into several broad groups like energy, metals, financial, and agriculture.
- These essential sectors, as measured by the Index of Eight Core Industries (ICI), en- compass coal, electricity generation, crude oil production, cement, natural gas, steel, refinery products, and fertilizer production. Core industries collectively contribute 40.27% to India's Index of Industrial Production (IIP).
- Accurate prediction of commodity prices is crucial for economic planning, inflation control, efficient supply chains, and informed investments, impacting energy sector stability, consumer budgets, and environmental considerations.
- Our research focuses on improving commodity price prediction accuracy through careful input data selection and model choice.

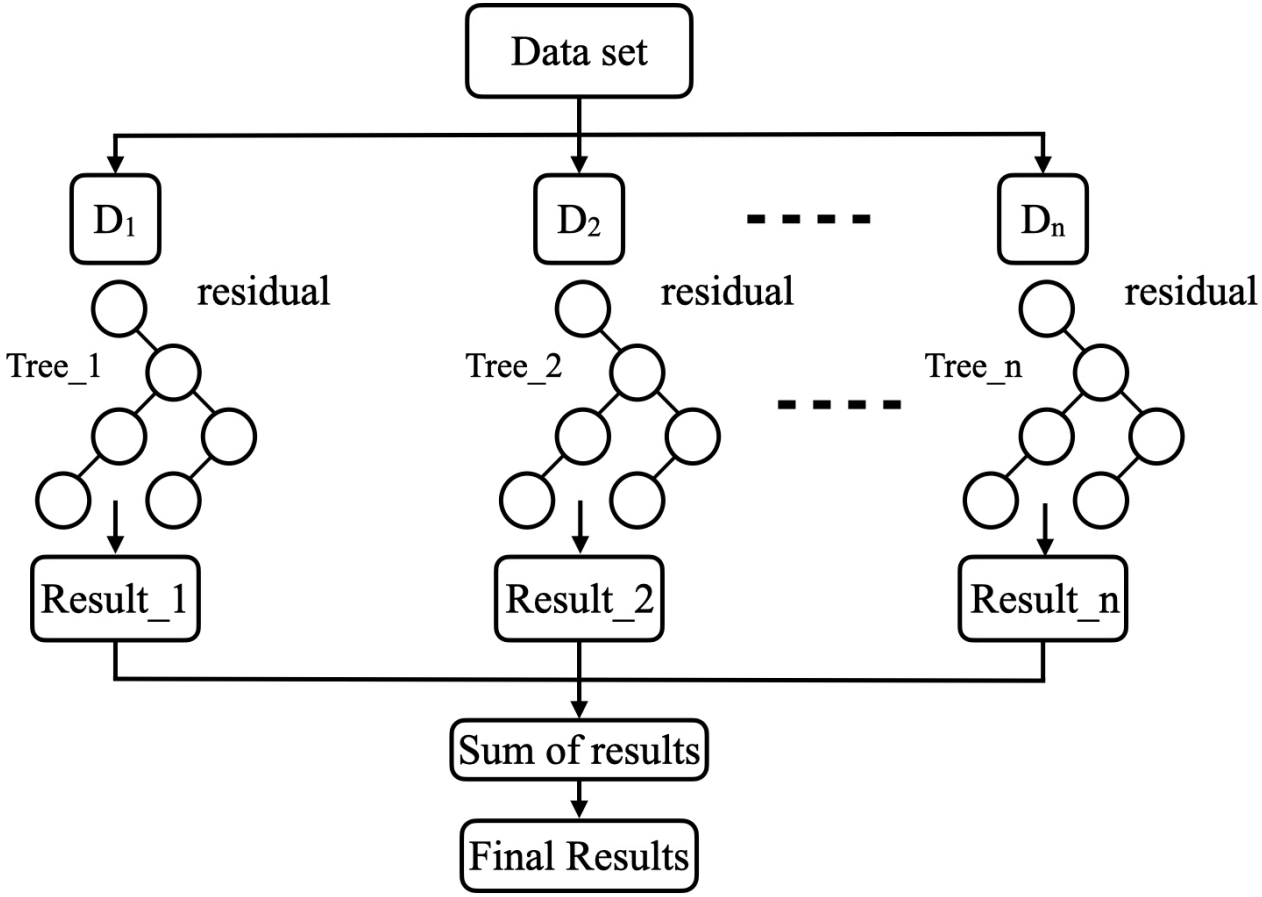
Proposed System

- Data preprocessing plays a fundamental role in ensuring that the raw data collected from various sources is clean, consistent, and ready for machine learning model training.
- The study's data covers petrol prices from January 2016 to March 2022 and electricity prices from May 2000 to March 2022. The primary variable of interest is the price, which is the target for prediction.
- Various factors influencing petrol prices are considered, including Crude Oil, USD/INR exchange rate, supply and demand dynamics, taxes and duties, and the annual inflation rate.
- Key factors like Tariff Rates, Coal Price, and Supply and Demand dynamics influence electricity pricing. This research contributes to enhanced energy management and data-driven decision-making in the local electricity market.

XGBoost Algorithm

- XGBoost is a highly regarded machine-learning technique created by Tianqi Chen in 2014. This ensemble method aggregates predictions from weak learners, typically decision trees, to form a precise model.
- Unlike single decision tree models, XGBoost efficiently ensembles them and can use parallel processing for speed. It offers control over tree depth (max_depth) for effective pruning and reduced overfitting.
- It's known for its exceptional predictive accuracy and robustness. XGBoost handles missing data and supports built-in cross-validation, simplifying model tuning.
- The algorithm starts with an initial prediction, iteratively creates decision trees to correct mistakes. It continues adding trees until it meets a stopping criterion.

