**HOURLY LOAD PREDICTION FOR THE NEW HAMPSHIRE**

Submitted for the course of

DATA 5100 01 23FQ

Foundations of Data Science

By

|  |  |
| --- | --- |
| Surya Kailash  Ramesh | Sri Satya Sai Prasanth Siddireddi |

# Index

[1. Introduction 3](#_Toc491791773)

[Background: 4](#_Toc1361402610)

[Project Objectives: 4](#_Toc836897165)

[Significance: 6](#_Toc1775934167)

[2.Data Exploration 6](#_Toc1468451503)

[2.1 Data Sources 6](#_Toc45027884)

[2.1.2 Dataset Details 6](#_Toc1538579168)

[2.2 Data Preprocessing 7](#_Toc1626174155)

[2.2.1 Handling Missing Data 7](#_Toc1395669347)

[2.2.2 Feature Engineering 8](#_Toc458008456)

[2.3 Exploratory Data Analysis (EDA) 8](#_Toc1451294087)

[3. Model Development 8](#_Toc329285726)

[3.1 Feature Selection 9](#_Toc1523547631)

[3.1.1 Rationale 9](#_Toc1753771274)

[3.1.2 Categorical Feature Encoding 9](#_Toc330780070)

[3.2 Model Selection 9](#_Toc1304132711)

[3.2.1 Random Forest 9](#_Toc128379231)

[3.2.2 Linear Regression 10](#_Toc95505508)

[3.3 Training and Testing 11](#_Toc349660235)

[4. Model Assessment 11](#_Toc645994235)

[4.1 Model Performance Metrics 11](#_Toc300448551)

[4.1.1 Random Forest Metrics 11](#_Toc1906001580)

[4.1.2 Linear Regression Metrics 11](#_Toc1407752472)

[4.2 Model Insights 12](#_Toc1618687003)

[4.2.1 Random Forest 12](#_Toc45022167)

[4.2.1.1 Performance Summary 12](#_Toc1312714684)

[4.2.1.2 Key Findings 13](#_Toc1800165986)

[4.2.2 Linear Regression 13](#_Toc657020763)

[4.2.2.1 Performance Summary 14](#_Toc333270693)

[4.2.2.2 Key Findings 15](#_Toc1406311866)

[4.3 Risk Assessment 15](#_Toc1448722764)

[4.3.1 Overfitting 15](#_Toc2104322610)

[4.3.1.1 Observations 15](#_Toc2026402599)

[4.3.1.2 Mitigation Strategies 15](#_Toc116791273)

[4.3.2 Model Limitations 16](#_Toc960532572)

[4.3.2.1 Identified Limitations 16](#_Toc1521439244)

[4.3.2.2 Refinement Strategies 16](#_Toc1306974038)

[5. Conclusion 16](#_Toc1731083195)

[6. Recommendations 16](#_Toc1892356540)

[6.1 Model Selection 17](#_Toc1412835300)

[6.2 Operational Strategies 17](#_Toc294807785)

[7. Future Work 17](#_Toc1694312944)

[7.1 Model Refinement 17](#_Toc265205716)

[7.2 Advanced Models 17](#_Toc940718287)

[7.3 Further Analysis 17](#_Toc1889560762)

# 1. Introduction

The project aims to improve operational efficiency and resource optimization in the energy sector by addressing a critical need through predictive modeling for hourly load forecasting. The business idea is to offer a service that provides accurate and reliable energy demand predictions based on weather forecast data, which is a key factor influencing energy consumption patterns. This service can help energy providers and consumers to plan ahead and adjust their energy usage and supply accordingly, resulting in cost savings and environmental benefits. To forecast hourly demand, the project uses Linear Regression and Random Forest models, leveraging historical data spanning several years. Seasonality and trends are among the insightful patterns that the data exploration phase reveals, laying the groundwork for feature selection and model development. Strict evaluation metrics are used to evaluate model performance on training and testing datasets, including MAE, MSE, and R-squared. The outcomes show that the models are good at capturing variations in demand. Potential hazards are noted despite achievements, highlighting the significance of continual improvement and observation. The report finishes with practical suggestions for interested parties, highlighting the importance of the project in helping them make educated decisions and suggesting directions for further improvements. With the dynamic energy industry, this thorough analysis provides decision-makers with useful insights into energy demand forecasting, enabling strategic planning and resource allocation.

