# PREDICTIVE ANALYTICS REPORT

Improving Singapore environment sustainability by reducing food, plastic, energy, and carbon emissions consumption

| | |
|---|---|
| Module Group | IT2386-03-01 |
| Supervisor | Ms Lim Ai Huey |
| Team Name | 3Js 1 Imposter |
| Team Leader | **Javerine Tan Jing Xuan (220429P)** |
| Strategist | Amber Yeo En (222809U) |
| Planner | Jess Lim Zhi Qi (211031L) |
| Scribe | Jonathan Yap Shi Hao (224548E) |
| Latest Report Amendment Date | 20/08/2023 |

# Table of Contents

# Executive Summary

## Background information

Many Singaporeans share the sentiment that **addressing climate change might seem futile**, given concerns about **costs, inconvenience, and limited impact**. In 2021, 67% worried about costs, 66% saw it as inconvenient, 27% struggled with sustainable habits, 24% found the status quo insufficient, and 22% deemed individual action insignificant. While reversing climate change demands persistent effort, there's a **drive to reduce emissions, plastic, food, and energy wastage**. Singapore's Environmental Performance Index (EPI) reflects this effort, with a 50.9 score in 2022, ranking 44th globally, showing a 3.7 increase over a decade. Despite a considered "poor" score, the **upward trend** signals improving sustainability. The urgency of environmental pollution's impact necessitates **global efforts in reducing waste and emissions.**

## Business Scenario

Our project aims to **provide actionable insights into current consumption trends**, **aiding community organizations, non-profits, and grassroots initiatives in resource reduction efforts.** Through predictive analysis, **we predict resource consumption patterns across sectors, empowering businesses to align with the Green Plan's Objectives**. By optimizing resource management and fostering sustainability, we contribute to a greener future and business growth. Our initiative also benefits academia, offering valuable data for studies in sustainability and predictive modeling. Emphasizing awareness, innovation, and sustainable practices, our project supports a more environmentally responsible future. Timely action is crucial for **reversing environmental impacts and securing a planet for future generations.**

## Stakeholders

Agencies involved in Singapore Green Plan 2030

## Business Goal

Improving Singapore **environment sustainability** by reducing food, plastic, energy, and carbon emissions consumption.

## Objectives Identified

| S/N | Objective identified | Team member |
|---|---|---|
| *1* | Predict the main factors contributing to food wastage and to reduce food waste in those areas. | Jess |
| *2* | Predict the top contributors to plastic waste to raise awareness and reduce plastic wastes in those areas. | Amber |
| *3* | Predict peak energy demand for electricity in Singapore, enabling proactive environmental monitoring and effective resource management strategies. | Javerine |
| *4* | Predict the carbon emission over the years by identifying the top contributors to carbon dioxide from industry, buildings, transport, households, and transport sectors. | Jonathan |

## Success Criteria

1. Develop predictive models of less than 0.05 validation ASE value.
2. Adjusted R-Square to be greater than 0.7.

# Business Objective: Energy

## Reason for choosing objective Energy:

Over the years, energy consumption has emerged as a significant global environmental concern with the surge in the demand for non-renewable resources like oil and natural gas. Currently, about **80%** of the world's total energy is derived from **fossil fuels**. However, the depletion of these resources is inevitable, and once exhausted, cannot be replenished. This will **impact numerous aspects of our lives** of many as we rely heavily on these resources to meet our daily energy needs.

Singapore has limited natural resources and renewable energy. In 2021, Singapore imported approximately **149.4 million tonnes of energy** products and **60.3%** of these imports continue to be in the form of Petroleum Products. These massive imports are to help us cope with our increasing electricity consumption. In just one year, from 2020 to 2021, our country-wide electricity consumption increased **by 5.3%** from 50.8 terawatt-hours (TWh) to 53.5 TWh. Aside from the issue of limited supply of non-renewable energy sources, the heavy usage of these non-renewable energy sources poses a greater issue. The extraction and combustion of fossil fuels releases carbon dioxide into the atmosphere, resulting in the rise in greenhouse gas levels, climate change and its detrimental effects.