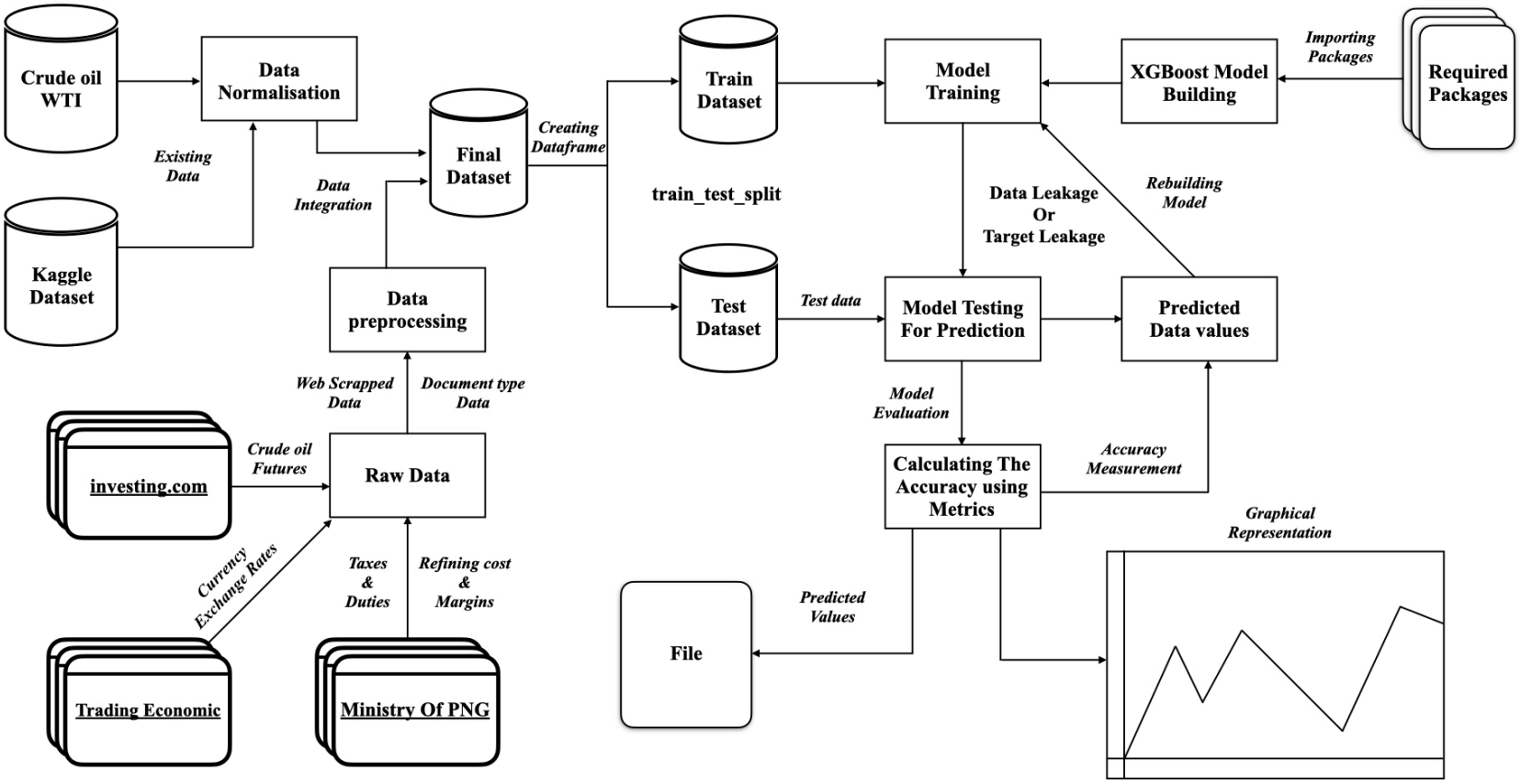
XGBoost Algorithm



System Architecture

- The system's architecture, begins with data collection from diverse sources, including government databases, financial markets, and Kaggle datasets in various formats.
- Data integration combines historical price data, influencing factors, and external variables into a cohesive dataset, ensuring consistency.
- Model training is where the algorithm learns from historical data, utilizing the training dataset. Model testing assesses its generalization to unseen data, revealing real-world performance and potential overfitting issues.
- Model evaluation goes further by using various techniques to assess strengths and weaknesses, including hyperparameter fine-tuning, aimed at maximizing performance metrics. Common regression evaluation metrics like MSE, RMSE, MAE, and R2 are used to measure model performance.

