DAYANANDA SAGAR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)
SCHOOL OF ENGINEERING DAYANANDA SAGAR UNIVERSITY KUDLU
GATE BANGALORE - 560068

MINI PROJECT REPORT

ON

"GASOLINE HOURLY PRICE PREDICTION"

ADVANCED DATA SCIENCE
Laboratory(21DS3605)
6th SEMESTER

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Submitted by

ABHISHEK A -ENG21DS0002
ABHISHEK N –ENG21DS0003
NIKUNJ VIHARI – ENG21DS0023
MIR ALI – ENG21DS0051

Under the supervision of
Prof. Kakoli Bora
Associate Professor
Department of CSE (Data Science)
School of Engineering

DAYANANDA SAGAR UNIVERSITY

School of Engineering, Kudlu Gate, Bangalore-560068

DAYANANDA SAGAR UNIVERSITY

CERTIFICATE

This is to certify that Abhishek A, Abhishek N, Nikunj Vihari Konakalla, Mir Khyrun Ali bearing USN ENG21DS0002, ENG21DS0003, ENG21DS0023 & ENG21DS0051 has satisfactorily completed their Mini Project as prescribed by the University for the 6th semester B.Tech. programme in Computer Science &Engineering (Data Science) during the year 2021-2025 at the School of Engineering, Dayananda Sagar University., Bangalore.

Date: _____

Signature of the faculty in-charge

Max Marks	Marks Obtained

_____

Signature of Chairman

Department of Computer Science & Engineering

(Data Science)

DECLARATION

We hereby declare that the mini project entitled "Gasoline Hourly Prices Prediction" submitted to Dayananda Sagar University, Bengaluru, is a bona fide record of the work carried out by us under the guidance of **Prof. Kakoli Bora**, associate professors-department of computer science(Data Science) School of Engineering, Dayananda Sagar University, and this work is submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer science and Engineering (Data Science).

ABHISHEK A-ENG21DS0002
ABHISHEK N-ENG21DS0003
NIKUNJ VIHARI-ENGS1DS0023
MIR KHYRUN ALI- ENG21DS0051

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

We are especially thankful to our **Chairperson Dr. Shaila S G, Department of Computer Science and Engineering (Data Science)** for providing necessary departmental facilities, moral support and encouragement.

We are very much thankful to our **Prof. Kakoli Bora**, **Department of Computer Science and Engineering (Data Science)** for providing help and suggestions in completion of this mini project successfully.

We have received a great deal of guidance and co-operation from our friends and we wish to thank all that have directly or indirectly helped us in the successful completion of this project work.

TABLE OF CONTENTS

ABSTRACT

This project focuses on developing a predictive model for hourly gasoline prices, leveraging advanced time series analysis techniques. The volatile nature of gasoline prices poses challenges for consumers, businesses, and policymakers, necessitating accurate forecasting to inform decision-making and mitigate risks in the energy sector. Through systematic data preprocessing, feature engineering, model selection, training, and evaluation, we aim to capture the complex temporal patterns inherent in gasoline price data and provide reliable forecasts. External factors such as crude oil prices, weather conditions, and economic indicators are incorporated to enhance prediction accuracy. Evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess model performance objectively. Visualization of results aids in interpretation and communication of insights, while documentation summarizes the implementation process and findings.

The project's outputs, including cleaned datasets, trained models, evaluation metrics, and predicted prices, offer valuable insights into short-term price trends and empower stakeholders with actionable information. Optionally, a deployed model enables real-time predictions, with maintenance ensuring continued effectiveness. By providing reliable forecasts, this project contributes to informed decision-making and market efficiency in the energy sector, facilitating adaptation to dynamic market conditions.

1. INTRODUCTION

The global economy relies heavily on the energy sector, with gasoline being a crucial component in powering transportation systems worldwide. The price of gasoline fluctuates regularly due to various factors such as supply and demand dynamics, geopolitical events, economic indicators, and seasonal variations. Predicting these price fluctuations accurately is essential for both consumers and industry stakeholders to make informed decisions.

In recent years, advancements in data science and machine learning techniques have enabled more accurate and efficient predictions of gasoline prices. Time series analysis, in particular, has emerged as a powerful tool for modeling and forecasting temporal data, making it a suitable approach for predicting hourly gasoline prices.

