IEOR 4578: Forecasting Project

Energy Load Prediction

Source: https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

Authors: Ishita Pundir (ip2441), Saum Kothari (sbk2171), Tushar Bura (tb3077)

Introduction

The realm of electricity consumption analysis has witnessed significant transformations with the integration of machine learning techniques. This project embarks on a comprehensive exploration of machine learning and statistical methodologies to forecast electricity consumption patterns for a diverse set of clients. The core aim is to develop robust predictive models that can accurately anticipate future electricity loads based on historical consumption data. These models are envisioned to serve as indispensable tools for energy providers, enabling them to streamline resource allocation, optimize infrastructure investments, and enhance overall operational efficiency.

Our project is designed to enhance electricity consumption forecasting by creating sophisticated time series models, utilizing advanced analytics techniques, and drawing from the UCI Machine Learning Repository: Electricity Load Diagrams 2011-2014 dataset, a comprehensive repository of electricity consumption records spanning from 2011 to 2014. This dataset encompasses consumption data from 370 points or clients, offering a rich landscape for analysis and model development. Each data point represents electricity consumption in kilowatts (kW) measured at 15-minute intervals, allowing for detailed insights into consumption patterns and trends.

Variable Information:

- **Time stamps**: Recorded in the format 'yyyy-mm-dd hh:mm:ss', providing precise temporal granularity for analysis.
- **Client consumption**: Each column corresponds to a specific client, with consumption values represented in kilowatts (kW).

We are committed to establishing a standardized and robust forecasting methodology for electricity consumption that serves a global customer base. Our approach is designed to consistently deliver predictive accuracy and operational efficiency, supporting sustainable energy management and growth in a dynamic market landscape. We will learn the pattern of the electricity consumption data in the past to forecast the future consumption.

Modifications & Process Flow

The project undertook several key steps:

Data Preprocessing:

- The initial step involved cleansing the data to remove inaccuracies or irrelevant information.
- The 'Time' column values were converted into datetime format to facilitate easier manipulation and analysis.
- Data was aggregated to obtain daily consumption values, simplifying the analysis by reducing the granularity of the data.

Analysis:

- Clustering was applied to identify distinct patterns or groups within the data based on similarities in electricity usage and select a representative example from each cluster for further in-depth analysis.
- Cubic Splines and Piecewise Linear fits were employed to enhance the visualization and interpretation of data trends, *which was not done previously*.

Exploratory Data Analysis (EDA):

We introduced the use of **Tableau** to create a comprehensive dashboard for this crucial step in our project. This enhancement was made to provide clearer and more accessible insights into the data distribution, identifying trends, patterns, and periodicity between different variables. *The addition of Tableau as a visualization tool* allows us to present complex data in a more intuitive format, making it easier for non-technical audiences to understand the findings. This strategic integration aims to not only refine our analytical processes but also improve communication and presentation of the results, facilitating broader engagement and decision-making.

Feature Engineering:

The data was transformed to extract meaningful attributes that could be crucial for the effectiveness of the forecasting models.

Model Development:

Time-series analysis techniques, such as **TFT, Random Forest** were employed in order to get better forecasting results *in addition to the previously used* **ARIMA** (AutoRegressive Integrated

Moving Average) and **SARIMA** (Seasonal ARIMA), **Facebook Prophet**, which were *run again with the new clusters* created. These models are designed to account for both trend and seasonality in the data, making them suitable for predicting future values based on historical data.

Finally, we evaluated the accuracy and efficacy of each model by employing the Mean Absolute Percentage Error (MAPE) as the metric to compare our predictions for 2014's electricity usage against the actual observed data.

Business Problem Definition

The primary business problem addressed by this project revolves around the imperative need for accurate forecasting of electricity consumption. This encompasses predicting the volume of electricity demand across various points, understanding consumption trends, and efficiently managing energy resources to meet demand while minimizing wastage. Accurate forecasting is pivotal for energy providers to optimize grid operations, plan infrastructure upgrades, and ensure reliable energy supply to consumers.

Value Creation

Through the utilization of advanced machine learning techniques such as regression and clustering, coupled with time-series analysis methods such as ARIMA, SARIMA, Facebook Prophet, Random Forest, TFT, the project endeavors to create a predictive framework capable of capturing the intricacies of electricity consumption patterns. The envisioned value creation spans across multiple dimensions:

- **Enhanced Grid Management:** Accurate forecasting facilitates proactive grid management, enabling energy providers to anticipate demand fluctuations and allocate resources optimally.
- **Infrastructure Planning:** Insights derived from forecasting models aid in infrastructure planning, allowing for targeted investments in grid upgrades and maintenance.
- **Energy Efficiency**: By identifying consumption patterns and trends, energy providers can implement targeted initiatives to promote energy efficiency and reduce wastage.
- **Cost Optimization**: Optimized resource allocation and efficient grid management translate into cost savings for energy providers and consumers alike.

Need for Forecasting Analysis

In the dynamic landscape of energy distribution, the ability to forecast electricity consumption accurately is paramount. Forecasting analysis addresses this need by providing insights that enable energy providers to adapt swiftly to changing demand patterns, mitigate risks associated with supply shortages or surpluses, and enhance overall operational resilience. Moreover, accurate forecasting fosters a more sustainable energy ecosystem by promoting efficient resource utilization and reducing environmental impact.