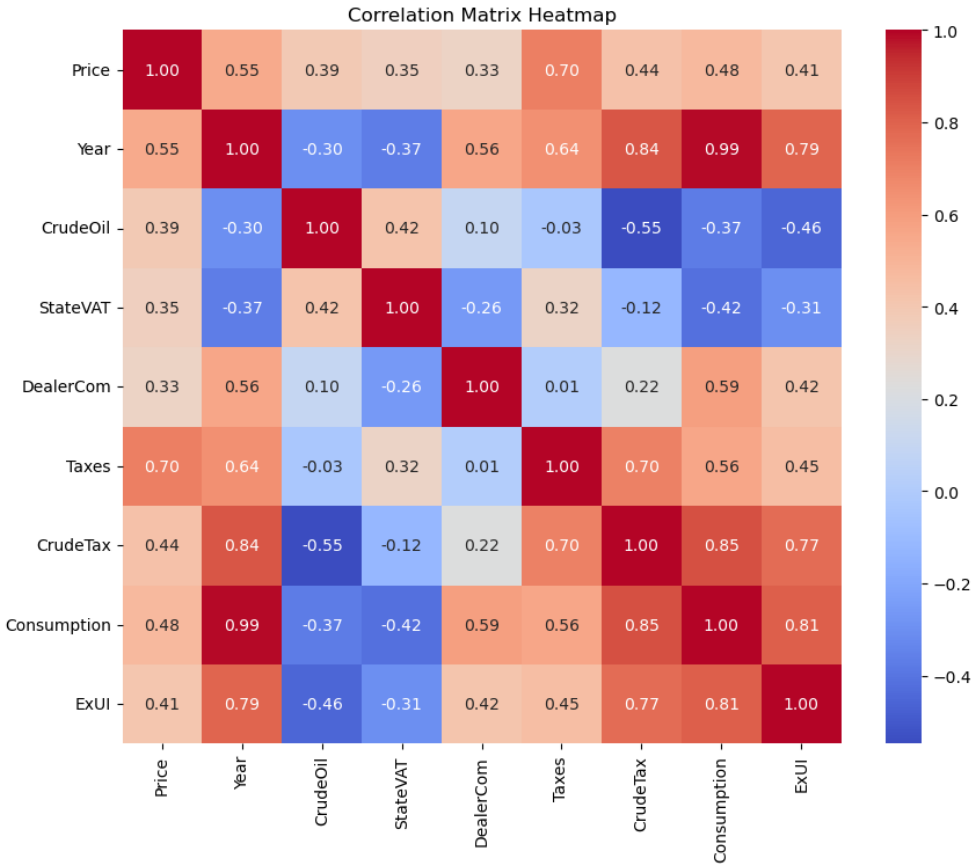
System Architecture



Implementation

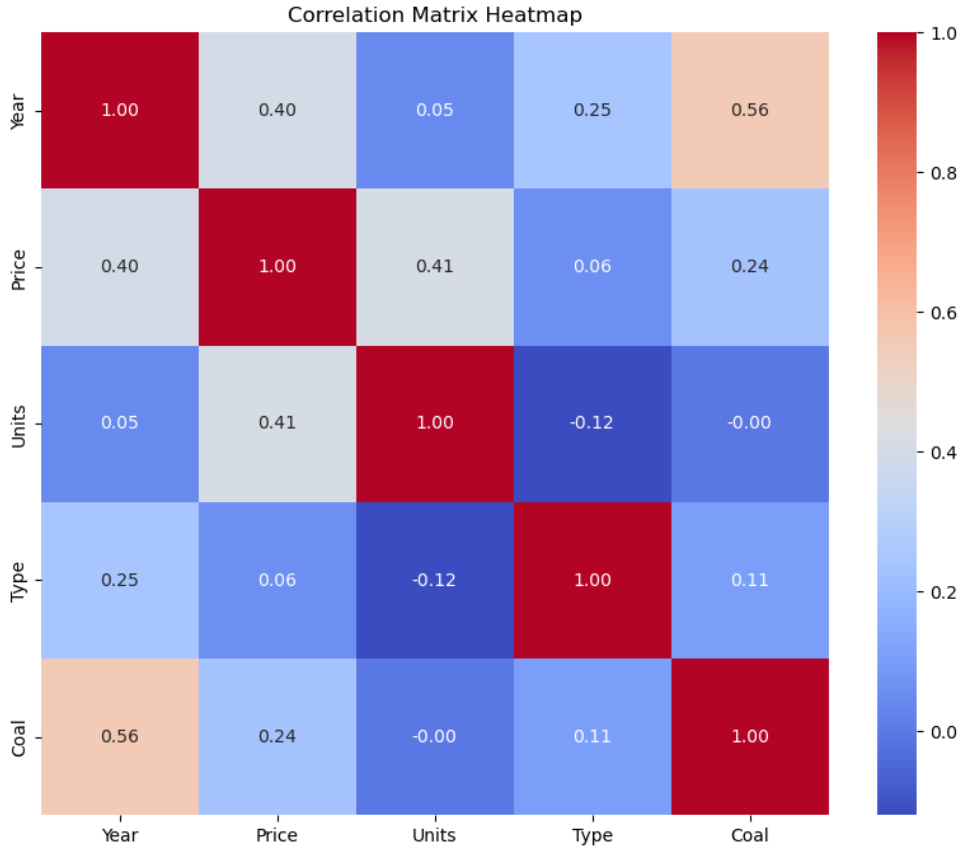
- A correlation matrix is a useful statistical tool for quantifying and visualizing relationships between variables in a dataset, revealing patterns and dependencies.
- These matrices are foundational for feature selection and model building, identifying important features with strong positive or negative correlations to the target variable. Correlation coefficients range from -1 to 1.
- The visualizations of testing data and predictions aid in communicating model accuracy, enabling stakeholders to assess the reliability of predictions and make informed decisions in commodity markets.