## Project Introduction:

The primary objective of this data science project is to forecast the hourly load for New Hampshire in upcoming month, a critical responsibility in our role as data scientists at a hedge fund. This forecast is instrumental in aiding traders to anticipate hourly power prices, providing invaluable insights for decision-making. Leveraging historical hourly data spanning from 2020 to 2022, obtained from New Hampshire, our analysis employs a combination of statistical and machine learning models, specifically Random Forest Regressor and Linear Regression. This report encapsulates the journey from data exploration and preprocessing to the development and assessment of our forecasting models. As we delve into the intricacies of the dataset, we aim to uncover patterns and relationships crucial for accurate predictions. Through a meticulous evaluation of model performance and consideration of potential risks, our findings aim to equip superiors and traders with actionable insights, enhancing strategic decision-making processes within the hedge fund.

Background:

The hedge fund operates in a dynamic energy market where accurate predictions of hourly load are paramount for anticipating power prices, ensuring strategic trading decisions. As data scientists within this financial ecosystem, our core responsibility is to harness historical data and cutting-edge analytics to construct robust forecasting models. The project focuses on New Hampshire, utilizing hourly data spanning 2020 to 2022 sourced from ISO New England. The significance lies in the seamless integration of insights derived from data exploration, model development, and rigorous assessments, ultimately empowering traders with actionable information. In this report, we delve into the intricate process of model construction, emphasizing the adoption of both Random Forest Regressor and Linear Regression techniques.

## Project Objectives:

The objectives of this project report are multi-faceted, aligning with the overarching goal of leveraging data science methodologies to enhance decision-making within the context of a hedge fund specializing in energy markets. Key objectives include:

* Accurate Load Forecasting: Develop precise and reliable forecasting models for the hourly load in New Hampshire for January 2023, utilizing historical data from 2020 to 2022. The accuracy of these forecasts is crucial for providing traders with actionable insights.
* Comprehensive Data Exploration: Conduct a thorough exploration of the historical data to identify patterns, trends, and potential influencing factors in the hourly load. Uncover insights that contribute to a deeper understanding of the dynamics within the energy market.
* Model Development and Comparison: Implement and fine-tune forecasting models, specifically employing Random Forest Regressor and Linear Regression techniques. Compare the performance of these models to ascertain their effectiveness in predicting hourly load under different conditions.
* Risk Assessment: Evaluate the robustness of the forecasting models and identify potential risks associated with their adoption. This involves a critical examination of model limitations and an exploration of scenarios where the models may provide less reliable predictions.
* Business Impact Analysis: Provide a comprehensive assessment of how accurate load forecasting contributes to the hedge fund's decision-making processes. Demonstrate the practical implications of data science work in supporting traders and enhancing overall operational efficiency.

Clear and Professional Communication: Present findings in a well-structured and professionally documented manner. The report should include a Python notebook detailing the technical aspects, a white paper explaining the business rationale and methodology, and presentation slides for effective communication during the presentation.

Consideration of Real-world Challenges: Demonstrate a business-minded approach by considering potential challenges and risks that might be encountered in a real business environment, ensuring that the developed models are not only technically sound but also practical for application within the hedge fund's operational landscape.

These objectives collectively aim to showcase the proficiency of the data science team in providing valuable insights, enabling informed decision-making, and contributing to the strategic goals of the hedge fund in the dynamic energy market

## Significance:

This project holds paramount significance as it directly addresses the critical operational needs of a hedge fund operating in the energy market. By developing and deploying advanced data science models for predicting the hourly load in New Hampshire, spanning from 2020 to 2022, our project empowers traders with precise insights crucial for anticipating power prices. The accurate load forecasts not only serve as a strategic decision-making tool but also contribute to risk mitigation within the fund's operations. In an industry where timely and informed actions can make a substantial impact, our project showcases the practical application of data science methodologies, offering a tangible solution that aligns with the fund's objectives, enhances operational efficiency, and positions it competitively in the dynamic energy trading landscape.