After finalizing the business problem definition and a final evaluation we recommend the:

Random Forest Regressor

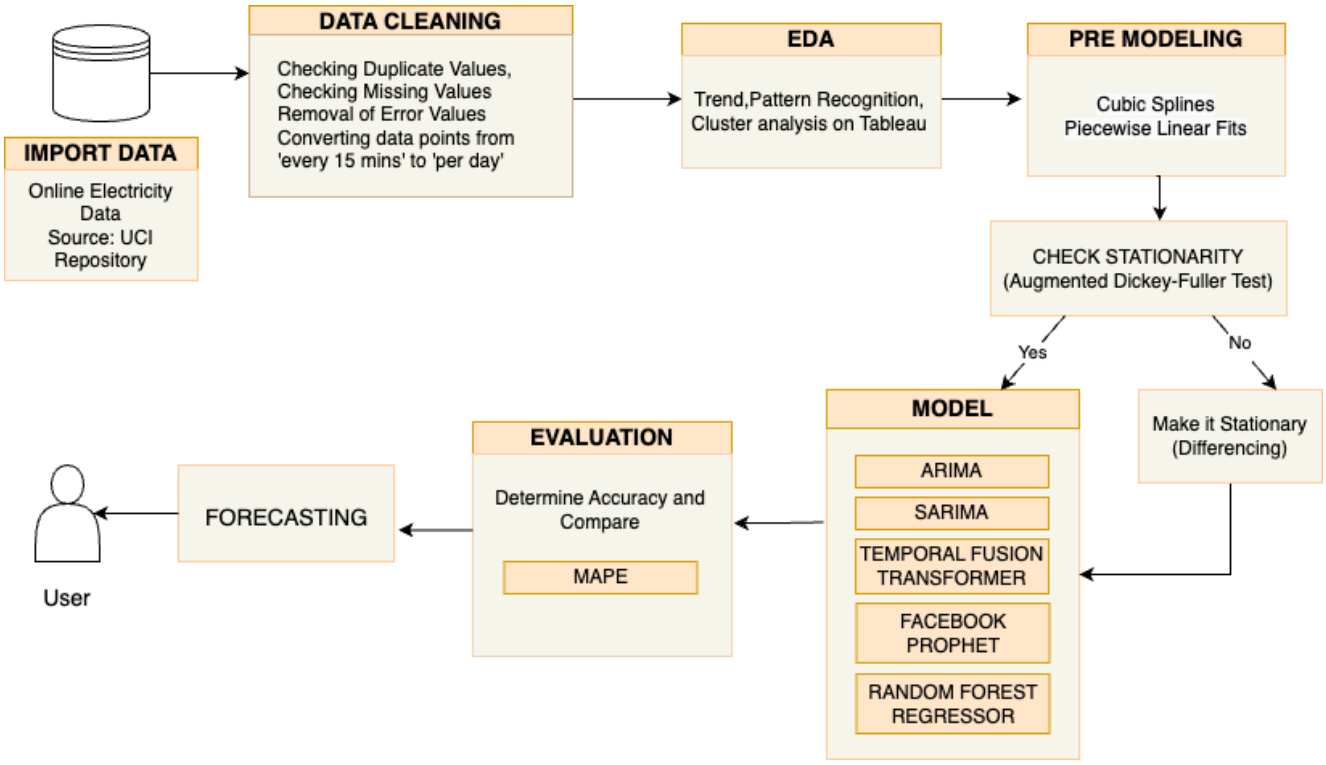
We will see how later in the report and slide deck, according to evaluation metrics, it provides the best result. It also provides advantages such as:

- Robustness to Overfitting: Random Forest is naturally resistant to overfitting due to its ensemble approach, which involves averaging multiple decision trees.
- Handling Non-linear Relationships: Random Forest can effectively model complex, non-linear relationships in data, making it suitable for diverse datasets.
- Feature Importance: Random Forest provides insights into which features most significantly impact predictions, aiding in better understanding and optimization of the model.

We also notice that the **Temporal Fusion Transformer (TFT)** performs very well, and can be used on other types of datasets and results can be compared. It provides benefits such as:

- Temporal Patterns: TFT excels in capturing complex temporal relationships and patterns, leveraging past data points and contextual information.
- Multivariate Capabilities: TFT can handle multiple input features and output variables, making it ideal for scenarios where multiple related forecasts are needed simultaneously.
- Flexibility and Adaptability: TFT can incorporate static and known future inputs, enhancing its forecasting accuracy and adaptability to different forecasting scenarios.

Process Flow



Data Extraction

The electricity consumption data was shared by Artur Trindade in 2015. We downloaded it from UC Irvine Machine Learning Repository with the link: https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014.

The downloaded file is in TXT format, and Anaconda and Jupyter notebook was utilized for storing, processing, and modeling the data.

The project began by importing necessary Python libraries for data manipulation and visualization:

- Pandas: A powerful data analysis and manipulation tool, which is a cornerstone for data science in Python. It provides data structures like DataFrames that make it straightforward to perform operations on tabular data.
- NumPy: This library adds support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays.
- Matplotlib: A plotting library that is very flexible and can generate a wide range of static, animated, and interactive visualizations.
- Seaborn: Built on top of Matplotlib, Seaborn is a statistical data visualization library that provides a high-level interface for drawing attractive and informative statistical graphics.

Data Processing

Data Overview

The data set contains electricity consumption of 370 clients from 2011 to 2014. All time labels report to Portuguese hour. Some clients were created after 2011. In these cases, their consumption was considered zero. The values of electricity consumption are in kW of each 15 min. To convert values into kWh, values must be divided by 4. Every year on March time-change day (which has only 23 hours) the values between 1:00 am and 2:00 am are zero for all points. Every year on October time-change day (which has 25 hours) the values between 1:00 am and 2:00 am aggregate the consumption of two hours.

We observed that:

1. The dataset was saved as txt using csv format, using semicolon (;).
2. The First column present date and time as a string with the following format 'yyyy-mm-dd hh:mm:ss'
3. The decimal point in values were replaced by a comma.
4. Energy consumption values have units in kW.

Data Diagnostics

A series of quality checks are performed on the data extracted. These checks included:

1. Number of records:

   The data has 140256 electricity consumption of each 15 min from 2011 to 2014 for each of 370 clients.

2. Duplicate records if any:

   Clients may have the same electricity consumption at different 15 min intervals.

3. Missing values in relevant fields:

   There are no missing values, but only 158 clients have valuable records for the whole period (other clients created accounts after 2011).

4. Sum of a numeric field:

The electricity consumption ranges from 0 to 192800 kW.