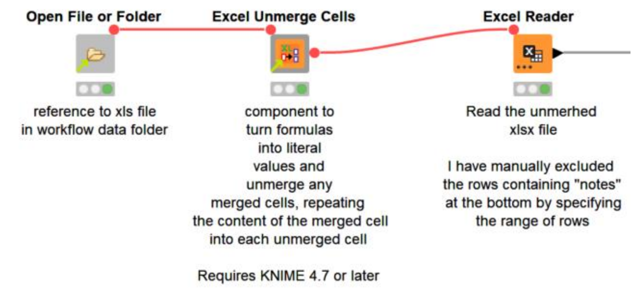
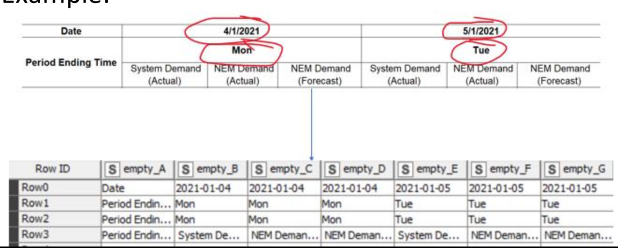
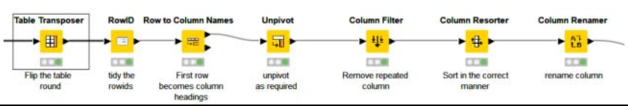
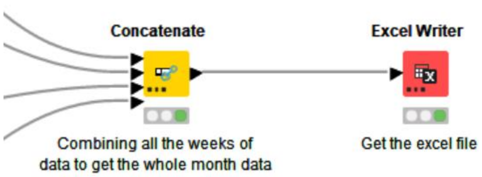
Accurately **predicting peak electricity demand** can play a crucial role in addressing these **environmental challenges** by facilitating proactive environmental monitoring and promoting sustainable energy consumption practices. By forecasting peak demand with precision, it will **help to identify opportunities to optimize electricity usage, implement demand response programs and develop strategies to help reduce consumption during high-demand periods**. Effectively managing energy resources in this manner can reduce the environmental impact associated with excessive energy production and consumption, leading to a more **sustainable and greener future**.

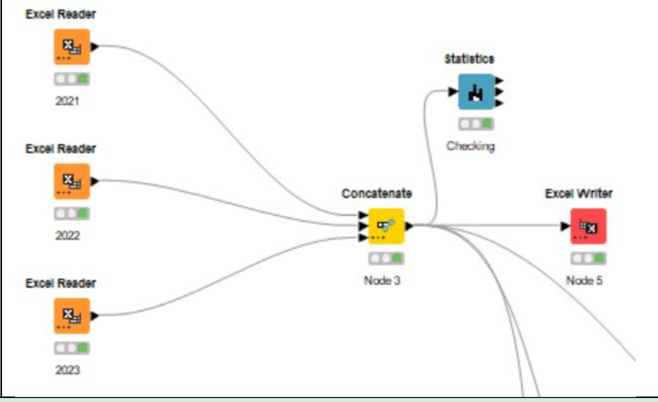
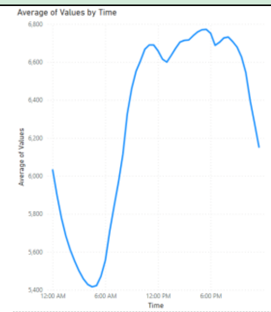
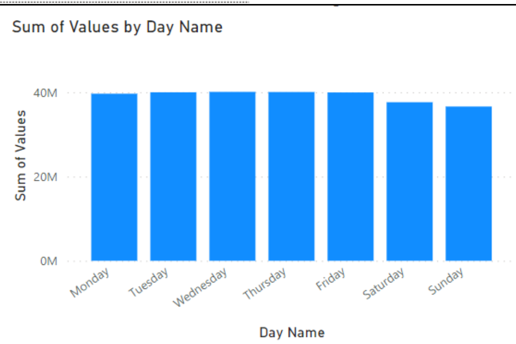
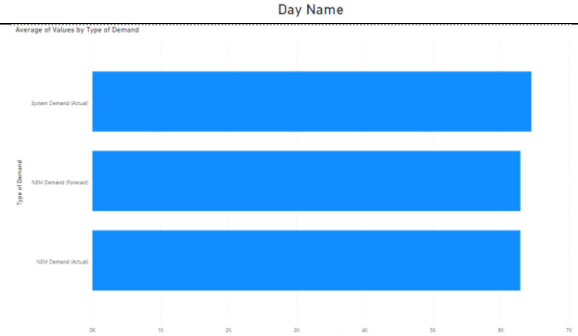
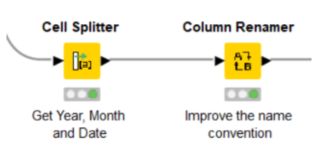
## Transdisciplinary