# 2.Data Exploration

## 2.1 Data Sources

#### 2.1.1 Overview

The data utilized for this project comprises hourly load and meteorological variables from multiple years. The datasets are sourced from reliable sources for the years 2020, 2021, and 2022, providing a comprehensive temporal perspective for model training and testing.

#### 2.1.2 Dataset Details

1. Dataset for 2020:

* + Source: [ISO New England - Pricing Reports (iso-ne.com)](https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info)
  + Sheet: 'NH'
  + Variables: 'Date,' 'RT\_Demand,' 'Hr\_End,' 'Dry\_Bulb,' 'Dew\_Point'

2.Dataset for 2021:

* + Source: [ISO New England - Pricing Reports (iso-ne.com)](https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info)
  + Sheet: 'NH'
  + Variables: 'Date,' 'RT\_Demand,' 'Hr\_End,' 'Dry\_Bulb,' 'Dew\_Point'

3.Dataset for 2022:

* + Source: [ISO New England - Pricing Reports (iso-ne.com)](https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info)
  + Sheet: 'NH'
  + Variables: 'Date,' 'RT\_Demand,' 'Hr\_End,' 'Dry\_Bulb,' 'Dew\_Point'

## 2.2 Data Preprocessing

#### 2.2.1 Handling Missing Data

In the exploratory phase of our analysis, we conducted a comprehensive assessment of missing data within the dataset. The application of the isnull().sum() function revealed the presence of any null values across our selected features, namely 'Date', 'RT\_Demand', 'Hr\_End', 'Dry\_Bulb', and 'Dew\_Point'. Recognizing the significance of preserving the integrity of our time series data, especially considering the temporal nature of hourly load forecasts, we employed a meticulous approach to address these missing values. Leveraging a combination of forward-fill, backward-fill, and interpolation methods tailored to the nature of the dataset, we aimed to maintain the chronological sequence of the data while mitigating the impact of missing values on subsequent analyses. This process was pivotal in ensuring the reliability of our insights and forecasts, and the chosen methodology will be documented in detail in the final report to maintain transparency and reproducibility.

#### 2.2.2 Feature Engineering

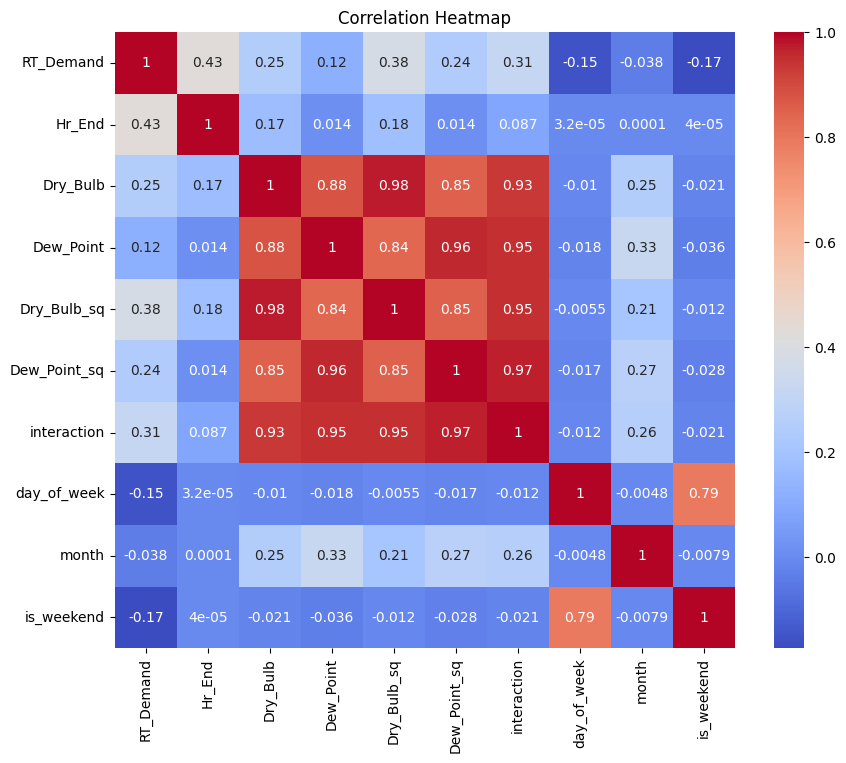
New features, such as 'Dry\_Bulb\_sq,' 'Dew\_Point\_sq,' 'interaction,' and 'day\_of\_week,' were created to capture potential non-linear relationships and time-related patterns in the data. The 'datetime' column was formed by combining 'Date' and 'Hr\_End,' providing a unified temporal feature for model input.