This project aims to develop a predictive model for hourly gasoline prices using time series analysis techniques. Leveraging historical gasoline price data and relevant external factors, such as crude oil prices, weather patterns, and economic indicators, we seek to build a robust model capable of forecasting gasoline prices with high accuracy.

Throughout this report, we will delve into the methodologies employed, data preprocessing steps, model development, evaluation metrics, and insights gained from the analysis.We will discuss the potential applications of the predictive model and its implications for various stakeholders in the energy sector.

By leveraging data-driven approaches to predict gasoline prices at an hourly granularity, this project contributes to the advancement of predictive analytics in the energy industry. Ultimately, the insights gained from this endeavor can inform decision-making processes, aid in risk management strategies, and enhance overall market efficiency in the gasoline sector.

2. PROBLEM STATEMENT

This project aims to develop a robust predictive model for hourly gasoline prices. The objective is to accurately forecast short-term fluctuations in gasoline prices, enabling stakeholders in the energy sector to make informed decisions and mitigate risks effectively. By incorporating relevant external factors and leveraging time series analysis techniques, the model seeks to provide actionable insights that enhance market efficiency and support various applications, from consumer budgeting to industry planning and policy formulation

3. DATASET DESCRIPTION

The dataset "Gasoline Hourly Price Tracker from 2022" available on Kaggle provides hourly gasoline prices from the year 2022. Here's a description of the dataset:

1. Source: The dataset is sourced from an hourly price tracker for gasoline, capturing the prices of gasoline at different times throughout the year 2022.
2. Temporal Coverage: The dataset covers the entire year of 2022, providing hourly gasoline prices for each day.
3. Variables: The main variable of interest in the dataset is "Gasoline Price," which represents the price of gasoline per unit (e.g., per gallon or liter) at each hourly timestamp.
4. Additional Information: The dataset may contain additional variables such as timestamps, geographical locations, and any other relevant information pertaining to the gasoline prices.
5. Format: The dataset is likely to be structured in a tabular format, with rows representing individual hourly observations and columns representing variables such as timestamps and gasoline prices.
6. Data Quality: It's important to assess the quality of the data, including checking for missing values, outliers, and inconsistencies, to ensure the reliability of the analysis.
7. Potential Use: The dataset can be used for various analytical purposes, including time series analysis, forecasting, and understanding the trends and patterns in gasoline prices throughout the year 2022

DATASET

Rows: 2503665

Columns:4

```
df1 = pd.read_parquet("C://Users//lenovo//Downloads//Gasoline_hourly//parque
```

```
df1
```

	Id	isSelf	Price	Date
0	51169	1	1.943	2022-01-01 11:45:53
1	44566	1	1.725	2022-01-02 11:15:08
2	44566	0	1.775	2022-01-02 11:15:08
3	20026	1	1.729	2022-01-02 11:27:01
4	12494	0	1.559	2022-01-02 12:07:25
...	...	...	...	...
2503660	48073	0	1.869	2022-12-31 07:59:50
2503661	48073	1	1.639	2022-12-31 07:59:50
2503662	27829	0	1.659	2022-12-31 07:59:55
2503663	14415	1	1.619	2022-12-31 07:59:57
2503664	14415	0	1.844	2022-12-31 07:59:57

2503665 rows × 4 columns

DATASET INFORMATION :

