Daily Gas Consumption Prediction in Ireland Using Time Series Analysis and Machine Learning

Sharjeel Nawaz (23389541)
Data Analytics for Artificial Intelligence – H9DAI
MSCAI1B
School of Computing
National College of Ireland

1 Background Research

Predicting gas consumption is essential for managing energy distribution, reducing waste, and optimizing operational efficiency. Accurate predictions of daily gas usage enable energy providers to allocate resources effectively, anticipate demand surges, and enhance customer service. Gas consumption patterns are influenced by various factors, including temperature, seasonal trends, and consumption behavior across different user groups. In the context of Ireland, gas consumption is highly seasonal, with higher demand during colder months due to heating needs. Weather conditions, such as temperature and rainfall, significantly shape daily consumption patterns. Understanding these influences is crucial for developing predictive models that can accurately forecast future consumption.

This project focuses on predicting daily gas consumption in Ireland using historical gas consumption data. The dataset contains several key features, including the date of the consumption record, the type of gas meter (e.g., Non-Daily Metered), and the amount of gas consumed on a specific day, measured in Gigawatt-hours (GWh). Additionally, weather data is integrated into the analysis to capture the impact of weather conditions on gas usage. Key weather variables include maximum temperature, minimum temperature, rainfall, and sunlight hours. Incorporating these weather features aims to improve the accuracy of predictions by accounting for their influence on gas demand, particularly during colder months when heating needs increase.

Time series forecasting models play a critical role in predicting gas consumption by analyzing temporal data points. Models like ARIMA (Autoregressive Integrated Moving Average) and its seasonal variant SARIMA are commonly used for forecasting energy usage, as they capture patterns of seasonality, trends, and autocorrelations in time series data. These elements are essential for accurately predicting gas consumption. In addition to time series models, machine

learning algorithms such as Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) have been applied to energy consumption forecasting. These algorithms are particularly adept at handling complex, non-linear relationships in the data, enabling them to learn intricate patterns, such as the relationship between weather conditions and gas consumption. These machine learning methods are valuable for identifying and modeling non-linear dependencies between features and target variables (Hyndman, 2018).

Feature engineering is a key process in improving the accuracy of predictive models. Temporal features such as the day of the week, month, and week number are critical for capturing seasonal and weekly trends in gas consumption. Furthermore, weather-related features help account for the influence of temperature fluctuations and weather patterns on gas demand. By combining domain knowledge with these features, the predictive model becomes more robust and capable of generating accurate forecasts (Liaw, 2002).

2 Data Analytics

In this project, I focused on predicting daily gas consumption in Ireland. To achieve this, I conducted an in-depth data analysis and applied all necessary preprocessing steps to ensure the models were both accurate and robust. The analysis provided valuable insights into the data, which helped me perform all the preprocessing effectively, and this was crucial for both regression and forecasting models used to predict future gas consumption.

The key questions addressed in the data analysis include:

- What is the gas consumption rate across different months?
- Does the gas consumption data exhibit seasonal trends, and what are these trends?
- How does gas consumption vary across the weeks within a month?
- What is the relationship between gas consumption and the days of the week?

These questions will be answered in the following sections

Analysis of Target Variable Distribution and Meter Type Selection:

After examining the features of the dataset, I conducted a deeper analysis into the distribution of the target variable. During this analysis, I discovered that each meter type has a distinct distribution, each show different trends over the months. The graph below illustrates how each meter type exhibits unique trends, highlighting the relationship between gas consumption and months.