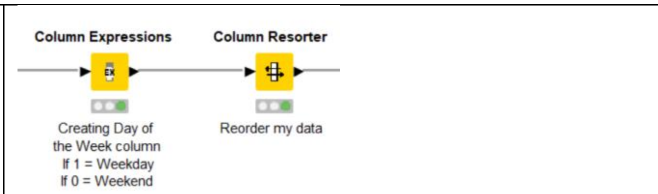
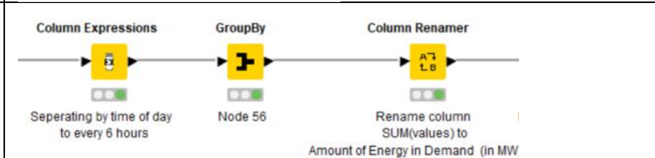
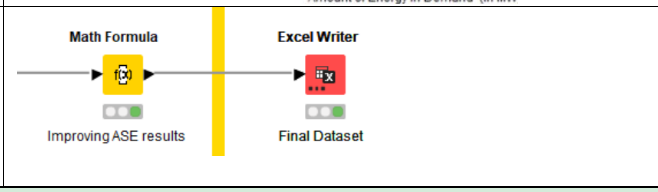
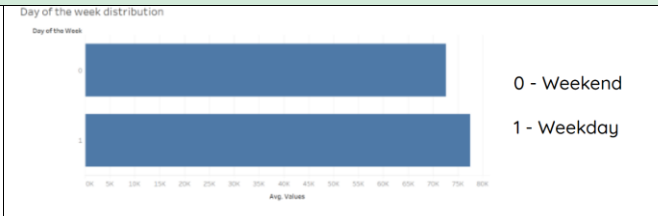
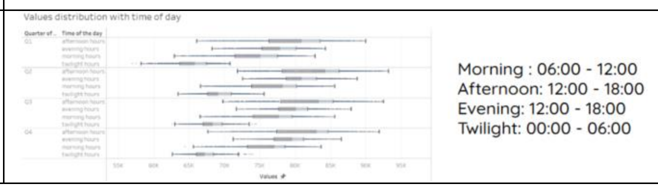
Dr. Koh suggested various methods for finding useful data. She encouraged us to explore publicly available datasets from reputable sources such as government agencies, research institutes, and data repositories. By leveraging these sources, we can access a wide range of data that aligns with our research question or objective.

Furthermore, Dr. Koh highlighted the significance of data preprocessing and data cleaning techniques in SAS. She advised us on the importance of carefully examining and handling missing values, outliers, and inconsistencies to ensure the accuracy and integrity of our analyses. By applying data manipulation and transformation functions in SAS, we can prepare the data for further analysis.