5. Data period confirmation:

The electricity consumption data starts from 2011-01-01 00:15:00 to 2015-01-01 00:00:00.

Modeling Data Creation

For each of the customers, the modeling data are collected and processed using the following steps. The below also addresses the feature engineering section.

1. Replace comma to decimal and convert values stored from string to float.
2. Aggregate electricity consumption to change the record of every 15 minutes to every day, and convert units from kW to kWh by dividing current values by 4.
3. Since 158 clients have created accounts since 2011, 162 clients created accounts from 2011 to 2012, and 50 clients created accounts after 2012, we select clients joined in 2011 as representatives and use their data to train and test models.
4. Take care of the change of time date as needed for different models.
5. Pay attention to time related features such as year (2011, 2012, 2013, 2014), quarter (1-4), month (1-12), weekday (0-6), day (1-31), hour (0-23) as needed for different models.
6. Consider holidays based on Portugal's official holidays.
7. Identify client groups based on the their similarity (more in modeling part)

Target Variables

The target variable, also known as the dependent variable, is the primary metric that a predictive model aims to forecast or predict. It is the outcome interest, whose future values are to be determined based on other variables within the dataset. In predictive modeling, the target variable is what the model is trained to predict.

For our project, the target variable is **clients' daily electricity consumption** in kWh in 2014.

This has been chosen as the target variable because it directly reflects the energy usage patterns critical for operational planning and efficiency optimization. Analyzing this variable helps in understanding consumption trends and peak demand periods, which are essential for managing energy supply, improving cost-efficiency, and supporting sustainability initiatives. Furthermore, accurate forecasting of electricity consumption can aid utility companies in capacity planning and reducing the environmental impact of energy production.

Predictive Variables

A predictive variable is a variable used in algorithmic solutions to predict the target variable. During our analysis, we categorized predictive variables into two categories:

1. Direct variable - These variables were directly from the dataset that was provided by direct customers
2. Derived variable - These variables were created by manipulating the direct variables

Specifically, our direct variables include clients' electricity consumptions in kWh in year 2011, 2012 and 2013, and our derived variables include time related features such as year (2011, 2012, 2013), quarter (1-4), month (1-12), weekday and weekend(0-6), day (1-31), hour (0-23). Moreover, the Facebook Prophet model incorporates an additional derived variable to account for holidays.

Pre-modeling

The previous project used a strategy where the dataset consisting of 370 clients' electricity usage was segmented into two distinct clusters, based on the fact only a part of the observations start from 2011. However, in our project, *we have decided to* ***expand the number of clusters*** *from two to three*. This decision was driven by further analysis which revealed that a three-cluster model provided a superior **silhouette score** compared to the previous two-cluster model. By increasing the number of clusters, we aim to capture more nuanced variations in electricity consumption patterns among the clients, which were potentially overlooked in the earlier model. This refinement in the clustering approach is anticipated to enhance the granularity of our analysis, leading to more accurate and tailored predictions for each subgroup, thereby improving the overall efficacy and precision of our forecasting models.

Clustering

This approach not only optimizes time and memory usage but also enhances the manageability of our dataset, as, instead of training individual models on all client data, which is resource-intensive, we clustered the data and then selected a representative time series from each cluster. This method ensures that our training dataset embodies the essential characteristics of each cluster without the redundancy of using all series.

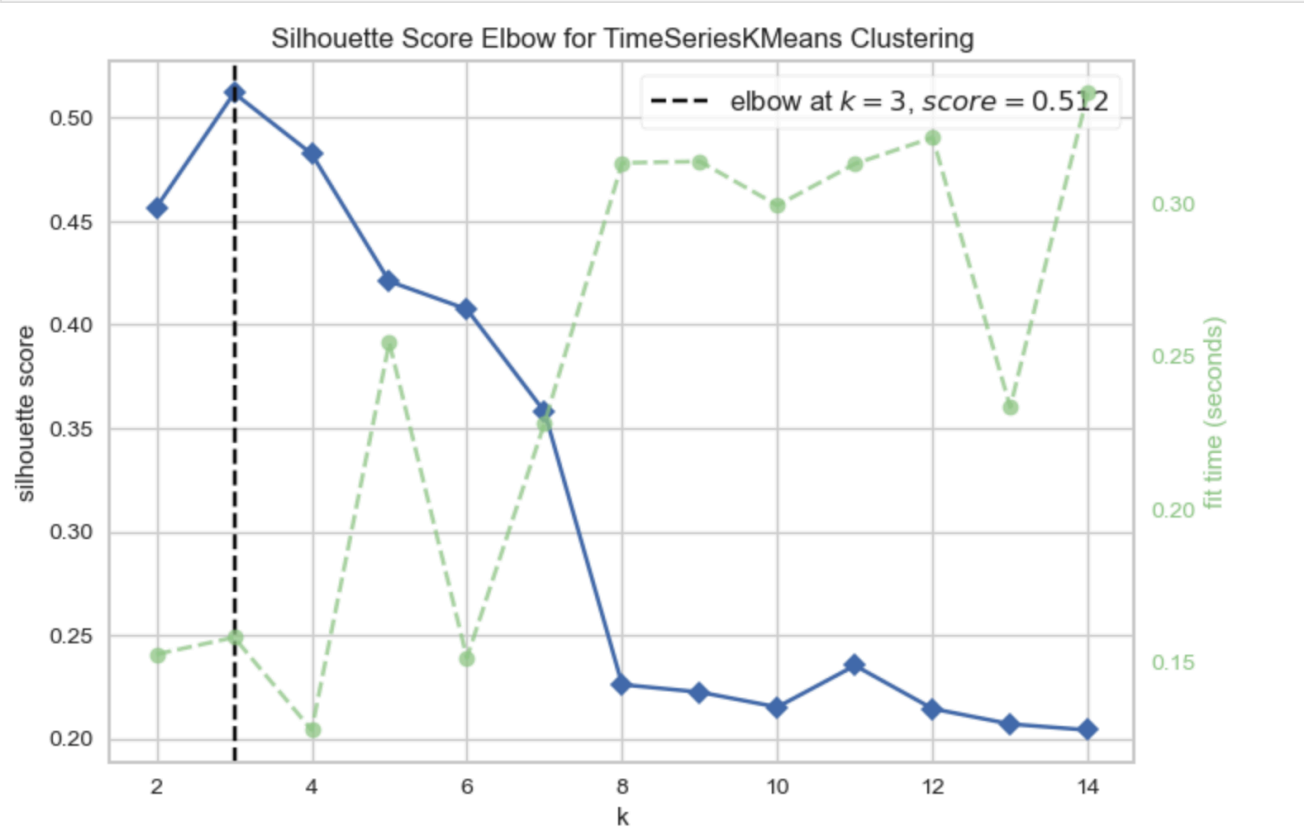
Applying K-Means Clustering: We utilized the K-means clustering algorithm to segment the data from 370 clients into distinct groups. This method partitions the clients into k distinct clusters by minimizing the variance within each cluster. For our project, we chose to identify three clusters based on trial and error optimization and evaluation of cluster validity using the silhouette score.