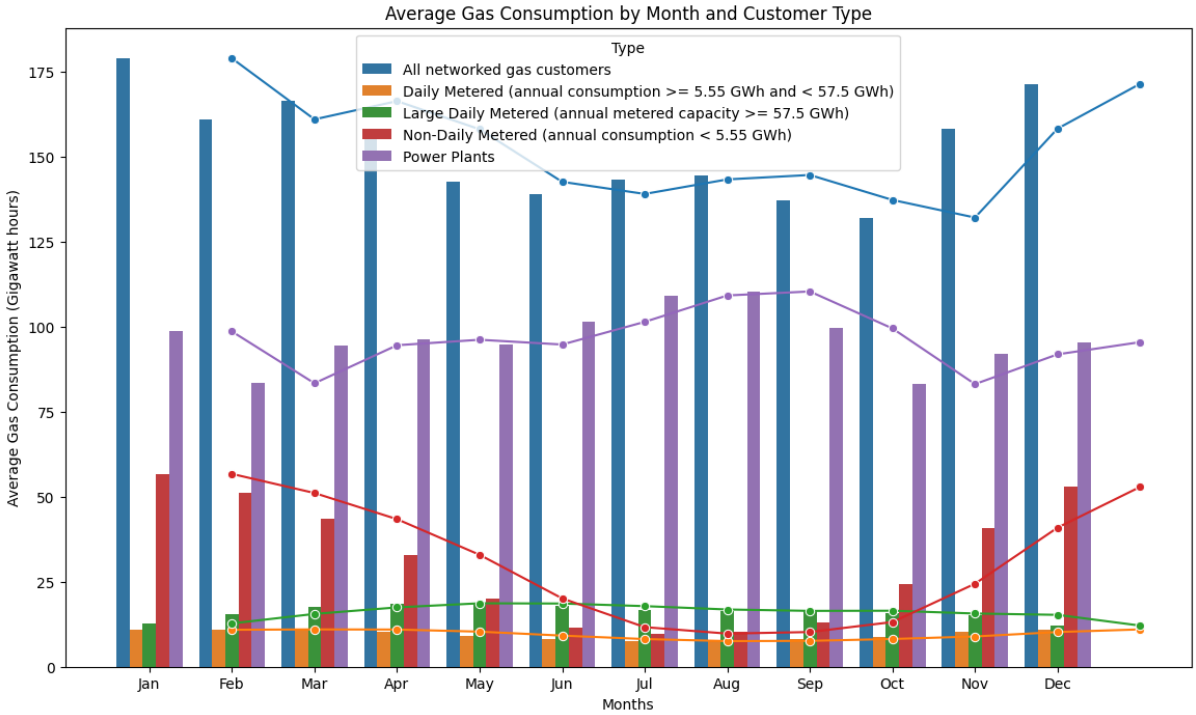


Figure 1

Focus on Specific Meter Type: Non-Daily Metered (Annual Consumption < 5.55 GWh): To narrow the focus of my analysis, I decided to work with a specific meter type for predicting gas consumption. After reviewing the data, I selected the "Non-Daily Metered (annual consumption < 5.55 GWh)" type. The distribution of gas consumption for this meter type is shown in the graph below.

Relationship Between Gas Consumption and Months:

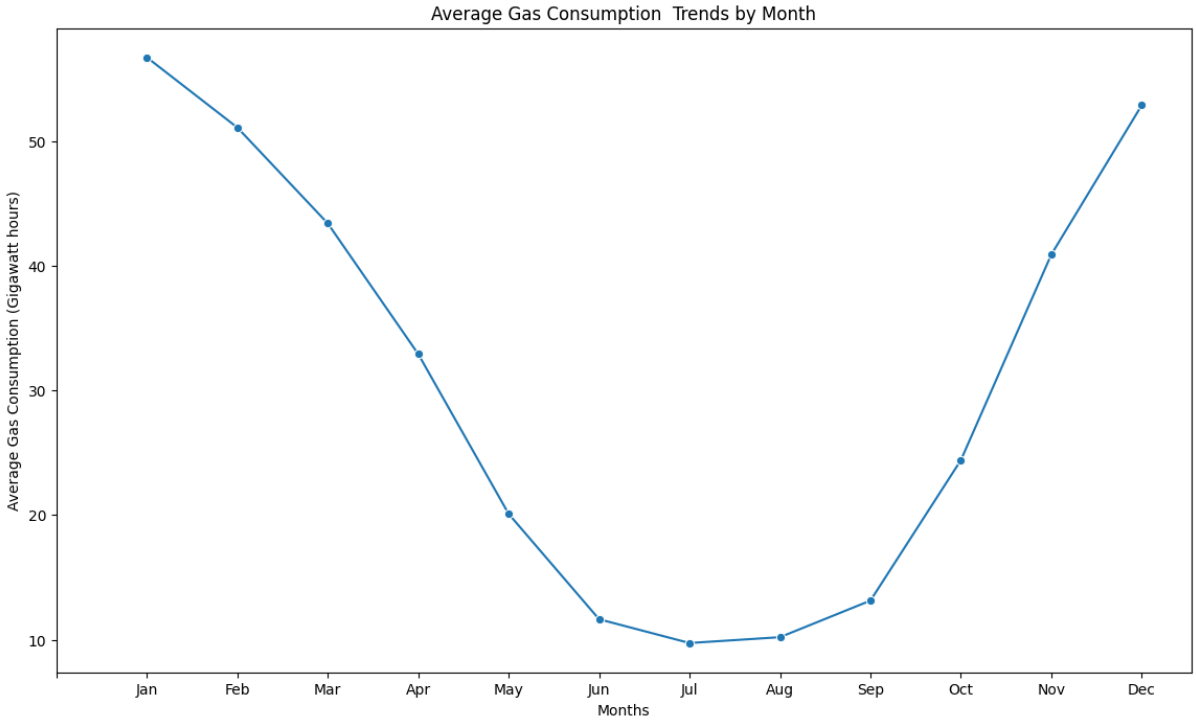


Figure 2

The graph illustrates a clear relationship between gas consumption and the months of the year. It shows that during months with lower temperatures, gas consumption is higher, while in months with warmer temperatures, consumption decreases.

Seasonal Trend in Gas Consumption:

This trend (the relationship between months and gas consumption) persists consistently across the entire time span of the dataset, as seen in the following graph.