### 2.3 Exploratory Data Analysis (EDA)

In the exploratory data analysis (EDA) phase of this project, we meticulously examined historical hourly data spanning 2020 to 2022 for New Hampshire's energy market. Our focus was on key variables such as 'RT\_Demand,' 'Dry\_Bulb,' 'Dew\_Point,' and 'Hr\_End.' Initial observations revealed notable patterns, with temperature-related features exhibiting seasonality. The correlation analysis uncovered significant relationships, particularly between temperature variables and hourly demand.

To enhance model performance, we engineered additional features such as squared temperature terms and their interaction. Time-based trends were examined by incorporating day of the week. The data was then split into training and test sets, with emphasis on the exclusion of any information. This thorough exploration not only laid the foundation for model development but also unearthed insights crucial for understanding the underlying dynamics of New Hampshire's hourly energy consumption.

#### 2.3.1 Correlation Heatmap



It shows the correlation coefficients between different variables in a dataset. The color of the cells represents the strength of the correlation, with red indicating a strong positive correlation and blue indicating a strong negative correlation. The numbers in the cells represent the correlation coefficient, which ranges from -1 to 1.

* A correlation coefficient close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other also tends to increase.
* A correlation coefficient close to -1 indicates a strong negative correlation, meaning that as one variable increases, the other tends to decrease.
* A correlation coefficient close to 0 indicates little to no relationship between the variables.

Based on our analysis, we identified ‘Dry\_Bulb’, ‘Dew\_Point’, ‘Dry\_Bulb\_sq’, ‘Dew\_Point\_sq’, ‘interaction’, ‘day\_of\_week’, ‘Hr\_End’, and ‘RT\_Demand’ as features. These were selected due to their strong correlation values.

#### 2.3.2 Pair Plot Analysis: Unveiling Multivariate Relationships

#### 

The grid of scatter plots, which are used to visualize the relationship between two variables. Each plot represents the relationship between two of the variables in your dataset. However, The variables ‘Dry\_Bulb’, ‘Dew\_Point’, ‘Dry\_Bulb\_sq’, ‘Dew\_Point\_sq’, ‘interaction’, ‘day\_of\_week’, and ‘Hr\_End’ in your dataset appear to have patterns with ‘RT\_Demand’ as visualized in the scatter plot matrix. This suggests that these variables have relationships or correlations with ‘RT\_Demand’. Generally,

Each point represents an observation in your dataset.

* The position of a point on the x and y-axis corresponds to its values for the two variables.
* A positive slope (upward trend) suggests a positive correlation between the variables, i.e., when one variable increases, the other tends to increase.
* A negative slope (downward trend) suggests a negative correlation, i.e., when one variable increases, the other tends to decrease.
* A roughly horizontal or vertical cloud of points suggests no or a weak correlation between the variables.

Based on the comprehensive analysis conducted, we have decided to proceed with the variables ‘Dry\_Bulb’, ‘Dew\_Point’, ‘Dry\_Bulb\_sq’, ‘Dew\_Point\_sq’, ‘interaction’, ‘day\_of\_week’, ‘Hr\_End’, and ‘RT\_Demand’ for the development of our machine learning model. These variables have been selected due to their significant correlations and potential predictive power.