Implementation



Correlation matrix of petrol dataset

Implementation



Correlation matrix of electricity dataset

Implementation

- Our machine learning and data analysis work is within the Jupyter Notebook environment, we relied on a carefully chosen set of Python libraries.
- Pandas was crucial for data handling and organization, while NumPy supported numerical operations. Scikit-Learn was indispensable for machine learning tasks, with XGBoost serving as our primary algorithm. Matplotlib and Seaborn were used for data visualization.
- With pandas, we can read and create structured DataFrames from different file formats, enabling efficient data manipulation and transformation. Using panda's file reader and writer methods for handling data in various formats, ensuring consistency for analysis.
- We merge and integrate datasets to create a comprehensive dataset containing necessary features and target variables, and then export it in CSV format for model training and testing.

Results

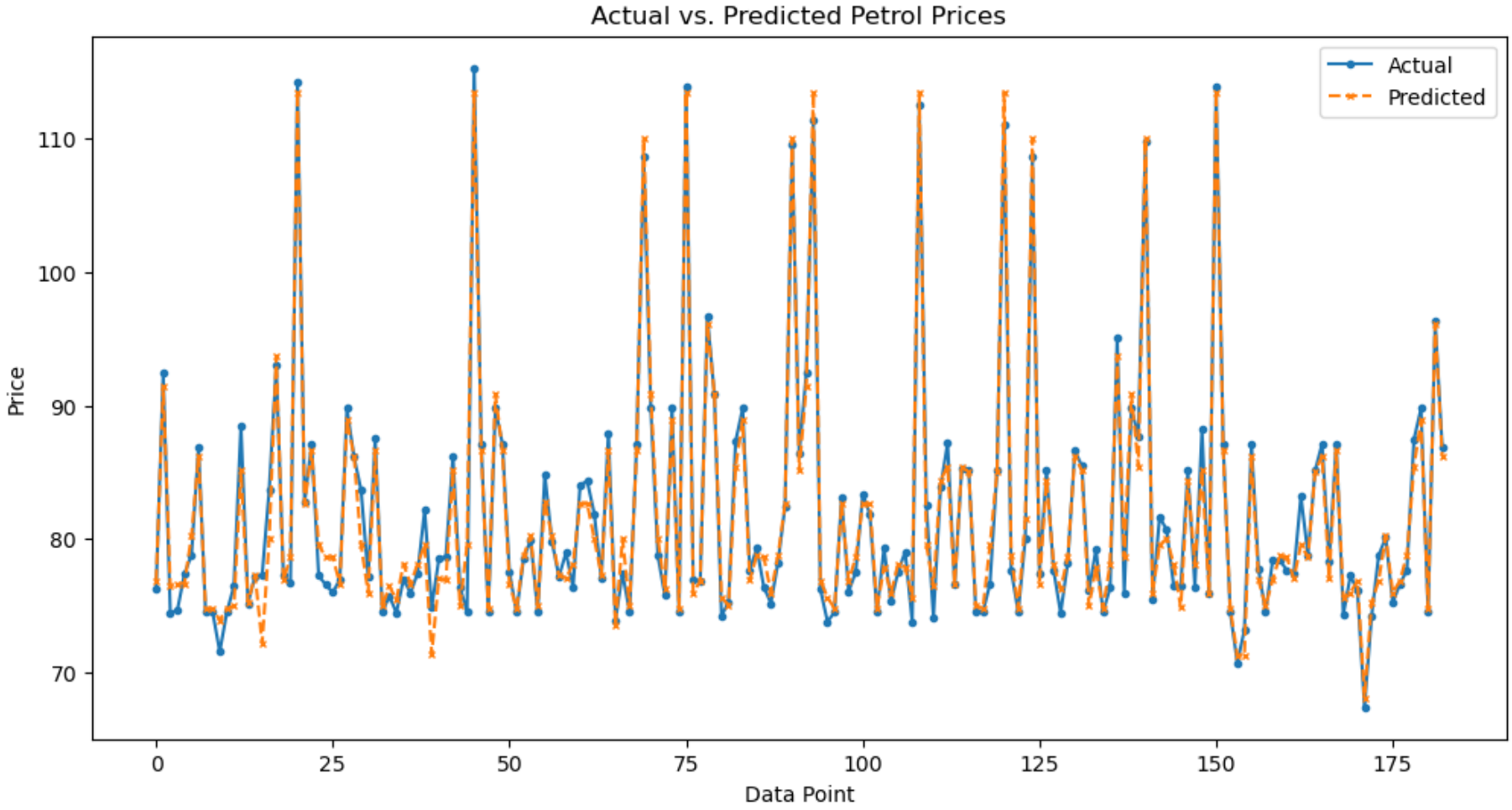
- The proposed model has experimented with 5 years of historical data for fuel price prediction in Andhra Pradesh, India. The data is set to have 80% as a training dataset and 20% as a testing dataset. The observed accuracy is above 95% on testing data.
- A similar procedure is applied to the electricity price prediction model. The observed accuracy is above 90% on testing data.
- The Petrol model demonstrates high accuracy with a low Mean Squared Error (MSE) of 2.0592, indicating close alignment between predictions and actual prices.
- Similarly, the Electricity model exhibits accuracy, boasting a low MSE of 0.2919, signifying precise predictions that closely match actual prices.