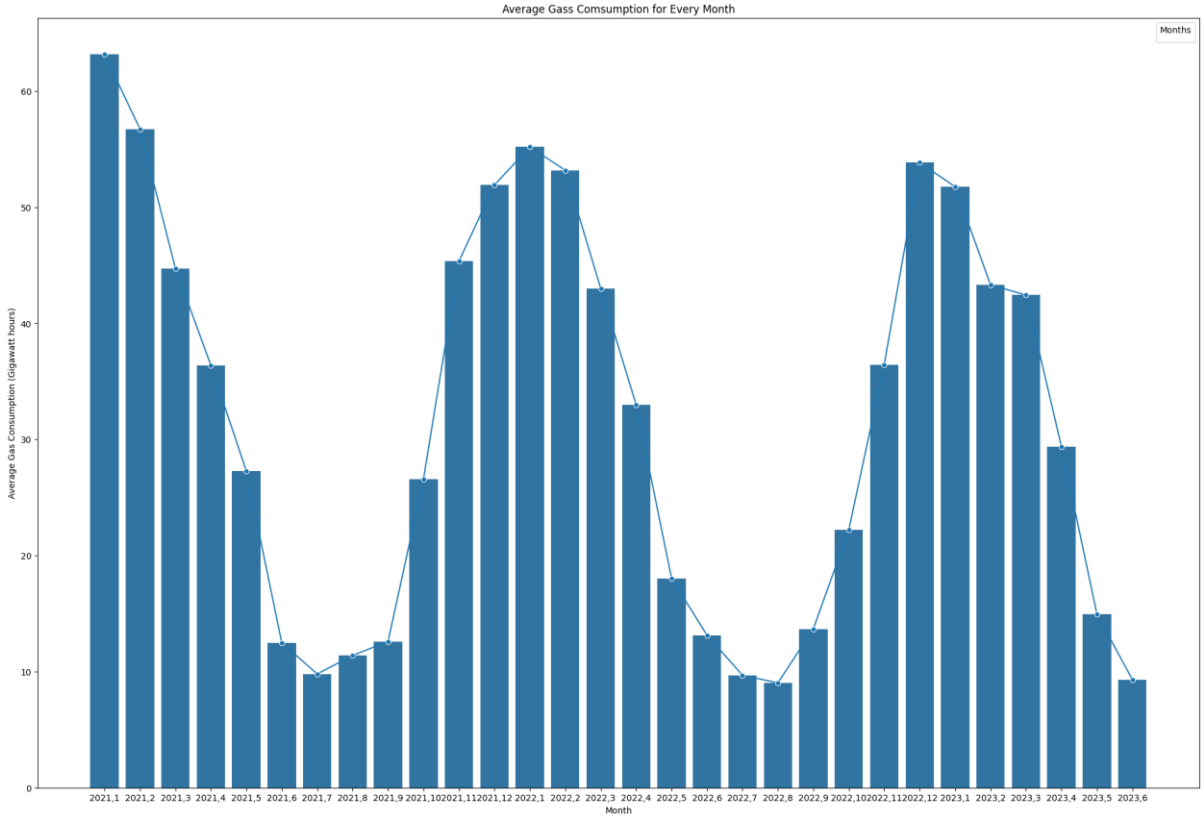


Figure 3

The graph confirms clear seasonal trends, with patterns repeating consistently over the months. Moving forward, the focus will be on analyzing trends within months and weeks to uncover finer relationships and identify useful patterns.

Weekly Trends:

To analyze trends within months, I divided the data of each month into five-week intervals and examined the relationship between gas consumption and week numbers. Upon analysis, I found no direct relationship between week numbers and gas consumption. However, since gas consumption varies over time (months), adding week numbers as a feature creates smaller time intervals within each month, which could enhance model accuracy. Weekly variations in gas consumption were observed, making this feature a valuable addition to the model.

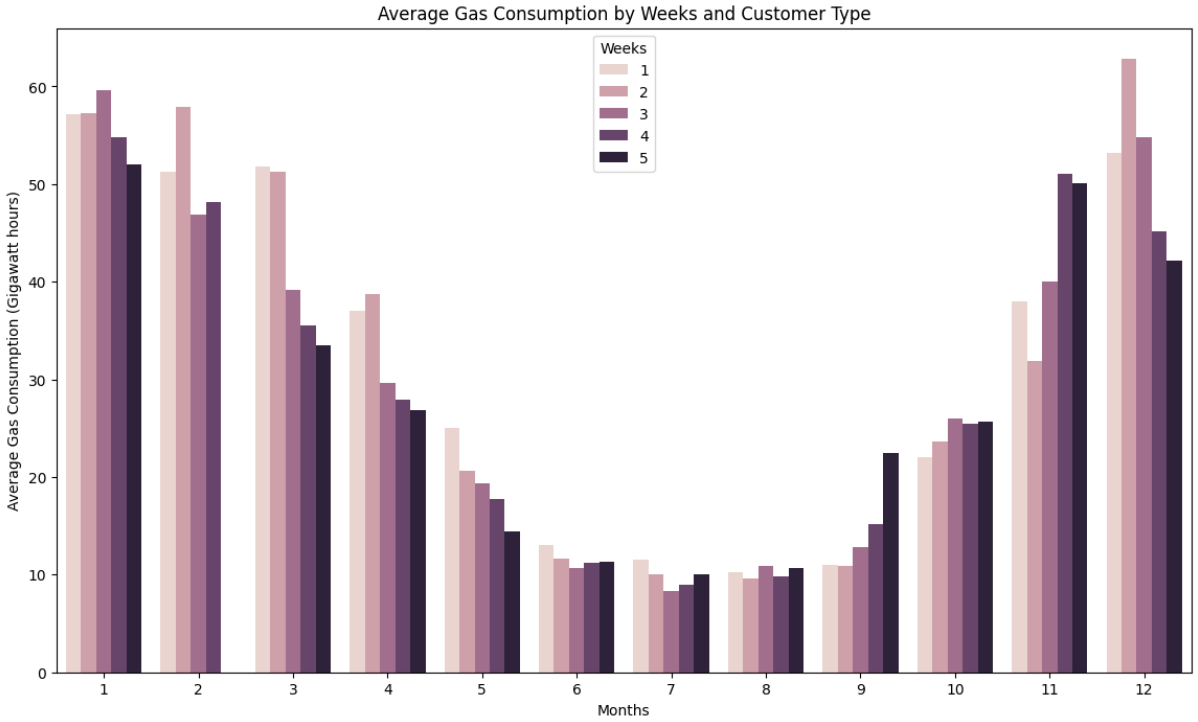


Figure 4

Impact of Weekdays on Gas Consumption:

It is known that gas usage could be influenced by the days of the week. To explore this, I analyzed the data and discovered that each weekday has a different level of gas consumption, with weekends showing the lowest consumption across the entire week in every month (as shown in the graph below). Consequently, I added day names as a feature to the model to capture this variation.

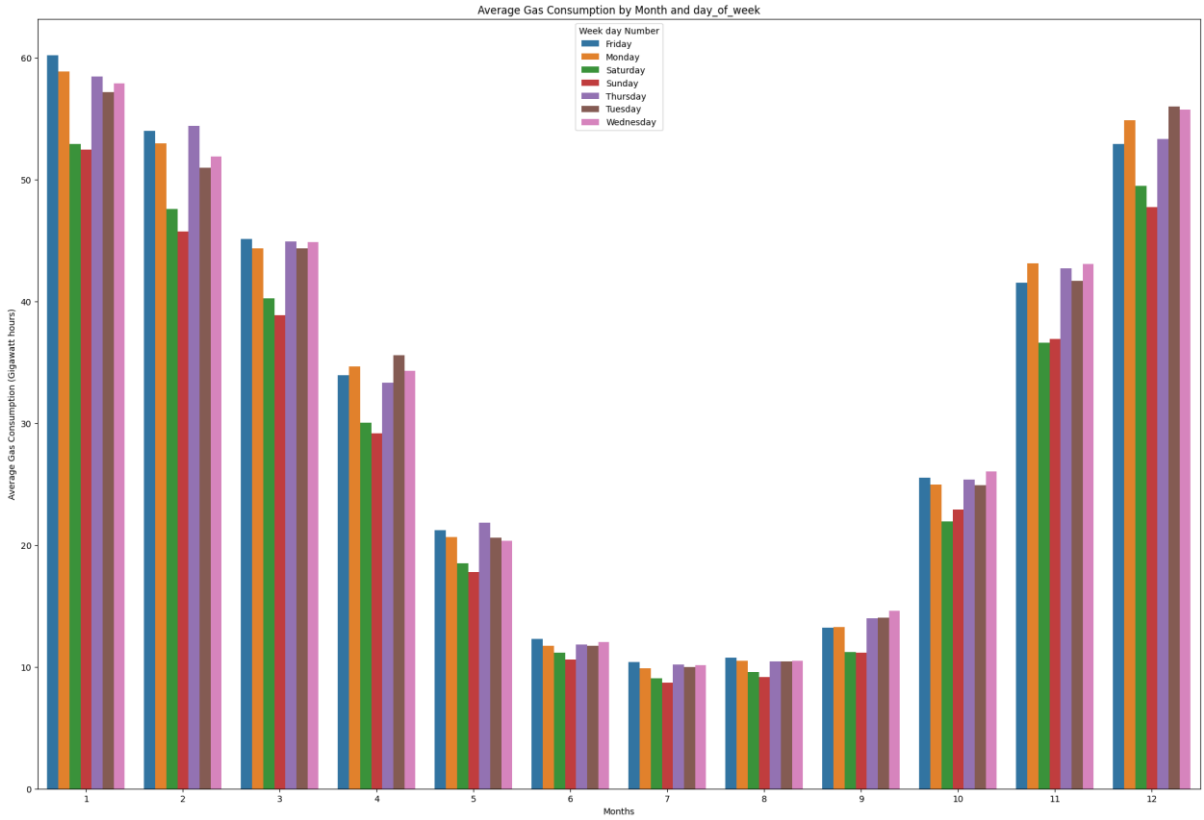


Figure 5

Preprocessing

In-depth analysis plays a crucial role in effective data preprocessing. The insights gained from the analysis significantly contributed to the preprocessing steps outlined below:

Feature Engineering

Feature engineering is a key step in machine learning, as it helps the model learn better and improves its accuracy. The analysis performed earlier guided the creation of additional features that enhanced model performance. The newly created features allowed the model to better capture trends and relationships in the data, ultimately leading to improved predictions.

List of Features Created:

- **Day_of_week**: Represents the day of the week, such as Thursday, Friday, Saturday, or Sunday.
- **Week Number**: Represents the week number within the month. The month was divided into five weeks to capture finer temporal patterns.
- **Month**: Since gas consumption shows a strong relationship with months, the month name was added as a feature to help the model capture monthly variations.

Complimentary Data:

As shown in the analysis, gas consumption is strongly influenced by weather factors. To account for the impact of weather, I integrated weather data based on the date, ensuring that relevant weather features were considered in the analysis.

List of Features from Weather Data:

- **maxtp**: Maximum temperature of the day.
- **mintp**: Minimum temperature of the day.
- **rain**: Amount of rain on the day.
- **sun**: Number of hours of sunlight on the day.