## Data Preparation

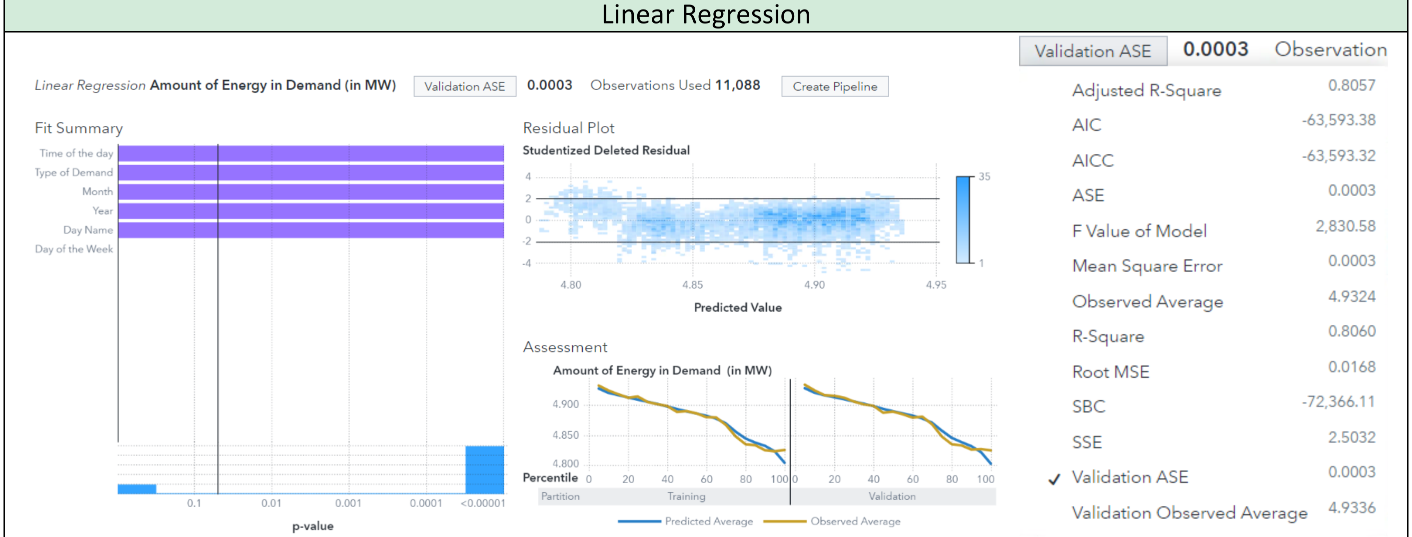
<table>
<tr><th colspan="3">Data Overview</th></tr>
<tr><td colspan="3">Data Source: https://www.ema.gov.sg/resources/statistics/half-hourly-system-demand-data<br>Year collected: Jan 2021 to Mid July 2023<br>Number of rows before cleaning: 133057<br>Number of rows after cleaning: 7392<br>Target column: Amount of Energy in Demand (in MW)<br>Feature columns: Year, Month, Day Name, Type of Demand, Day of the week, Time of the day</td></tr>
<tr><td>Steps</td><td>Image</td><td>Details</td></tr>
<tr><td colspan="3">Data cleaning</td></tr>
<tr><td>1</td><td>
<br>Example:<br>
</td><td>Convert unmerged cells to unmerged while preserving values. These nodes are part of a set created by KNIME for me. Requires KNIME 4.7 or later.<br><br>Open File or Folder is like Excel Reader, but it references to the xls file in workflow data folder. This enables Excel Unmerge Cells node to read and unmerge cells, duplicating content. Generate a new .xlsx file auto-read by Excel Reader.</td></tr>
<tr><td>2</td><td>
</td><td>Getting the Date, Day, Time, Values and Type of Demand as a Column</td></tr>
<tr><td>3</td><td colspan="2">Repeat Step 1 and 2, based on the number of weeks in a month. (Step 1 and 2 is one week of cleaning of data.)</td></tr>
<tr><td>4</td><td>
</td><td>The Concatenate node merges four/five weeks of data into a single month. To validate complete and accurate data collection, employ the Statistics node to detect any missing date or values. After thorough cleaning and collection, utilize the Excel Writer for the final clean data in Excel format.</td></tr>
<tr><td>5</td><td colspan="2">Repeat step 4, 12 times to collect one year of data. Then use a concatenate node to combine all the months of data into one to get one year of data.</td></tr>
</table>

| | | |
|---|---|---|
| 6 |  | Upon gathering data from all years, using the **concatenate** node to merge it into the ultimate dataset. Validate completeness, particularly for dates and values, via the **Statistics** node. Once done use the **Excel Writer** for the final dataset in Excel Format. |
| Data Understanding | | |
| 1 |  | Looking at the timing, we can see that about every there is a varying of the demand of electricity used. The top time of the amount of electricity used is at 5.30PM. |
| 2 |  | Electricity consumption remains consistent from Monday to Friday, but notably decreases on weekends. To differentiate between weekdays and weekends, a binary grouping approach assigns 1 for weekdays and 0 for weekends, aiding prediction. |
| 3 |  | The NEM Demand Actual is much higher than the NEM Demand (Forecast). As its NEM Demand (Forecast) is a forecast, it will be removed in the Data Transformation from the Type of demand. |
| Data Transformation | | |
| 1 |  | To see if the year and month have a significance on when the amount of energy is in demand, using the cell splitter, split the date column to get the year and month. To allow easier understanding of the columns, rename the columns to year and month using the **Column Rename**. |

<table>
<tr><td>2</td><td></td><td>Seeing if weekdays or weekends have any significance for energy in demand. And reordering to allow easier finding of data, to keep the dataset organized.</td></tr>
<tr><td>3</td><td></td><td>This second transformation was done after realizing that the timing was to miscue, therefore grouping the timing by every 6 hours which is commonly done to group the time of day.</td></tr>
<tr><td>4</td><td></td><td>To improve the ASE reading and value, log transformation of the target column using the Math Formula node. Once done, using the Excel Writer get the excel file, to be use later in the model.</td></tr>
<tr><td colspan="3">Data Understanding</td></tr>
<tr><td>1</td><td></td><td>After grouping the day of the week by weekday, we can see they are about the same, but weekday uses more amount of electricity, compared to weekend. We can see the primary that the primary day of the week contributes to electricity demand is weekday.</td></tr>
<tr><td>2</td><td></td><td>The skewness of the box plot provides insights into the time of the day distribution. The longer the whisker on one side, it indicates skewness towards that direction.</td></tr>
</table>