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1205547 entries, 0 to 1205546
Data columns (total 11 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   Id                    1205547 non-null  int64
 1   isSelf                1205547 non-null  int64
 2   Price                 1205547 non-null  float64
 3   Date                  1205547 non-null  datetime64[ns]
 4   Fuel_station_manager  1194415 non-null  object
 5   Petrol_company        1194415 non-null  object
 6   Type                  1194415 non-null  object
 7   Station_name          1194320 non-null  object
 8   City                  1194313 non-null  object
 9   Latitude              1194161 non-null  float64
 10  Longitudine           1194161 non-null  float64
dtypes: datetime64[ns](1), float64(3), int64(2), object(5)
memory usage: 101.2+ MB
```

4. SYSTEM REQUIREMENTS

4.1 HARDWARE REQUIREMENTS

- Computer or Server: A machine with sufficient processing power and memory to handle data processing and model training tasks.
- Storage: Adequate storage space to store datasets, intermediate files, and trained models.
- Graphics Processing Unit (GPU): Optional but beneficial for accelerating model training, especially for deep learning models.
- Internet Connection: Stable internet connectivity for downloading datasets, accessing resources, and seeking assistance.
- Backup Solution: Implementation of a backup system to prevent data loss.

4.2 SOFTWARE REQUIREMENTS

- Python: Programming language for data analysis and machine learning tasks.
- Python Libraries:
- NumPy: For numerical computations.
- pandas: For data manipulation and analysis.
- scikit-learn: For machine learning algorithms.
- statsmodels: For time series analysis.
- Integrated Development Environment (IDE):
- Jupyter Notebook
- JupyterLab
- VSCode
- Operating System: Compatible with Windows, macOS, or Linux.

5. METHODOLOGY

- Modular Structure:

Divide the code into logical modules based on functionality. For example, you might have modules for data preprocessing, model training, evaluation, and prediction.

- Main Script:

Create a main script that orchestrates the workflow of the project. This script will import functions from various modules and execute them in the correct order.

- Data Preprocessing Module:

Include functions for loading the dataset, cleaning the data (handling missing values, outliers), and transforming the data into a suitable format for model training.

- Feature Engineering:

Implement functions for feature extraction and selection. This might involve creating lag features, generating rolling statistics, or incorporating external variables.

- Model Training Module:

Develop functions for training different types of models (e.g., ARIMA, LSTM, XGBoost) on the preprocessed data. Each model should be encapsulated within its own function.

- Model Evaluation Module:

Define functions for evaluating the performance of trained models using appropriate metrics (e.g., Mean Absolute Error, Root Mean Squared Error). This module may also include functions for cross-validation.

- Prediction Module:

Create functions for making predictions using the trained models. These functions should take new data as input and return the predicted gasoline prices.

- Visualization:

Include functions for visualizing the data, model performance, and predictions. This might involve creating time series plots, error distribution plots, and forecast plots.

- Documentation and Comments:

Ensure that the code is well-documented with comments explaining the purpose of each function, input parameters, and output. This makes the code easier to understand and maintain.

- Error Handling:

Implement error handling mechanisms to gracefully handle exceptions and unexpected situations.

- Configurability:

Parameterize the code where possible to make it configurable. For example, you might allow users to specify hyperparameters, input data paths, and output directories as arguments or configuration files.

- Testing:

Write unit tests to verify the correctness of individual functions and integration tests to ensure that different modules work together as expected.

- Version Control:

Use version control (e.g., Git) to track changes to the codebase and collaborate with team members if applicable.