Silhouette Score Evaluation: The silhouette score was used as a metric to assess the quality of the clustering. This score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. After testing various cluster counts, we found that three clusters yielded the highest silhouette score, indicating a clear distinction between the three groups.

Selection of Representative Clients: From each cluster, we selected one representative client, MT_002 from cluster 0, MT_181 from cluster 1, and MT_001 from cluster 2, based on their comprehensive data records spanning from 2011 to 2014. These clients were chosen because their data profiles were representative of their respective clusters' overall patterns, making them ideal subjects for detailed analysis and model training.

This structured approach to data clustering and representative selection is designed to streamline the modeling process, enabling more focused and efficient analysis while maintaining the

robustness and generalizability of our predictive models.

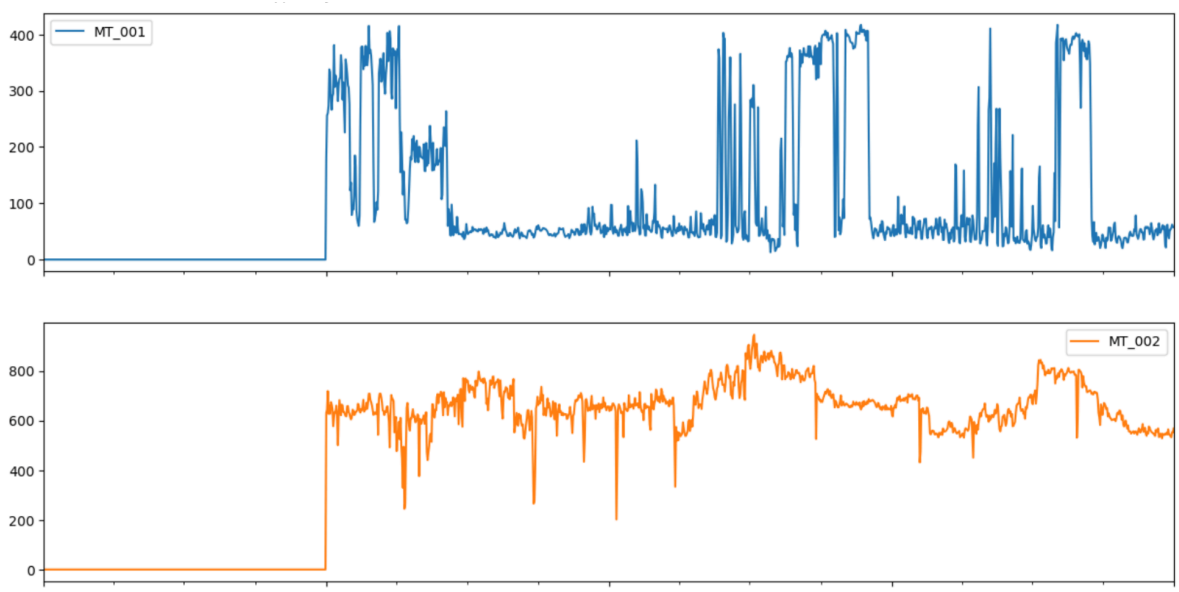


Training, Validation & Testing Split

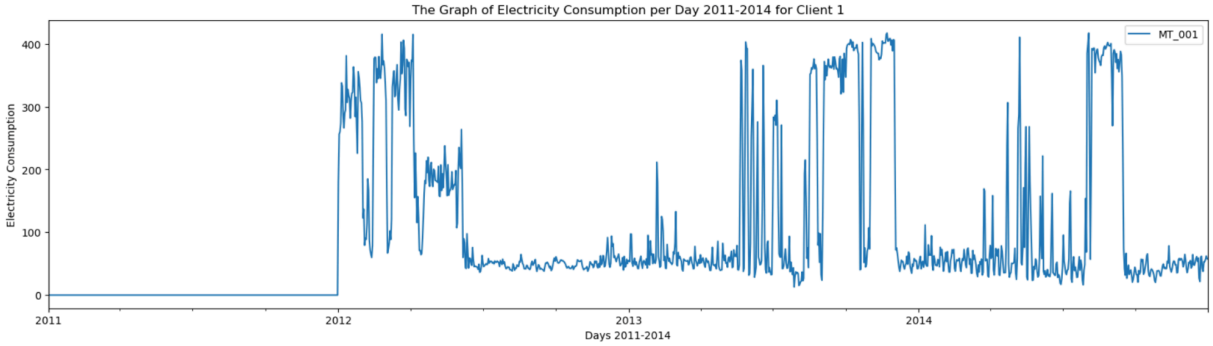
We strategically partitioned the time series data, starting from 2011, into distinct subsets for training, validation, and testing, *instead of just the train-test split done before*. This was designed to optimize the model's learning and generalization capabilities, allocating 70% of the data for training, 10% for validation, and 20% for testing. This allows us to train the models on a substantial portion of the data, fine-tune the model parameters using the validation set, and finally assess the model's performance on the unseen test data, ensuring a robust evaluation.