Results

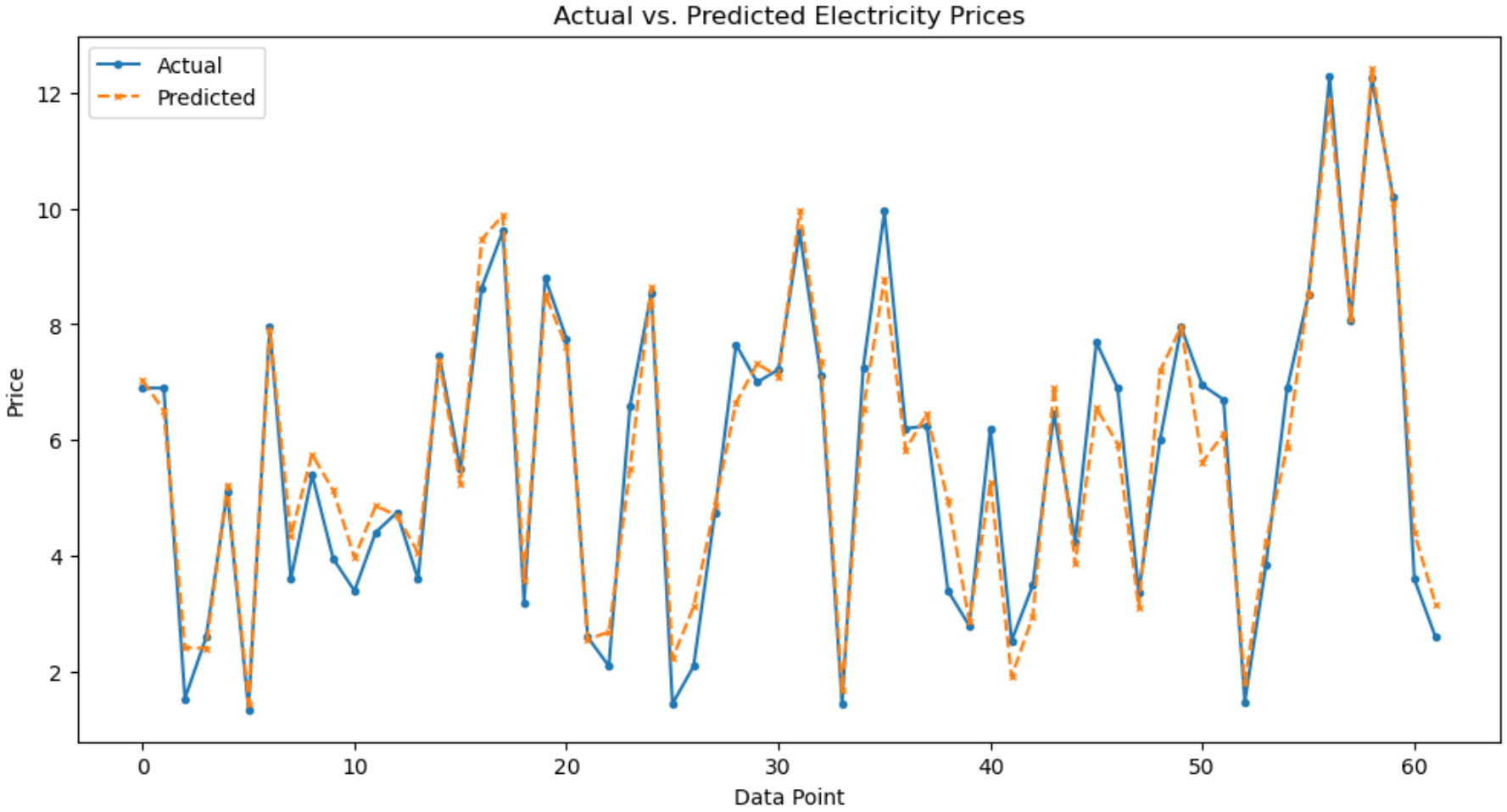
	Petrol	Electricity
Best Perms	[n_estimators=600, learning_rate=0.02, max_depth=3, random_state=42]	[n_estimators=500, learning_rate=0.02, max_depth=7, random_state=42]
Best R-Squared score	0.976514340012527	0.960227743127287
mean_square_error	2.05928441628387	0.291912879718121
y_test (first five)	[76.31, 92.42, 74.46, 74.65, 77.37]	[6.90, 6.90, 1.54, 2.60, 5.11]
y_pred	[76.844124, 91.427574, 76.45471 76.65402, 76.65402]	[6.923163, 6.564761, 2.5438266 2.332338, 5.305329]