Outlier Detection:

Upon reviewing the target variable, I found that there were no outliers present. Therefore, no outlier removal was necessary for the data. format.

3 Machine Learning Algorithms

For the task of predicting daily gas consumption, we proposed two approaches: a Regression Approach and a Forecasting Approach. Both approaches were selected based on their suitability for our problem, as it involves predicting continuous values, making it a regression and forecasting task.

Regression Approach:

For the regression approach, I trained six models, including Linear Regression, RANSAC Regressor, Random Forest Regressor, Gradient Boosting Regressor, Ridge Regression with Cross-Validation, and Support Vector Regressor (SVR). These models were trained using tacking approach, so i can select the model which performs better. Based on this approach, Linear Regression emerged as the best model, with Gradient Boosting Regressor ranking as the third-best model after Ridge Regression (linear model). The results can be seen in the graph below.

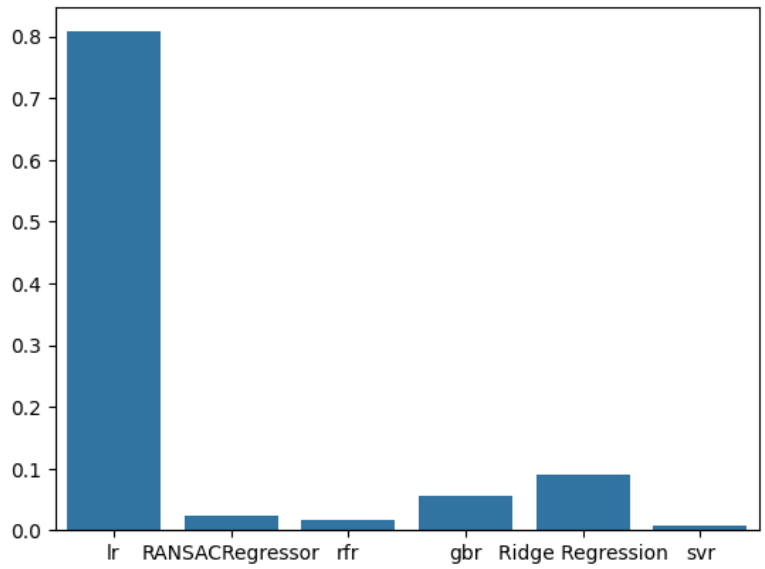


Figure 6

To further evaluate the effectiveness of this approach and assess the models separately, I selected the best linear model (Linear Regression) and the best non-linear model (Gradient Boosting Regressor) for training, aiming to achieve the best accuracy. I trained these models separately on the data and tested their performance to compare their results. *(these results will be discussed in evaluation section)*

Feature Selection:

For feature selection, I applied the Permutation Importance algorithm using the Gradient Boosting Regressor model. This approach helped identify the most important features contributing to the prediction. After selecting the significant features, I retrained the model using only these features, which resulted in a slight increase in accuracy with a reduced number of features. The feature importance graph is provided below, showcasing the most influential features.

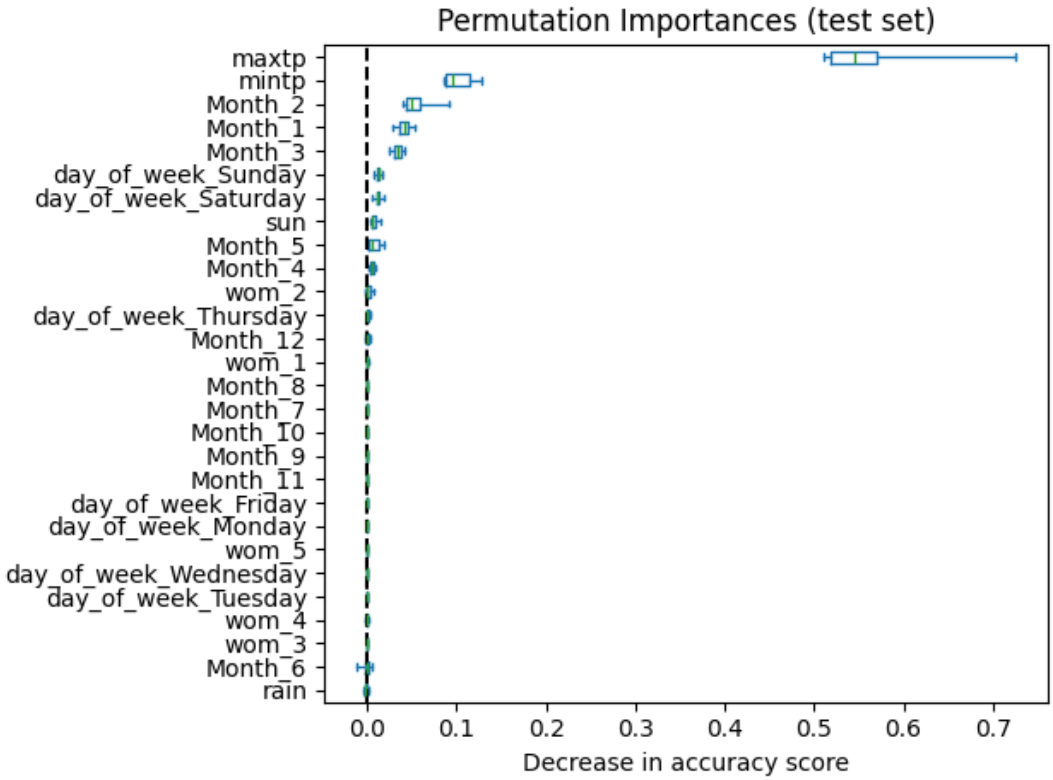


Figure 7

Forecasting:

So far, I have used regression models to predict gas consumption. Now, I aim to explore the use of a forecasting model. For this, I employed the SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous regressors) model, as the data exhibits seasonal trends. Additionally, I included extra features that were selected through Permutation Feature Selection, along with some manually adjusted features to better capture the underlying patterns in the data. The results of this model will be discussed in evaluation section

4 Evaluation and Discussion

Both the regression and forecasting models yielded significantly good results. The primary reason for these successful outcomes lies in the thorough analysis conducted on the data. This analysis helped me identify key trends, whether they were yearly, monthly, weekly, or within each weekday. By understanding these trends, I was able to create effective features that enabled the models to better capture the underlying patterns and factors influencing gas consumption.

The importance of this feature engineering is demonstrated by the results of the Permutation Feature Selection model, which identified all the features created through this analysis as

important (shown in the corresponding plot in the Machine Learning section above).Moreover, each model we trained with our feature engineering demonstrated strong performance, effectively capturing the underlying trends in the data. It can be seen in the model comparison table below, the SARIMAX model without feature engineering exhibits significantly higher errors compared to the SARIMAX model with feature engineering. This highlights the crucial role that data analysis, and features engineering played in improving model performance and reducing errors.

Moreover, through my analysis, I discovered that the data exhibits seasonal patterns, which led me to apply the SARIMAX model for forecasting, as it is well-suited for handling seasonality. The additional analysis also helped me better understand the distribution of the target variable. While the data was imbalanced, it did not contain any significant outliers, which further supported the decision to use these models

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-Squared
Linear Regression	44.31	6.66	0.84
Gradient Boosting Regressor	41.85	6.47	0.85
SARIMAX (without Feature Engineering)	86.67	9.31	0.69
SARIMAX (with Feature Engineering)	74.04	8.60	0.73

Prediction of Gradient Boosting Regressor on validation data along with true values.

Gradient Boosting Regressor:

Prediction of Gradient Boosting Regressor on validation data along with true values.

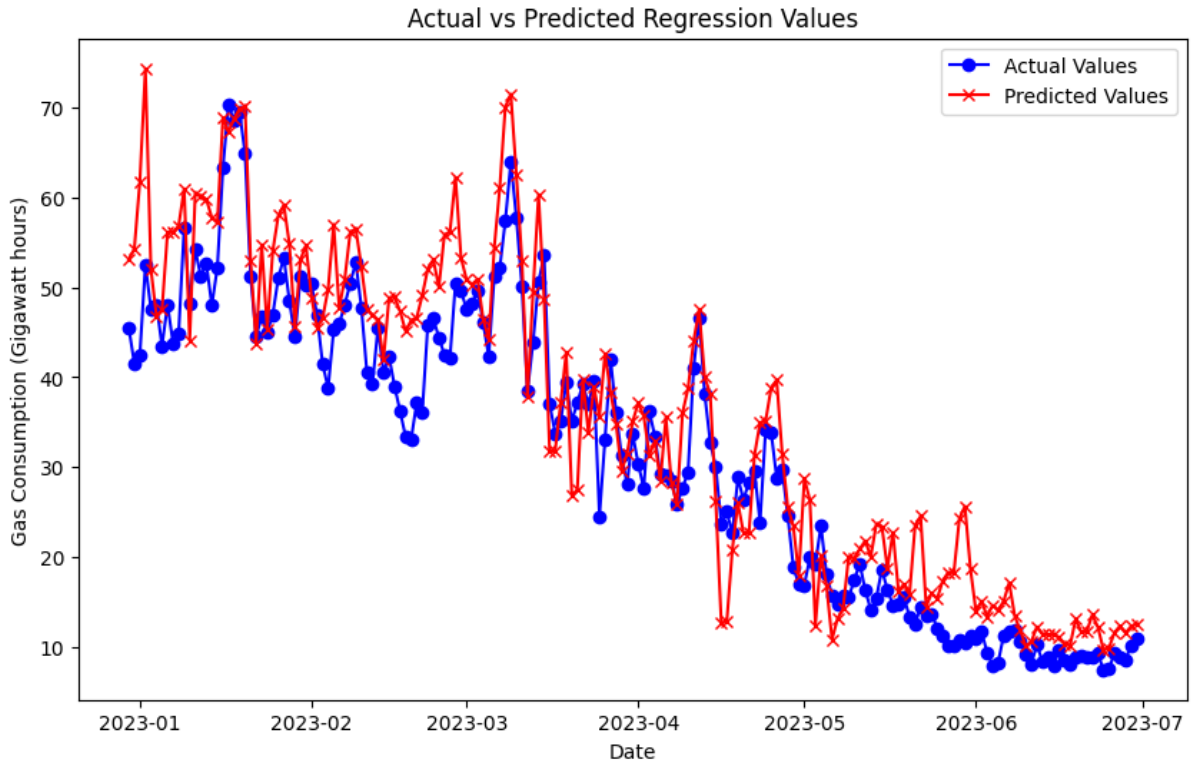


Figure 8

Linear Regression:

Prediction of Gradient Linear Regression on validation data along with true values.