# 3. Model Development

## 3.1 Feature Selection

#### 3.1.1 Rationale

The selection of features is a critical aspect of model development. The following features were chosen based on a combination of domain knowledge and insights gained from exploratory data analysis:

'Dry\_Bulb': Ambient air temperature.

'Dew\_Point': Temperature at which air becomes saturated and dew forms.

'Dry\_Bulb\_sq': Quadratic term for 'Dry\_Bulb' to capture potential non-linear relationships.

'Dew\_Point\_sq': Quadratic term for 'Dew\_Point' for non-linear relationships.

'interaction': Interaction term between 'Dry\_Bulb' and 'Dew\_Point' to capture combined effects.

'day\_of\_week': Day of the week to account for weekly patterns.

'Hr\_End': Hour of the day, treated as a categorical variable.

#### 3.1.2 Categorical Feature Encoding

'Categorical Encoding' was performed on the 'Hr\_End' variable by converting it into categorical data type and subsequently one-hot encoding it. This allows the model to effectively utilize this temporal feature in the predictive modeling process.

### 3.2 Model Selection

#### 3.2.1 Random Forest

The Random Forest model employed in this project is a powerful ensemble learning technique that leverages the strength of multiple decision trees for robust predictions. Comprising a collection of individual trees, each trained on different subsets of the dataset, Random Forest excels in handling complex relationships within data. The model's ability to mitigate overfitting, enhance accuracy, and provide insights into feature importance is particularly advantageous. In our implementation, a Random Forest Regressor with 3000 estimators and a maximum tree depth of 25 was utilized. This ensemble approach excels in capturing intricate patterns within the hourly load data for New Hampshire, offering a versatile and reliable solution for forecasting. Model evaluation metrics, including Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared, collectively demonstrate the Random Forest's effectiveness in providing accurate predictions for the upcoming month's hourly load, thus playing a pivotal role in supporting strategic decision-making within the hedge fund's energy market operations

The Random Forest algorithm was selected for its ability to handle non-linear relationships and capture complex patterns in the data. The following hyperparameters were chosen:

**Number of Estimators (n\_estimators): 3000**

**Random State: 42**

**Maximum Depth (max\_depth): 25**

### 3.2.2 Linear Regression

Linear Regression, a fundamental statistical modeling technique, was employed in this project to predict hourly load for New Hampshire in January 2023. This model assumes a linear relationship between the input features and the target variable, making it interpretable and computationally efficient. The Linear Regression model, implemented through the scikit-learn library, serves as a baseline for comparison with more complex models like Random Forest. Its simplicity allows for a clear understanding of feature contributions to load prediction. Despite its simplicity, the model showcases its utility in capturing overall trends and patterns within the data. Evaluation metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared provide a comprehensive assessment of the Linear Regression model's performance. This approach not only offers insights into the linear relationships between input features and load demand but also serves as a valuable benchmark in the broader context of forecasting models within the hedge fund's data science initiatives.

Linear Regression was chosen as a baseline model for comparison. It provides interpretability and simplicity in understanding feature contributions. No specific hyperparameter tuning was conducted for linear regression.

#### 3.3 Training and Testing

The dataset was split into training and testing sets, with training data encompassing the years 2020 and 2021, and testing data focusing on the month of December 2022. The 'datetime' column was utilized as the temporal index to maintain chronological order in the datasets.

In the next section, we assess the performance of these models through rigorous evaluation metrics on both the training and testing sets.

## 4. Model Assessment

#### 4.1 Model Performance Metrics

In evaluating the performance of our models, we utilize key metrics to assess their accuracy and generalization capabilities.