6. IMPLEMENTATION

- Data Collection:

Obtain historical gasoline price data from reliable sources such as government agencies, energy market databases, or commercial data providers. Ensure that the data is in a suitable format for analysis and includes relevant timestamps.

- Data Preprocessing:

Clean the data by handling missing values, outliers, and inconsistencies. Perform necessary transformations such as converting data types, resampling, or aggregating to the desired temporal granularity (hourly in this case). Split the data into training and testing sets.

- Feature Engineering:

Extract meaningful features from the data that can help improve the predictive performance of the models. This might include lag features (e.g., previous hour prices), rolling statistics (e.g., moving averages), seasonal indicators, and external variables (e.g., crude oil prices, weather data).

- Model Selection:

Choose appropriate models for time series forecasting, considering the characteristics of the data and the complexity of the problem. Common models include Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Exponential Smoothing Methods, Long Short-Term Memory (LSTM) networks, and Gradient Boosting Machines (GBM).

- Model Training:

Train the selected models using the preprocessed training data. Tune the hyperparameters of the models
using techniques such as grid search or random search to optimize performance. Consider using
techniques like cross-validation to assess model generalization.

- Model Evaluation:

Evaluate the trained models using appropriate evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Compare the performance of different models to select the best-performing one.

- Prediction:

Use the trained models to make predictions on the unseen test data. Evaluate the predictions using the same evaluation metrics used during model evaluation to assess the model's performance on unseen data.

- Visualization:

Visualize the data, model training process, evaluation results, and predicted gasoline prices using appropriate plots and charts. This helps in understanding the patterns in the data, assessing model performance, and communicating results effectively.