## Modeling

**Linear Regression**

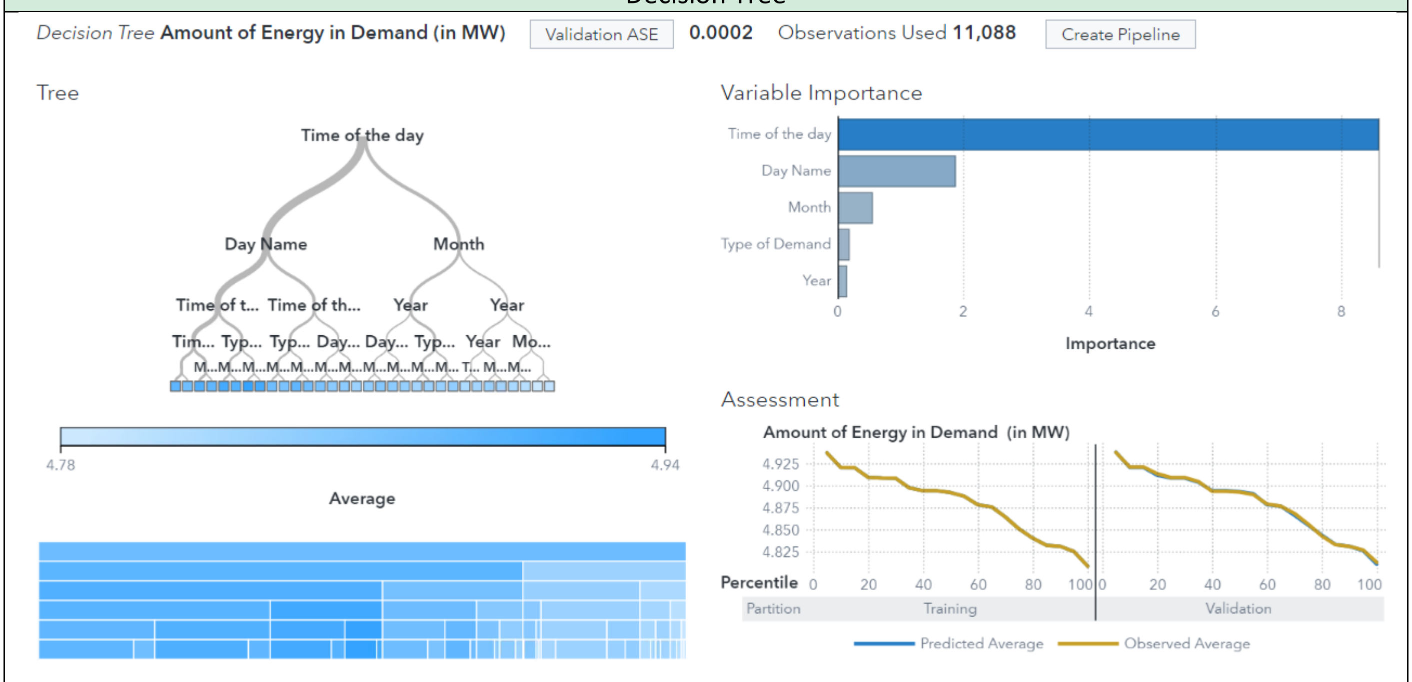



**Tuning**

| Parameters Tuned | Validation ASE | Training ASE |
|---|---|---|
| Variable Selection Method: None (Default) | 0.0003 | 0.0003 |
| Variable Selection Method: Forward | 0.0003 | 0.0003 |
| Variable Selection Method: Backward | 0.0003 | 0.0003 |
| Variable Selection Method: Stepwise | 0.0003 | 0.0003 |

| | |
|---|---|
| | After tunning the variable selection method to forward, backward and stepwise, there is no difference in the validation and training ASE and the adjusted R square value, hence I have kept the original of not using a variable selection method. |
| Details | Using linear regression is to establish relationships between input variables and the Amount of Energy. To ensure the model's reliability, we conducted a comprehensive validation process, by dividing our dataset into training and validation sets by 80 and 20 respectively. By assessing the model's performance on unseen data, we gained insights into generalization ability.<br>The results of our linear regression model indicated an Adjusted R-square value of 0.8057. This value signifies that approximately 80.5% of the variance in the Amount of Energy in Demand is accounted for by our chosen input variables. While not capturing the full complexity of the system, this demonstrates a significant degree of explanatory power. Additionally, our validation process yielded a validation ASE of 0.0003. This low value implies that the model's predictions deviated from the actual values by a very small margin. Importantly the Adjusted R-square value surpassed the 0.7 threshold, suggesting a strong fit between the model and the data.<br>The success criteria we established were met, as our ASE was well below 0.05 and the Adjusted R-square surpassed 0.7. These outcomes affirm the efficiency of the model in explaining Energy Demand variability. Thus, this model can be used for decision-making and resource allocation in managing electricity demand. |