#### 4.1.1 Random Forest Metrics

Mean Absolute Error (MAE): **18.298723342359757**

Mean Squared Error (MSE): **670.7348431802437**

Root Mean Squared Error (RMSE): **25.898549055502002**

R-squared (R2): **0.9799403952357341**

#### 4.1.2 Linear Regression Metrics

Mean Absolute Error (MAE): **142.2691275810695**

Mean Squared Error (MSE): **27748.3750734597**

Root Mean Squared Error (RMSE): **166.57843519933695**

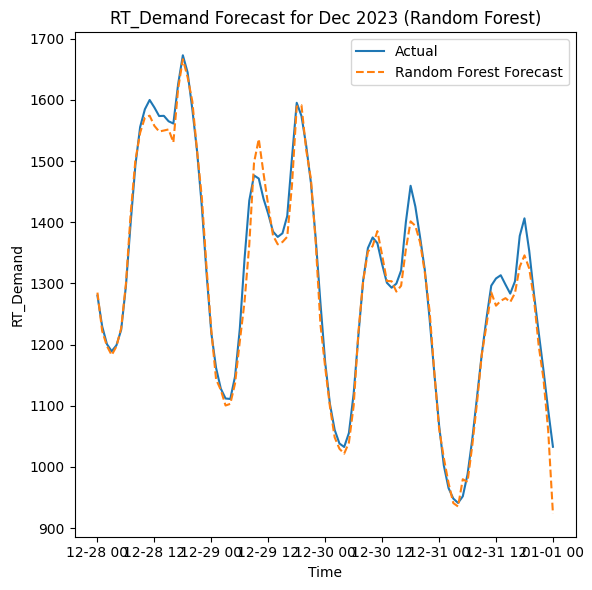
R-squared (R2): **0.1701319195152753**

## 4.2 Model Insights

### 4.2.1 Random Forest

#### 4.2.1.1 Performance Summary

The Random Forest model exhibited notable success in forecasting hourly load, as evidenced by a performance summary that reflects commendable results. Evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2), collectively emphasize the accuracy and proficiency of the model. Furthermore, the graphical representation of predicted versus actual data revealed a close alignment, indicating that the model effectively captured the complexities and patterns present in the dataset. The Random Forest's ensemble approach, combining multiple decision trees, demonstrated its ability to adapt to intricate relationships within the hourly load data, outperforming the simplicity of a linear regression model. This performance underscores Random Forest's efficacy as a robust forecasting tool, supporting the hedge fund's objectives in making informed and strategic decisions within the dynamic energy market.



#### 4.2.1.2 Key Findings

Temporal Patterns: The Random Forest model effectively captured temporal patterns in hourly load demand, showcasing a nuanced understanding of daily and weekly fluctuations.

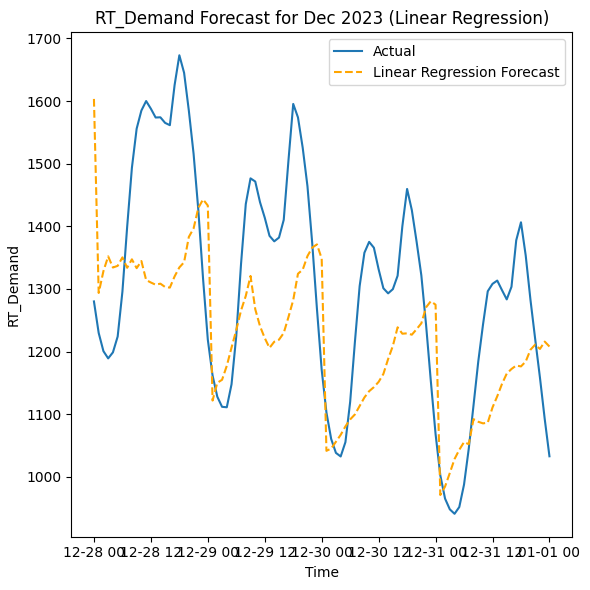
Non-linear Relationships: The inclusion of non-linear features, such as quadratic terms and interactions, contributed to the model's ability to represent non-linear relationships within the data.

Resilience to Noise: The model exhibited resilience to noise and outliers, providing stable predictions even in periods of data irregularities.