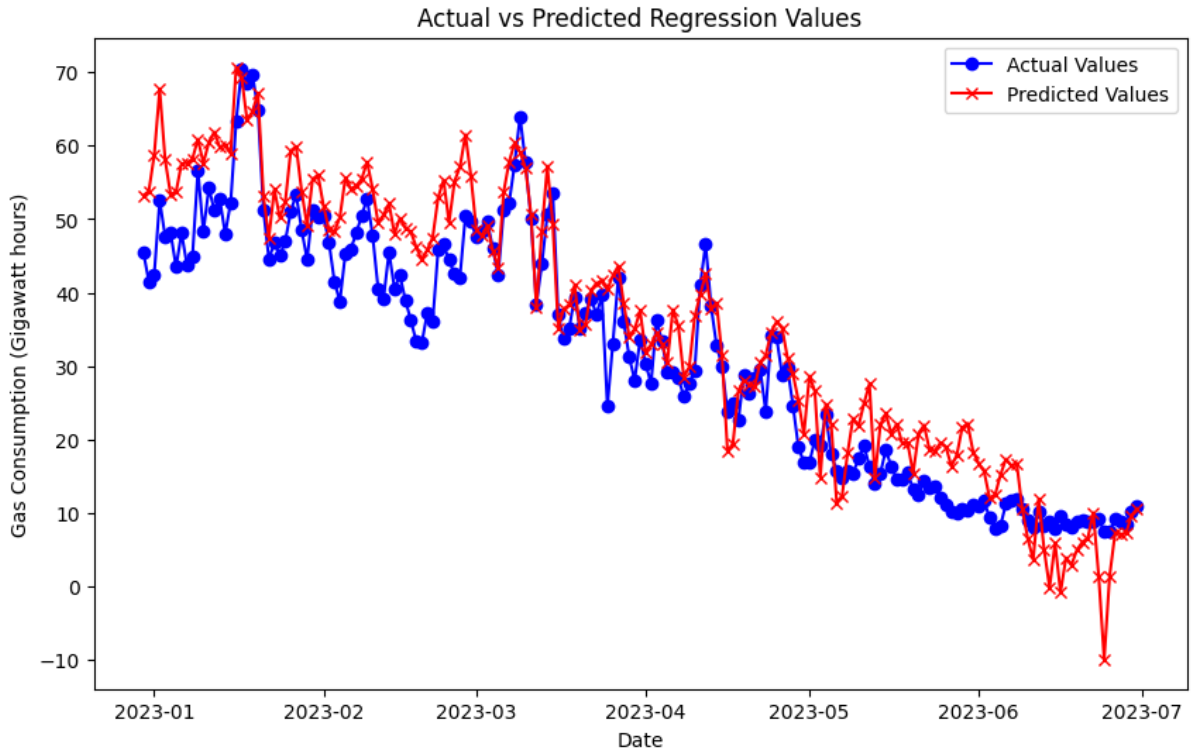


Figure 9

Forecasting (without Feature Engineering):

The SARIMAX model was applied to the validation data without the additional features created using feature engineering, and forecasted for 183 steps. The predicted values show higher errors and limited accuracy compared to true values.