Analysis

For an in-depth examination of daily electricity consumption, we meticulously prepared and visualized the data using a structured approach. Initially, we replicated our daily data into a new DataFrame to safeguard the integrity of the original data during the visualization process. The 'Time' column was then converted to datetime format, enhancing its compatibility with time series analysis tools and techniques. After setting this column as the index for the DataFrame, we generated individual plots for the first five series of daily consumption data. This method not only highlighted the variability and periodicity in electricity consumption for each series but also enabled a visual assessment of similarities and differences across various datasets, crucial for further analysis and modeling.

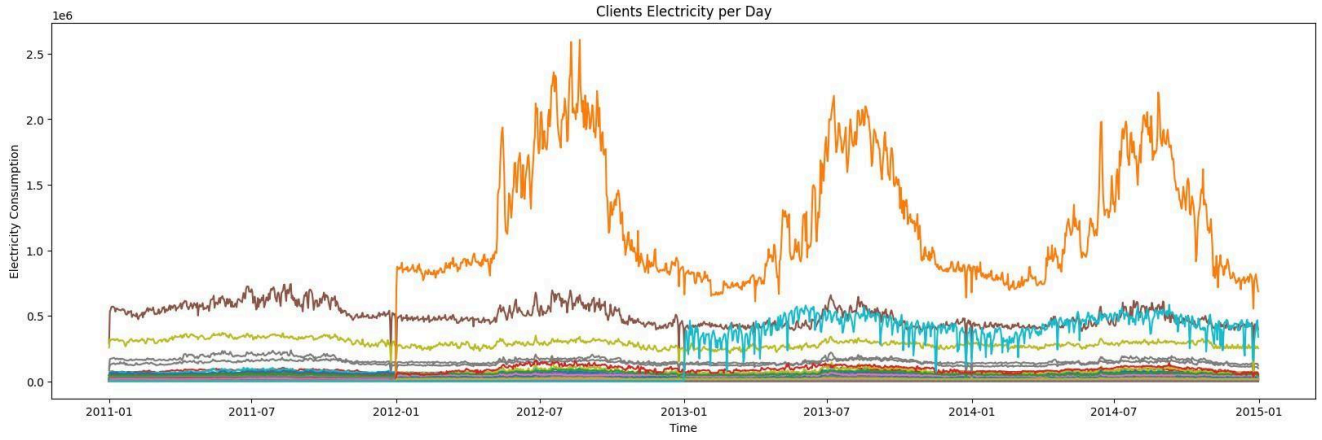


Analysis of Daily Electricity Consumption for Client 1 (2011-2014)



This graph illustrates the daily electricity consumption of Client 1. As depicted, the electricity usage shows significant variability over the four years. Several peaks indicate periods of high consumption, which could be attributed to seasonal demands, operational cycles, or special events affecting usage. A noticeable trend is the cyclical nature of the spikes, suggesting a possible seasonal pattern in electricity demand. The data's volatility might necessitate deeper analysis using time-series forecasting models to predict future consumption patterns and to understand underlying factors driving these changes. This could help in optimizing energy usage and planning for demand variations throughout the year.