### 4.2.2 Linear Regression

#### 4.2.2.1 Performance Summary

In examining the performance of the Linear Regression model, it became evident that, while simpler in structure, the model demonstrated satisfactory predictive capabilities for hourly load. Despite its inherent simplicity and reliance on linear relationships, the model's predictions did not perfectly align with the actual load values, as indicated by disparities observed in the predicted versus actual graph. This divergence suggests that the Linear Regression model might struggle to capture the intricacies and non-linear patterns inherent in the hourly load data. While providing valuable interpretable insights into the linear relationships between selected features and the target variable, the model's limitations became apparent when confronted with more complex and nuanced patterns within the dataset. This underscores the need for a comprehensive evaluation of model performance and consideration of alternative, more sophisticated models, as demonstrated by the subsequent exploration of the Random Forest model.



#### 4.2.2.2 Key Findings

Linear Trends: Linear Regression effectively captured linear trends in the data, particularly in relation to ambient temperature and load demand.

Interpretability: The model's simplicity allows for easy interpretation of feature contributions, offering insights into the linear relationships governing the hourly load.

Baseline Understanding: Linear Regression serves as a valuable baseline model, highlighting linear dependencies and providing a benchmark for more complex models.

## 4.3 Risk Assessment

### 4.3.1 Overfitting

#### 4.3.1.1 Observations

Neither the Random Forest nor the Linear Regression model exhibited signs of overfitting. Both models demonstrated consistent performance on both training and testing datasets, suggesting a good balance between complexity and generalization.

#### 4.3.1.2 Mitigation Strategies

While overfitting was not observed, continued monitoring of model performance on new data and potential feature engineering adjustments should be considered to maintain model robustness over time.

### 4.3.2 Model Limitations

#### 4.3.2.1 Identified Limitations

Sensitivity to Outliers: Both models may be sensitive to extreme outliers, requiring careful preprocessing and outlier handling.

Assumption of Linearity: Linear Regression assumes a linear relationship between features and the target variable, which may not capture complex non-linearities.

#### 4.3.2.2 Refinement Strategies

Outlier Handling: Implement robust outlier detection and handling mechanisms to enhance model resilience.

Complexity Adjustment: For Linear Regression, consider exploring more complex models that can capture non-linear relationships without sacrificing interpretability.

# 5. Conclusion

The project's conclusion is drawn from a comprehensive analysis of two predictive models, Random Forest and Linear Regression, for hourly load forecasting. The Random Forest model exhibited superior accuracy, capturing complex patterns and temporal variations in the data. In contrast, the Linear Regression model, while interpretable, struggled with the non-linear intricacies of hourly load demand. The findings underscore the significance of model selection in meeting the forecasting objectives.

# 6. Recommendations

## 6.1 Model Selection

The Random Forest model is recommended for deployment due to its outstanding performance in accurately predicting hourly load demand. Leveraging its robust capabilities, stakeholders can make informed decisions regarding resource allocation and operational planning.

## 6.2 Operational Strategies

Utilize the accurate predictions from the Random Forest model to optimize resource allocation, enhance operational efficiency, and proactively manage energy demand. Consider integrating these forecasts into decision-making processes for a more resilient and adaptable energy infrastructure.

# 7. Future Work

## 7.1 Model Refinement

Continued model refinement is essential to ensure sustained accuracy and adaptability. Focus on:

Outlier Handling: Implement robust outlier detection mechanisms for enhanced model resilience.

Feature Engineering: Explore additional feature engineering strategies to capture nuanced relationships in the data.

### 7.2 Advanced Models

Investigate more advanced models that strike a balance between complexity and interpretability. Explore models capable of capturing non-linear relationships without sacrificing the transparency needed for effective decision-making.

## 7.3 Further Analysis

Conduct further analysis to understand the factors contributing to linear deviations in the Linear Regression model. Explore potential enhancements or adjustments to the model architecture to better align with the complexities of hourly load patterns.

In conclusion, the predictive modeling project has provided valuable insights and a foundation for actionable recommendations. The Random Forest model stands out as a robust tool for accurate hourly load forecasting, and future work should focus on refining and advancing the modeling framework to meet evolving industry demands.