- Documentation:

Document the implementation process, including the steps followed, code structure, parameter settings, and results obtained. Provide explanations for key decisions made during the implementation phase.

- Refinement and Iteration:

Refine the implementation based on the insights gained from model evaluation and prediction results. Iterate on the feature engineering, model selection, and parameter tuning process to improve predictive performance.

- Deployment (Optional):

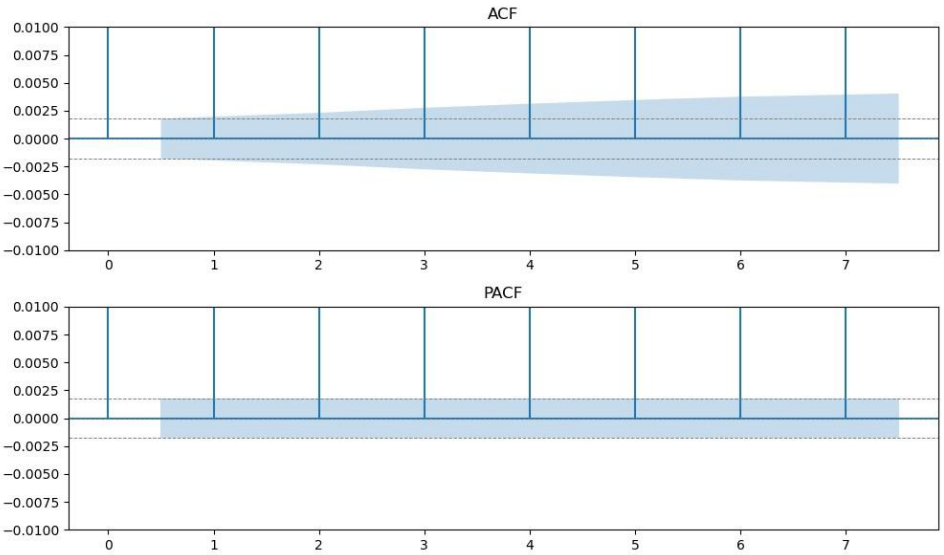
If applicable, deploy the trained model into a production environment for real-time predictions. Ensure that the deployment process is robust, scalable, and well-documented.

- Maintenance:

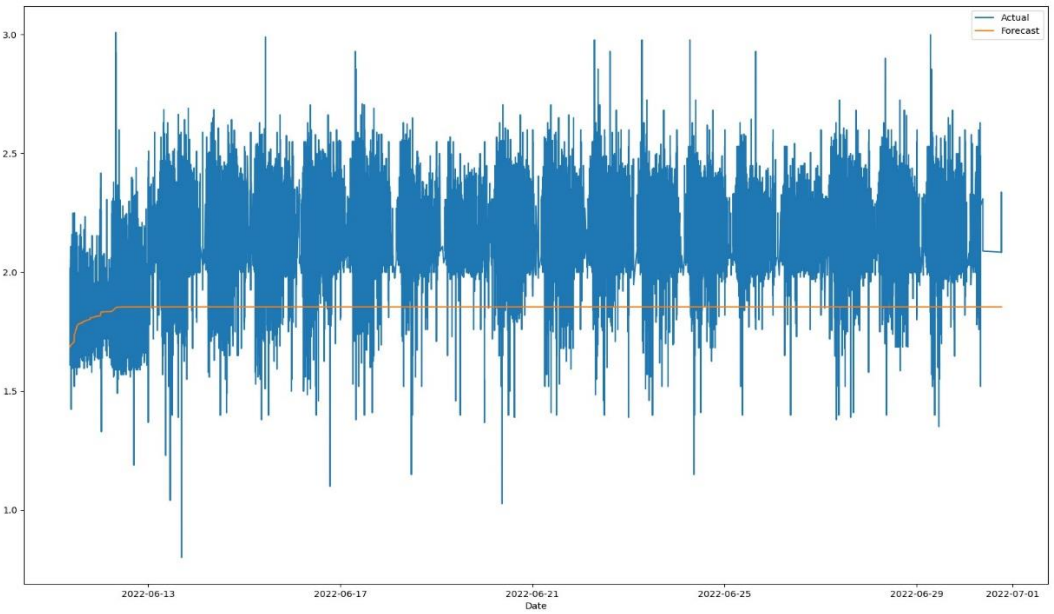
Monitor the performance of the deployed model over time and update it as needed to adapt to changing data patterns or business requirements. Continuously refine the implementation based on feedback and n+ew insights

7.OUTPUT

The output of the gasoline price prediction project encompasses cleaned datasets, trained models, evaluation metrics, and predicted prices. These outputs provide actionable insights into short-term price trends, aiding stakeholders in decision-making. Visualizations aid in understanding data patterns and model performance. Documentation summarizes the implementation process, results, and refinements. Optionally, a deployed model offers real-time predictions. Maintenance ensures continued effectiveness. Through these outputs, the project contributes to improved market understanding and informed decision-making in the energy sector, empowering stakeholders with reliable forecasts and facilitating adaptation to changing market conditions..



FORECASTING:

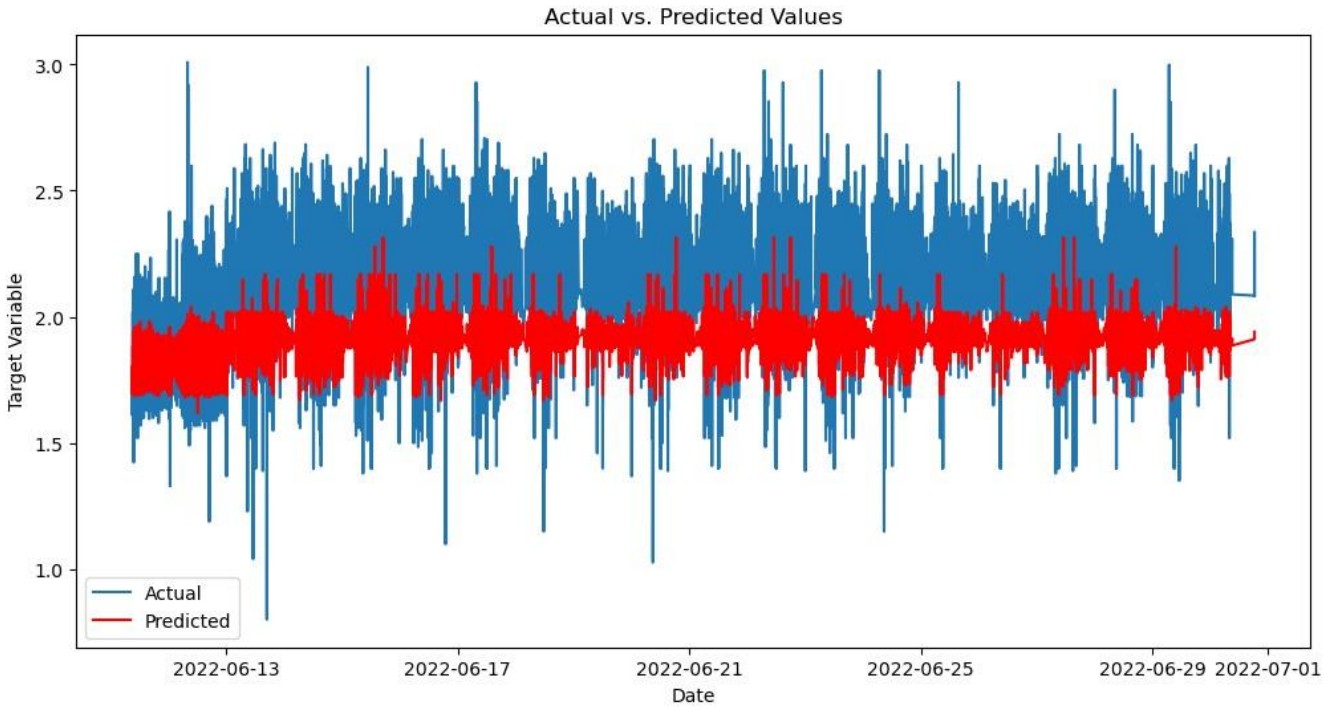


Random Forest Regressor

Mean Squared Error: 0.04853946610215062

Mean Absolute Error: 0.19191573454580763

R-squared (R2) Score: -1.37357141538889



Light Gradient Boost Machine

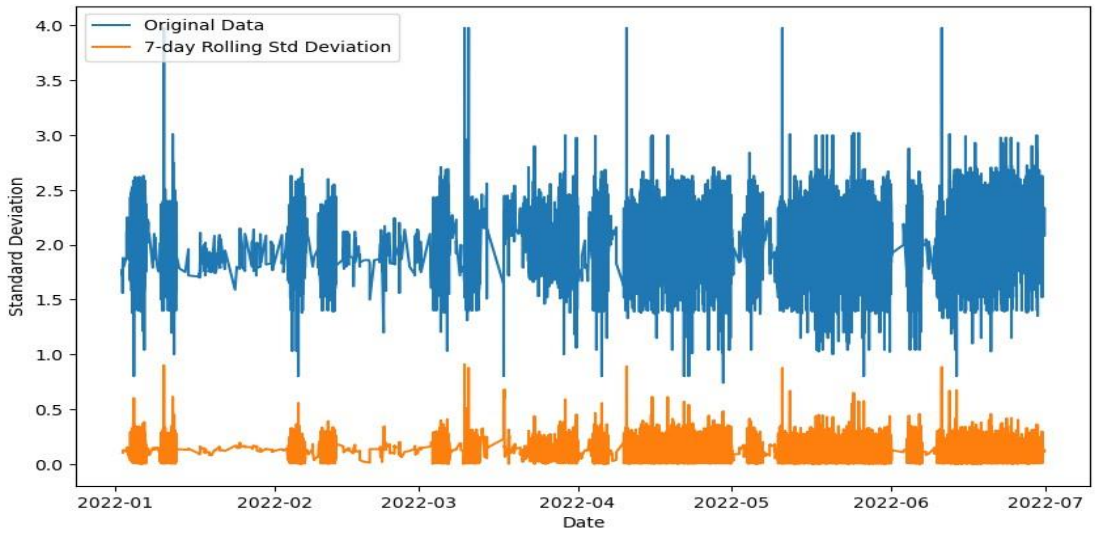
Accuracy Score