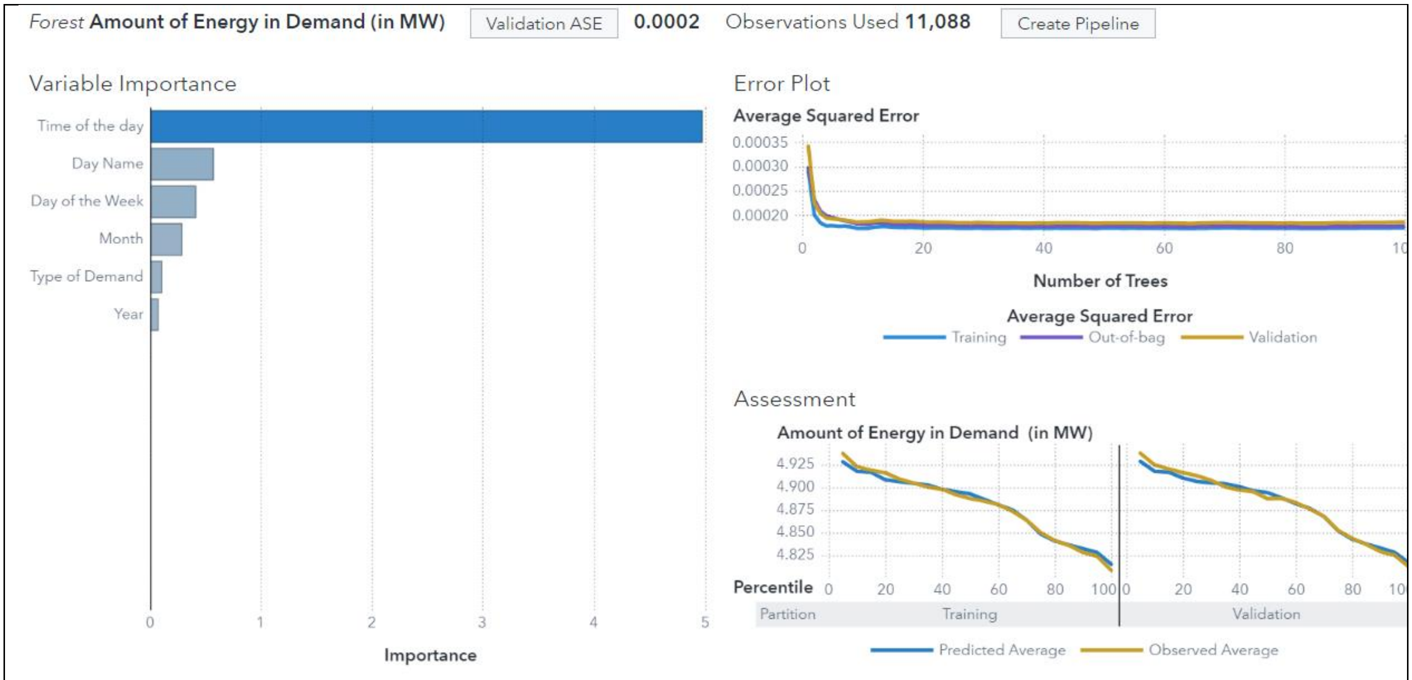
**Decision Tree**




| Tuning | Parameters Tuned | Validation ASE | Training ASE |
|---|---|---|---|
| | Maximum Levels: 6, Leaf Size: 5, Predictor Bins: 50 (Default) | 0.0002 | 0.0002 |
| | Maximum Levels: 6, Leaf Size: 10, Predictor Bins: 50 | 0.0003 | 0.0003 |
| | Maximum Levels: 6, Leaf Size: 5, Predictor Bins: 100 | 0.0003 | 0.0002 |
| | Maximum Levels: 12, Leaf Size: 5, Predictor Bins: 50 | 0.0002 | 0.0002 |