Parameter Optimization results

Results


Actual vs. Predicted Petrol Prices
Actual
Predicted
Price
Data Point
110
100
90
80
70
0
25
50
75
100
125
150
175

Results


Actual vs. Predicted Electricity Prices
Actual
Predicted
Price
Data Point
2
4
6
8
10
12
0
10
20
30
40
50
60

Conclusion

- The study focused on predicting petrol and electricity prices in Andhra Pradesh using the XGBoost algorithm.
- Comprehensive analysis, including data collection, preprocessing, and feature engineering, was conducted.
- XGBoost models demonstrated high predictive accuracy with R-squared scores of 0.9765 for petrol and 0.9459 for electricity.
- Low Mean Absolute Error and Root Mean Squared Error values underscored the precision of the predictions.
- The research has practical implications for stakeholders, aiding in energy management, economic planning, and investment decisions in Andhra Pradesh's commodities market.
- Future enhancements involve integrating additional data sources and exploring alternative modeling techniques for further refinement.

Future Enhancement

- The research has broad implications for various industries, providing stakeholders with valuable insights for informed decision-making during price fluctuations.
- Current findings focus on predicting petrol and electricity prices in Andhra Pradesh, offering significant insights into critical commodities.
- Future enhancements can improve prediction accuracy by integrating alternative data sources like satellite imagery, social media sentiment analysis, and weather data.
- Exploring advanced machine-learning techniques beyond XGBoost is identified as a promising avenue for further improvement.
- Consideration for extending the project's scope to include other essential commodities, such as metals, natural gas, and agricultural items, is highlighted.
- Geographical expansion beyond Andhra Pradesh is suggested to provide a more comprehensive understanding of the commodities market.

Thank you