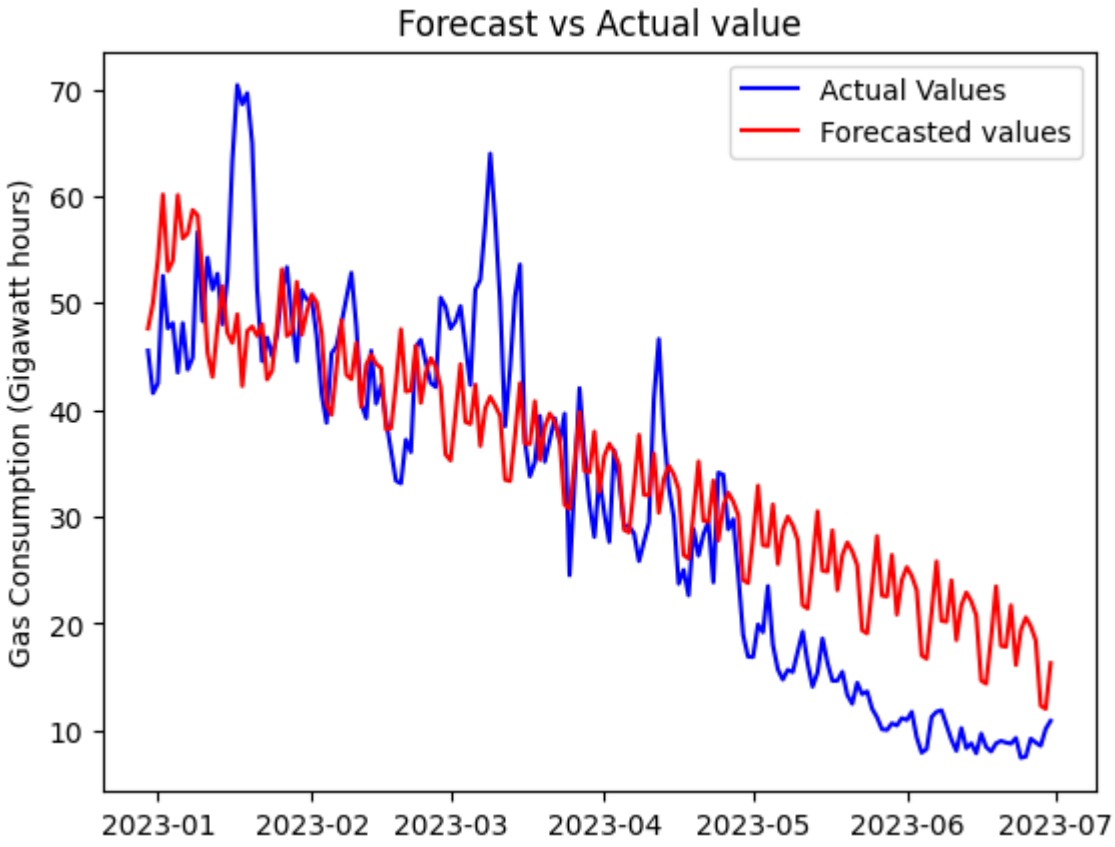
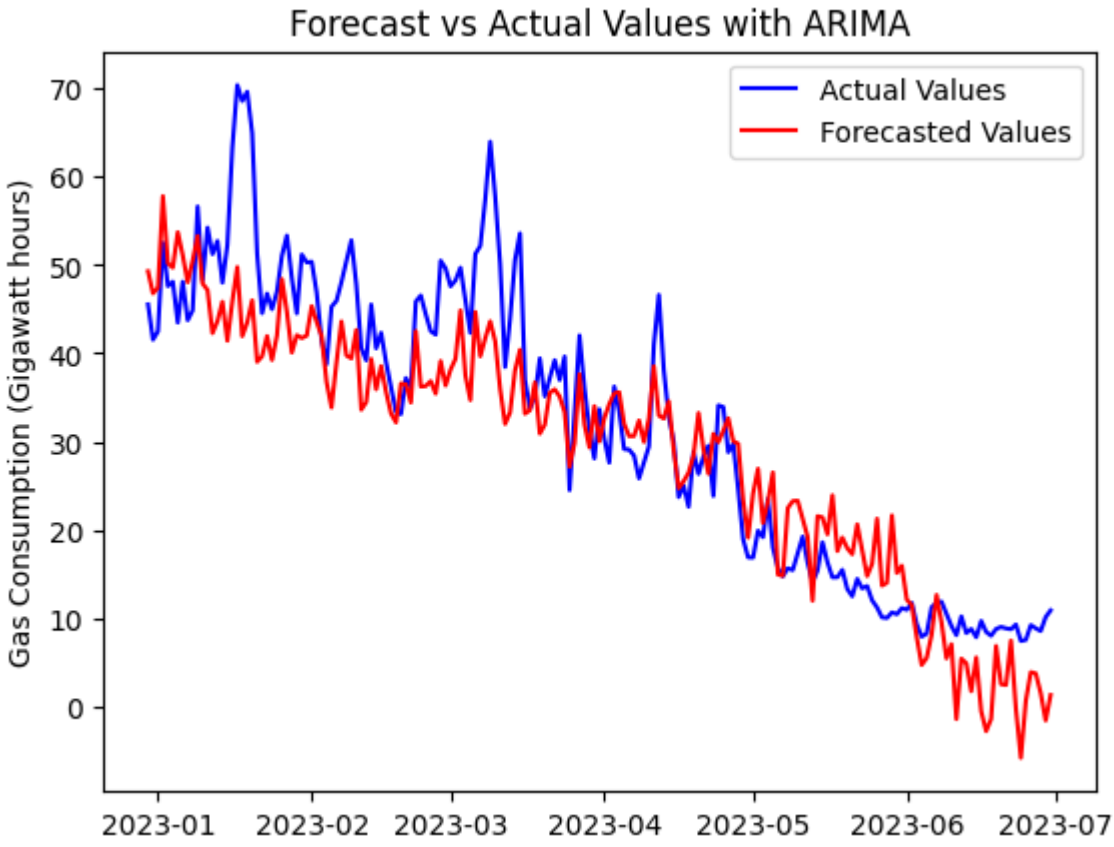


Figure 10

Forecasting (With Feature Engineering):

The SARIMAX model was applied to the validation data with additional features created through feature engineering ,and forecasted for 183 steps. The inclusion of these features improved the model's accuracy, resulting in lower errors and better alignment with the true values.



5 Conclusion

In this project, I utilized a range of data analytics, and machine learning techniques to predict daily gas consumption in Ireland. The primary goal was to understand the dataset through exploratory analysis and then perform feature engineering based on the insights gained. Afterward, both regression and forecasting models were trained on these features to improve prediction accuracy. Significant results were achieved due to the thorough data analysis and effective feature engineering.

A key insight was that different models respond better to different types of feature engineering. For example, the SARIMA model performed better with a different set of features compared to tree-based models. This highlighted the importance of model-specific feature engineering. In conclusion, the robustness of the models was greatly dependent on data analysis is and features engineering, so this project emphasized the importance of the analysis and feature engineering.

References

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22