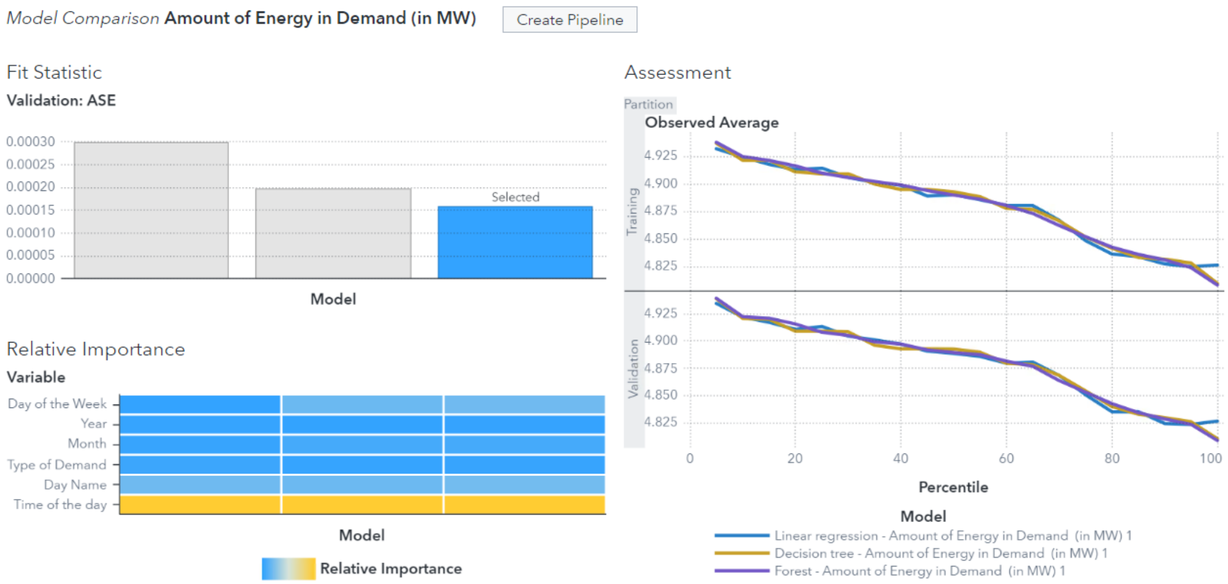
| | |
|---|---|
| Details | After tuning the parameters, I found that the original parameters Maximum Levels: 6, Leaf Size: 5, Predictor Bins: 50 yielded the best model. My validation and training ASE values are the same, signifying that the model is not overfitted. As the ASE is lower than 0.05, my model has met my success criteria. This implies that the model's prediction closely aligns with the actual data, affirming its accuracy and reliability. |