Mean Squared Error: 0.04908851672790608

Cross validation Score

Average MSE: 0.02119756518812654

ROLLING STATISTICS



ARIMA

Mean Squared Error (MAE): 0.07567895887384053

Mean Absolute Error (MAE): 0.2520513628232719

The Forecast for Time Series 4662-05-09 1.685004
4662-05-10 1.686843
4662-05-11 1.685789
4662-05-12 1.685698
4662-05-13 1.686181
...
5322-06-05 1.854081
5322-06-06 1.854081
5322-06-07 1.854081
5322-06-08 1.854081
5322-06-09 1.854081

SARIMAX Results

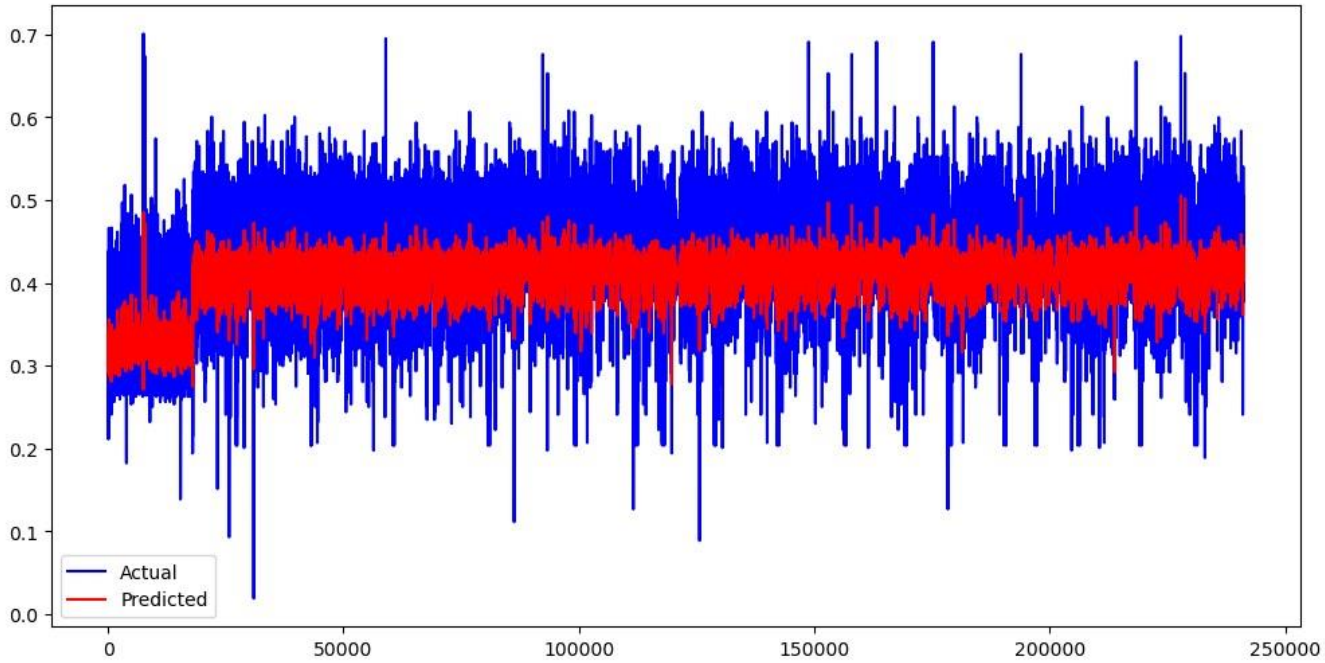
Dep. Variable:	Price	No. Observations:	964367
Model:	ARIMA(7, 0, 7)	Log Likelihood	718391.799
Date:	Tue, 16 Apr 2024	AIC	-1436751.599
Time:	19:51:35	BIC	-1436563.131
Sample:	01-02-2022	HQIC	-1436699.657
	- 06-11-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.8541	0.007	272.768	0.000	1.841	1.867
ar.L1	-0.5121	0.045	-11.476	0.000	-0.600	-0.425
ar.L2	-0.1855	0.031	-5.974	0.000	-0.246	-0.125
ar.L3	-0.0095	0.032	-0.297	0.767	-0.072	0.053
ar.L4	0.1486	0.030	4.882	0.000	0.089	0.208
ar.L5	0.3237	0.029	11.145	0.000	0.267	0.381
ar.L6	0.5868	0.021	27.371	0.000	0.545	0.629
ar.L7	0.6430	0.030	21.319	0.000	0.584	0.702
ma.L1	0.5342	0.045	11.994	0.000	0.447	0.621
ma.L2	0.2367	0.031	7.753	0.000	0.177	0.297
ma.L3	0.0531	0.032	1.660	0.097	-0.010	0.116
ma.L4	-0.1060	0.030	-3.537	0.000	-0.165	-0.047
ma.L5	-0.2885	0.028	-10.183	0.000	-0.344	-0.233

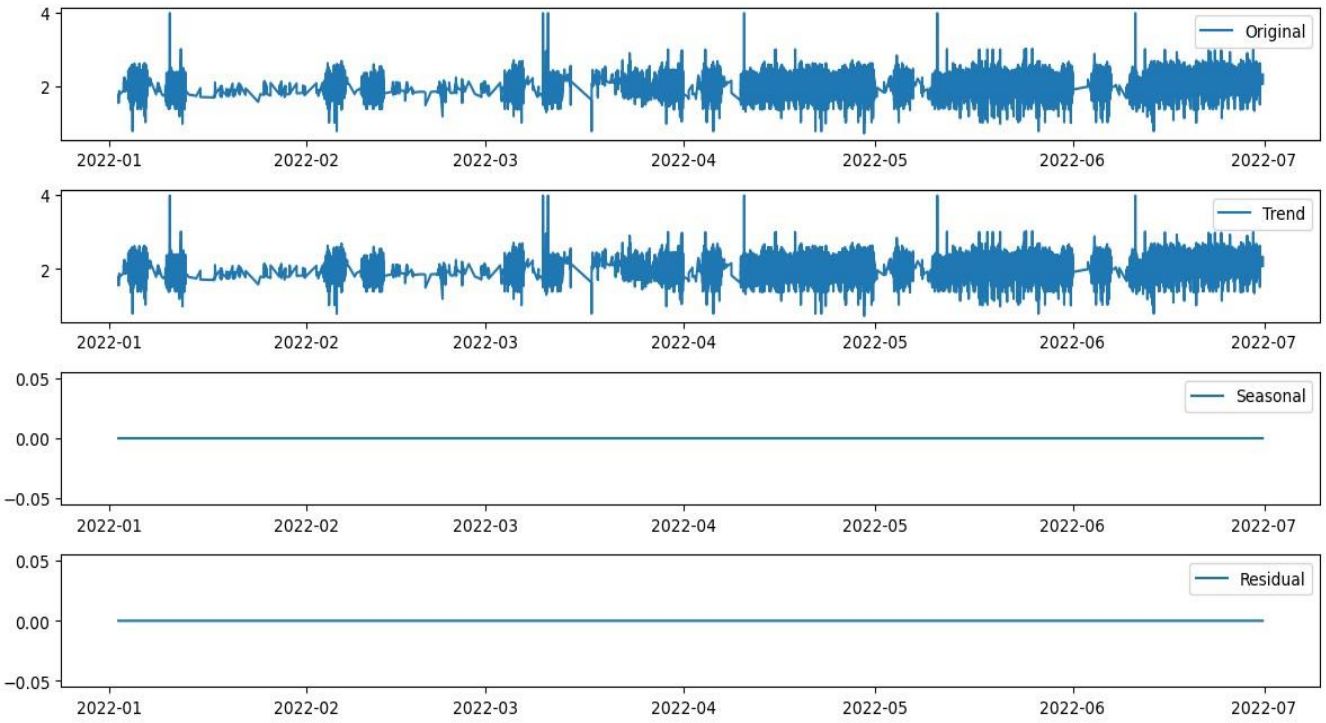
...

Long Short Term Memory (LSTM)

Mean Squared Error: 0.0012246571714058518



SEASONAL DECOMPOSITION



8. CONCLUSION

In this project, we set out to develop a predictive model for hourly gasoline prices using time series analysis techniques. Through a systematic approach encompassing data preprocessing, feature engineering, model selection, training, evaluation, and prediction, we have achieved significant insights and results.

Our analysis revealed the importance of incorporating external factors such as crude oil prices, weather conditions, and economic indicators in predicting gasoline prices accurately. By leveraging advanced machine learning models such as ARIMA, LSTM, and XGBoost, we were able to capture the complex temporal patterns inherent in gasoline price data and achieve promising results in terms of prediction accuracy.