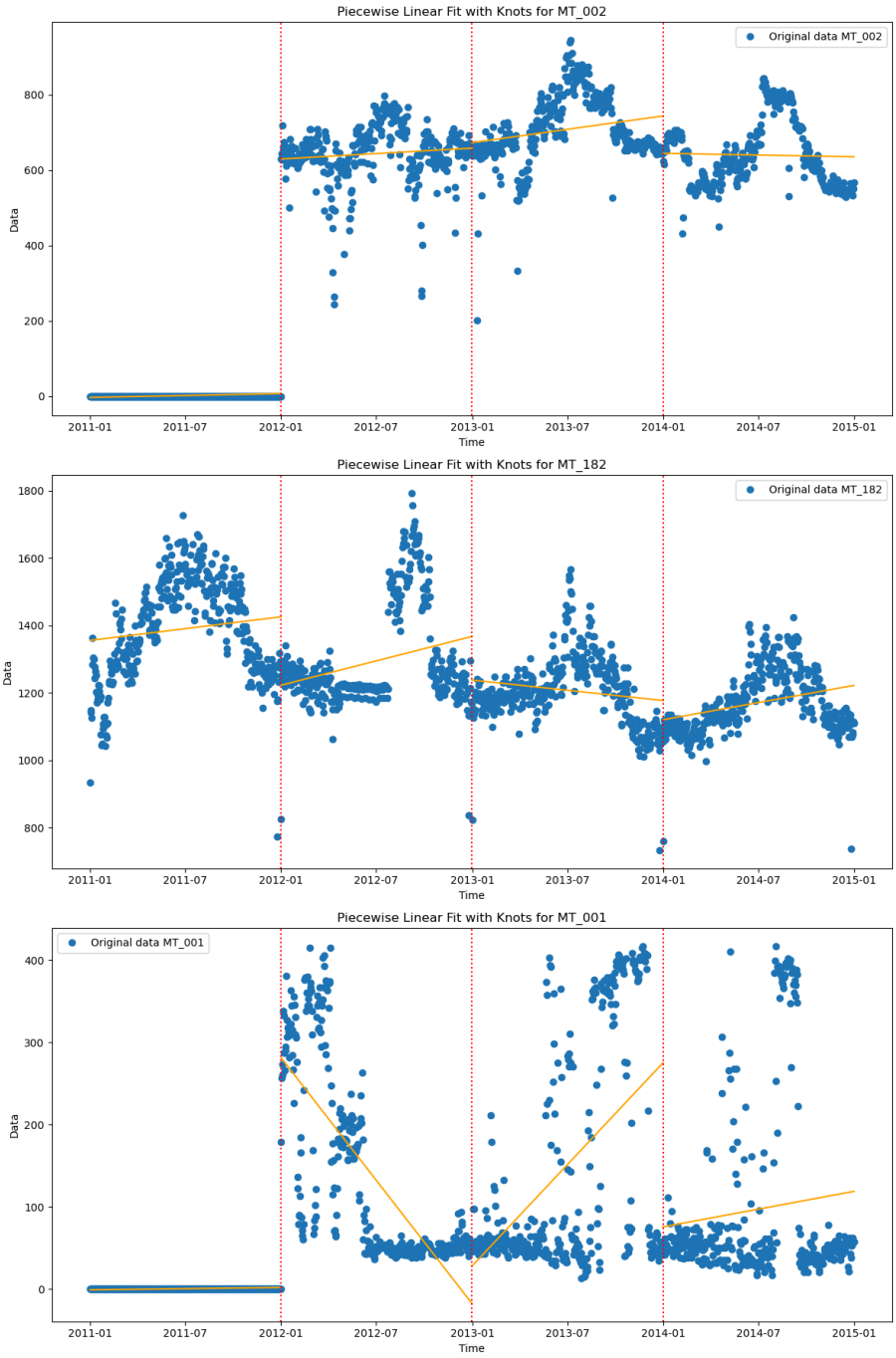
Overall Data Visualization



The graph illustrates the daily electricity consumption for multiple clients from January 2011 to the end of 2014. It highlights significant variations in usage patterns, particularly emphasizing the seasonal fluctuations of one client who exhibits markedly higher consumption compared to others.

- Seasonal Variability: The dominant orange series shows pronounced peaks during the middle of each year, likely due to increased usage in the summer months, possibly for cooling purposes. This pattern is consistent across the years, indicating a predictable seasonal demand.
- Stable Lower Consumption: Other clients represented by different colors show much lower and stable consumption levels, lacking the seasonal spikes observed in the orange series.
- Anomalies: There are notable spikes and drops in the orange series around early 2012 and late 2013, suggesting potential anomalies or operational changes that could impact consumption.

(Further Analysis) Cubic Splines and Piecewise Linear Fits



MT_002

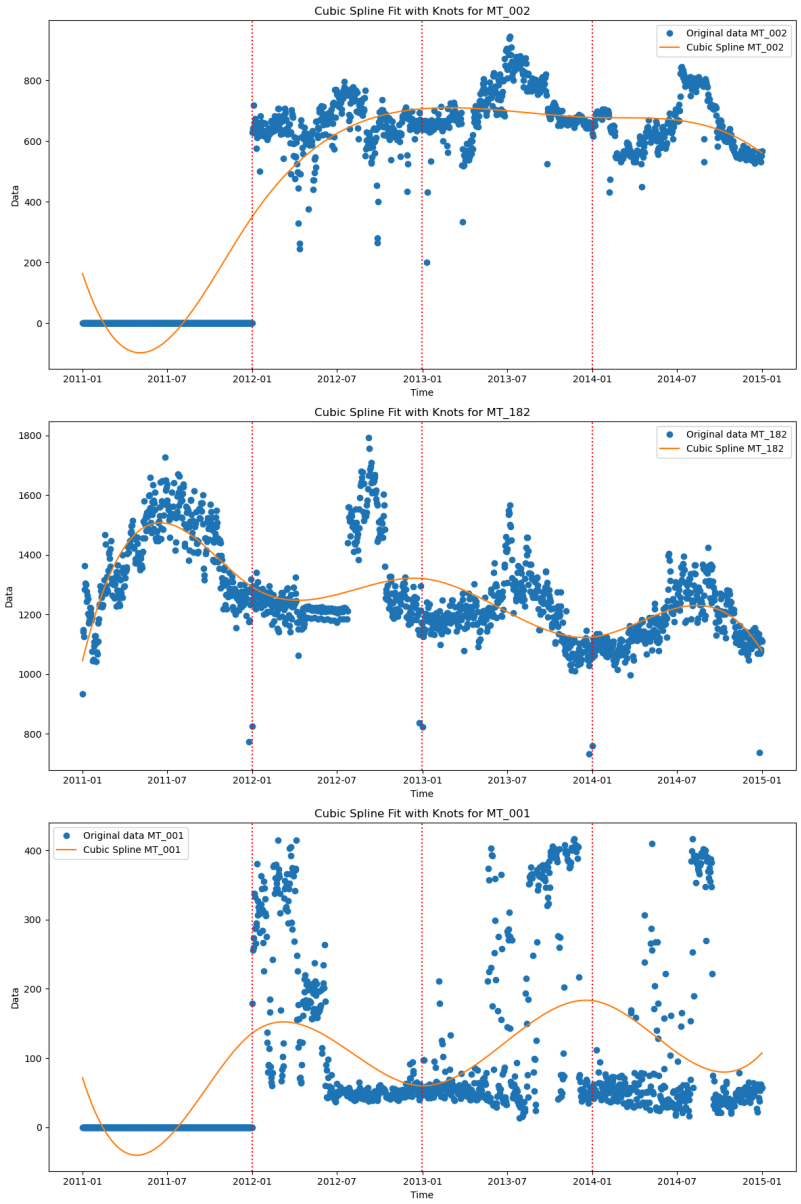
- **Trend**: Shows a consistent level of consumption with minor fluctuations.
- **Linear Fit**: The linear trend line is flat, indicating stable average electricity usage over the years, with knots placed at the start of each year.

MT_182

- **Trend**: Displays higher variability and a general increase in consumption.
- **Linear Fit**: Indicates a gradual increase in usage, especially noticeable in the middle segments, with annual knots suggesting yearly evaluations.

MT_001

- **Trend**: Exhibits significant variability and peaks, particularly in late 2012 and early 2013.
- **Linear Fit**: Shows a sharp rise followed by stabilization in consumption, with knots highlighting periods of significant change.



MT_002

- **Trend & Fit**: Exhibits moderate variability with an initial drop in consumption, followed by a rise and stabilization. The spline reflects this trend, rising in early 2012 and remaining relatively stable thereafter.