**Random Forest**



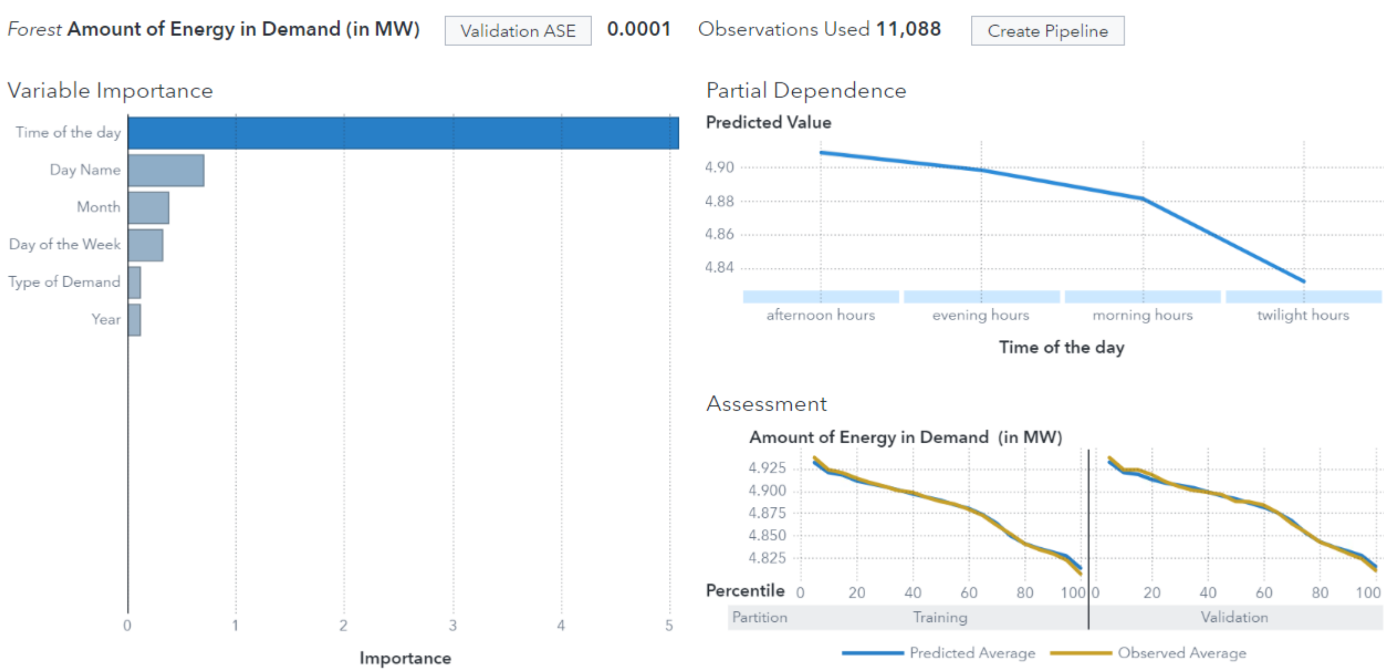

| Tuning | Parameters Tuned | Validation ASE | Training ASE |
|---|---|---|---|
| | Number of trees: 100, Bootstrap: 0.6, Min Value: 1, Max Branches:2, Max Levels: 6, Leaf Size: 5, Predictor Bins: 50 (Default) | 0.0002 | 0.0002 |
| | Number of trees: 50, Bootstrap: 0.8, Min Value: 1, Max Branches:2, Max Levels: 6, Leaf Size: 5, Predictor Bins: 50 | 0.0002 | 0.0002 |
| | Number of trees: 100, Bootstrap: 0.6, Min Value: 1, Max Branches:3, Max Levels: 6, Leaf Size: 5, Predictor Bins: 50 | 0.0001 | 0.0001 |
| | Number of trees: 100, Bootstrap: 0.6, Min Value: 1, Max Branches:2, Max Levels: 12, Leaf Size: 10, Predictor Bins: 40 | 0.0001 | 0.0001 |
| Details | After tuning the parameters, I found that the best model is the one with Number of trees: 100, Bootstrap: 0.6, Min Value: 1, Max Branches:3, Max Levels: 6, Leaf Size: 5, Predictor Bins: 50, which gave me the best Validation and Training ASE. Both my training and validation ASE are the same, signifying that my model is not overfitted. As the validation is lower than 0.05, my model has met my success criteria. | | |

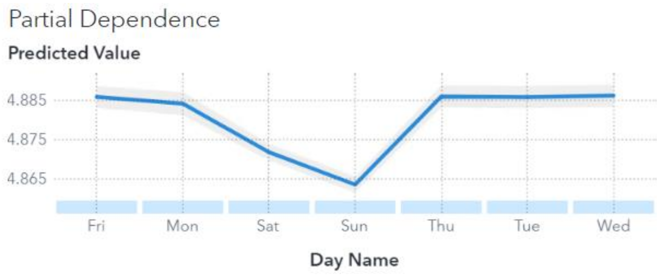
## Evaluation




The best model using the evaluation model selected the Random Forest Model as the best model compared to Decision Tree and Linear Regression, as it has the lowest validation ASE value compared to them. The time of the day variable is the most important factor for all the 3 models, as it has the highest relative importance. Other influential factors such as “Type of Demand” and “Day Name” hold importance, with their relative contributions closely after the pivotal “Time of the Day.”

## Best Model






The variable importance chart highlights Time of the day as the most significant variable followed by day name and month, among others. Analyzing peak energy demand timings reveals the highest predicted amount during the afternoon and evening hours, while the lowest demand occurs during twilight hours. Examining day names, Wednesday exhibits the highest energy demand, while Sunday records the lowest.

# Conclusion

## Energy

In conclusion, the objective of predicting peak energy demand for electricity in Singapore to enhance environmental monitoring and resource management was achieved. The analysis emphasizes the importance of variable such as Time of the day, day name and month in influencing energy demand patterns. Notably peak energy demand was projected to occur during afternoon and evening house with Wednesday showing the highest demand and Sunday the lowest.  Thus, accurate peak energy demand prediction based on the time of days and day empowers Singapore to prioritize eco-friendly energy generation during afternoon hours, optimizing environmental monitoring and resource management. Additionally, considering factors like the month further refines strategy, enabling the prioritization of specific days for eco-friendly energy initiatives.

## Overall

The results gained from food, plastic, energy and carbon emissions after developing predictive models will help assist our stakeholders that are focusing on improving sustainability by reducing food waste, energy, plastic usage and carbon emissions to gain valuable insights into the consumption trends and also the production of carbon emissions, allowing them to make informed decisions and drive sustainability efforts in community organizations, non-profits or grassroots initiatives regarding to effective resource reduction measures through our predictive analysis.

## Tools Used

| Description | Image |
|---|---|
| **Tableau** | |
| **PowerBI** | |
| **SAS** | sas |