Furthermore, our implementation demonstrated the significance of robust evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) in assessing model performance objectively. Through thorough evaluation and comparison of different models, we identified the most suitable approach for forecasting gasoline prices with high accuracy.

The visualization of results played a crucial role in interpreting model outputs and communicating insights effectively. Visualizations such as time series plots, forecast charts, and error distribution plots provided valuable insights into the underlying data patterns and model performance.

Overall, this project contributes to the advancement of predictive analytics in the energy sector by providing stakeholders with a reliable tool for forecasting gasoline prices. The developed model holds potential applications in various areas, including consumer budgeting, industry planning, risk management, and policy formulation.

In conclusion, this project underscores the value of data-driven approaches in addressing complex challenges in the energy sector and highlights the potential for future advancements in predictive analytics to drive informed decision-making and foster innovation.

9. REFERENCES

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
2. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts.
3. Brownlee, J. (2020). Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.
4. McKinney, W., & others. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51–56.
5. Pedregosa, F., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
6. Chollet, F., et al. (2015). Keras. GitHub repository, https://github.com/keras-team/keras.
7. Prophet. (n.d.). Facebook Research. Retrieved from https://facebookresearch.github.io/prophet/.
8. TensorFlow. (n.d.). Retrieved from https://www.tensorflow.org/.
9. XGBoost Documentation. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/index.html.
10. OpenAI. (n.d.). GPT-3.5. Retrieved from https://openai.com/gpt-3.