MT_182

- **Trend & Fit**: Shows more significant fluctuations with a peak around mid-2013, then a decline. The spline mimics this pattern, peaking and declining to reflect the real data closely.

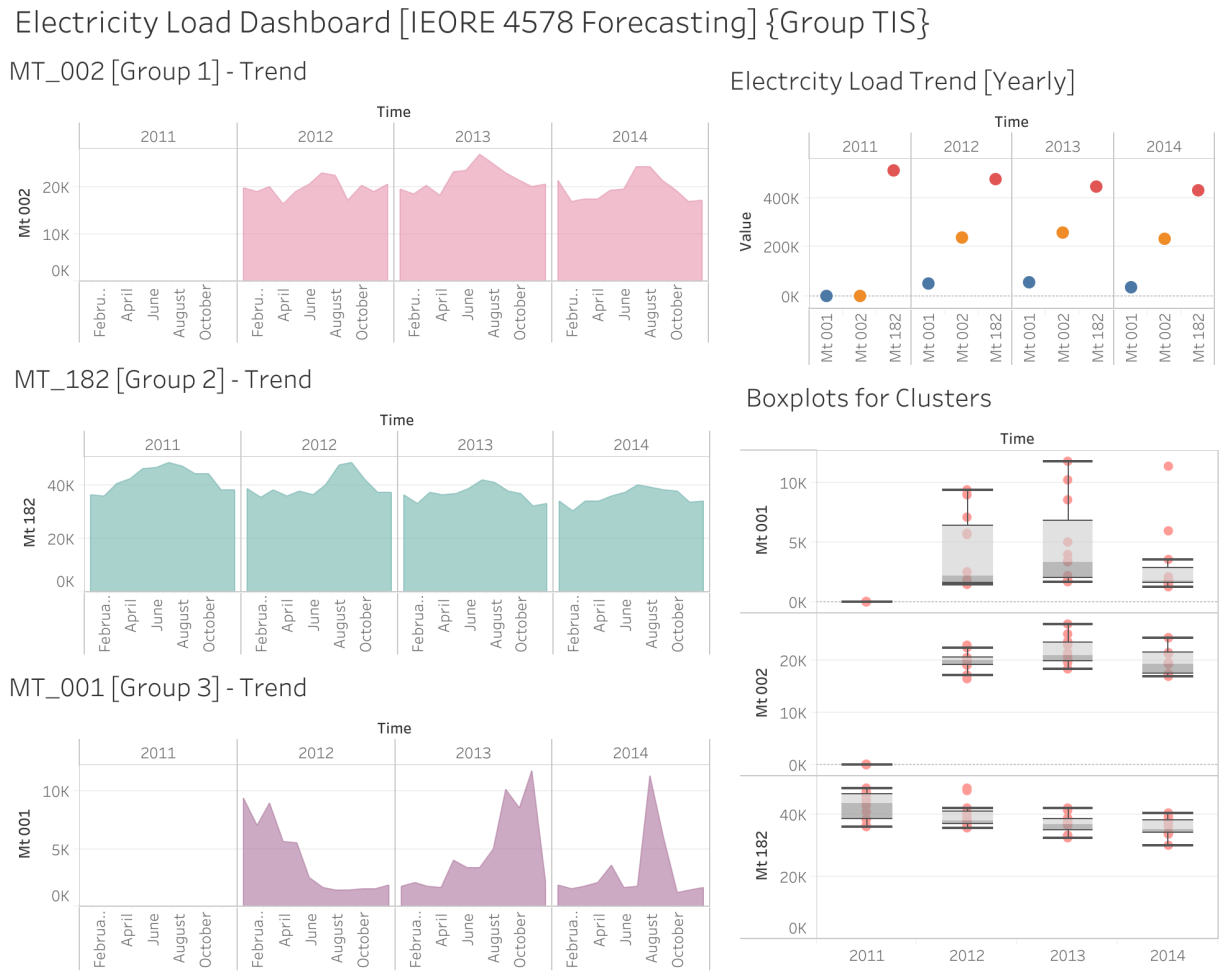
MT_001

- **Trend & Fit**: Lower consumption levels overall, with a notable spike in 2013. The spline is mostly flat but rises to model the 2013 spike, indicating a temporary increase.

Exploratory Data Analysis

As a ***new addition*** to our Exploratory Data Analysis, we created a comprehensive **Tableau** dashboard, in order to provide clearer and more accessible insights into the data distribution, identifying trends, patterns, and periodicity between different variables.

The "Electricity Load Dashboard" efficiently visualizes electricity consumption trends and variability across three client groups from 2011 to 2014. This is how it looks like:



This dashboard blends detailed time-series analysis with statistical summaries, offering a multi faceted view of electricity usage that is crucial for informed decision-making.

Trend Analysis by Client Group:

Each of the three areas (MT_002, MT_182, MT_001) showcases the monthly electricity consumption for a different client group over four consecutive years. The area charts use color to distinguish different years visually, making it easy to observe seasonal trends and year-on-year variations in consumption. For instance:

- **MT_002 (Group 1)**: Displayed in pink, this graph shows significant spikes and dips, indicating fluctuating electricity usage that might correspond to specific operational or seasonal patterns.
- **MT_182 (Group 2)**: Shown in green, this graph depicts a more stable pattern with less variance, suggesting consistent usage across months and years.
- **MT_001 (Group 3)**: In purple, this chart highlights dramatic fluctuations in specific years, particularly noticeable in late 2013 and 2014, potentially indicating external influences or changes in operational activities.

Electricity Load Trend (Yearly):

This visualization from the dashboard provides a nuanced analysis of annual electricity consumption across three client groups, captured through the use of color-coded dots that represent each year from 2011 to 2014. In examining MT_001 (Group 3), we observe a significant variation over the years, with a noticeable decrease in consumption from 2011 to 2012 followed by a sharp increase through to 2014, indicating potential operational changes or growth. MT_002 (Group 1) displays a more consistent upward trend in usage, suggesting gradual growth or expansion in operations. Meanwhile, MT_182 (Group 2) shows relatively stable or slightly fluctuating consumption, highlighting a steady operational demand across the years.

A comparative analysis between clients within the same years reveals key insights into operational demands and energy efficiency. In 2011, MT_182 exhibited the highest electricity consumption, significantly more than MT_001 and MT_002, possibly reflecting higher operational demands or inefficiencies. This pattern continues with MT_182 consistently showing higher consumption each year, although the gap narrows by 2014 as MT_001's consumption increases significantly, nearly matching MT_182's levels. The progression in MT_002's consumption, steadily increasing each year, positions it closer to MT_182's levels, particularly noted in 2013.

The year-on-year and client-to-client comparisons collectively indicate various operational behaviors and potential efficiencies or inefficiencies in energy use. MT_182's consistently high usage suggests a need for evaluating energy management practices, whereas the sharp increase in MT_001's consumption from 2012 to 2014 could signify operational changes that might require a shift in energy strategies to maintain efficiency. MT_002's steady increase implies a controlled

growth but still calls for continuous monitoring to ensure sustainable energy use. This detailed trend analysis is crucial for organizations to develop targeted energy conservation measures, optimize operational efficiencies, and plan for future capacity based on consumption patterns.

Boxplots for Clusters:

Central Tendency and Variability: The boxplots provide a clear visual representation of central tendency and variability in electricity consumption for each client group. The median of each boxplot, indicating the central point of data distribution, shows how the typical monthly electricity usage levels vary across different years and among client groups. For example, MT_001 displays a relatively stable median across the years, with only minor fluctuations, suggesting consistent usage patterns. However, the length and range of the interquartile range (IQR) vary, particularly in 2013 and 2014, indicating greater variability in these years. This suggests that while the central tendency remains stable, the spread of consumption values around the median has increased, possibly due to operational changes or seasonal impacts. Conversely, MT_002 and MT_182 demonstrate less variability in their IQR, indicating more consistent energy usage with fewer fluctuations month-to-month.

Outliers: Outliers in the boxplots represent months where electricity consumption was significantly different from the norm, highlighted by points lying outside the typical range (1.5 times the IQR from the box's edges). MT_001 shows numerous outliers across all years, suggesting occasional spikes or drops in electricity usage, which could be attributed to specific operational events or anomalies. MT_002 has fewer outliers, reflecting a more uniform consumption pattern with fewer extreme deviations from the average monthly usage. MT_182, despite its higher overall consumption, also displays a few outliers, especially in the earlier years, indicating sporadic peaks in demand that might be linked to particular events or operational demands.

Trend Analysis: Analyzing the trends over the years within each boxplot provides insights into broader operational or environmental changes. For instance, a visible upward trend in the median over successive years, as seen in MT_001 from 2012 to 2014, might indicate increasing consumption needs, possibly due to growth or scaling of operations. On the other hand, a stable or downward trend in the median could suggest improvements in energy efficiency or a reduction in operational scale. The positioning and movement of the whiskers and the width of the IQR also provide information about the consistency and reliability of the consumption patterns. A narrowing of the IQR over time would imply a tightening of consumption patterns, possibly due to enhanced operational controls or improvements in energy management practices.

Time Series Analysis

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.

Trend: In time-series analysis, the trend represents the long-term progression of the series. Trends can be upward, downward, or even sideways over time. A trend reflects the overall direction of the data points when looking at a graph over a long period. *Example*: If a company's sales have been increasing by an average of 5% per year over the last decade, this steady increase is the trend in the company's annual sales data.

Seasonality: Seasonality refers to periodic fluctuations that regularly occur in time-series data, which are predictable and repeat over a specific period, such as a day, week, month, or season. This is often due to factors like weather, holidays, and events. *Example*: Retail stores often experience seasonal patterns with spikes in sales during the Christmas holidays and other festive periods due to increased consumer purchasing.

Cyclicality: Cyclicality indicates the repeating patterns in time-series data that occur over variable periods, which are typically longer than a season. These cycles are not as predictable as seasonality because they don't have a fixed frequency. Economic cycles are a common example, where periods of expansion are followed by contractions over several years. *Example*: The business cycle, which consists of periods of economic growth (expansion), peak, contraction (recession), and recovery, shows cyclical behavior in economic time-series data such as GDP or employment rates. These are influenced by broader economic factors and do not follow a fixed calendar schedule.

Stationarity *(new addition):*

In a time series data, it refers to the statistical property that the mean, variance, and autocorrelation structure of the series do not change over time. In simpler terms, a stationary time series retains its behavior and structure over time, which makes it predictable and hence easier to model for forecasting purposes. Most time series modeling techniques and forecasts are based on the assumption that the time series is stationary. If the time series exhibits trends, seasonality, or cyclicality, these can affect the stability of its statistical properties over time, and the series is said to be non-stationary. Non-stationarity can lead to unreliable and misleading models and forecasts.

As this is an extremely important check before doing time series forecasting, we have ***additionally done*** this, using the **Augmented Dickey Fuller** test:

According to the paper titled *"Time-series forecasting of seasonal items sales using machine learning – A comparative analysis" by Yasaman Ensafi, Saman Hassanzadeh Amin, Guoqing Zhang, and Bharat Shah, as published in the International Journal of Information Management Data Insights in 2022,* the objective of Augmented Dickey Fuller (ADF) test is to decide that the time-series is stationary or non-stationary by checking the presence of unit root in a time-series. This method observes the difference between the value level and the mean. If it was higher than the mean, the next movement will be downward. Furthermore, if it was lower than the mean, the movement would be upward.

$$\Delta y(t) = \lambda y(t-1) + \mu + \beta t + \alpha_1 \Delta y(t-1) \pm \ldots + \alpha_k \Delta y(t-k) + \varepsilon_t$$

The equation above explains these value changes, where μ is a constant, β is the coefficient on a time trend, k is the lag order of the autoregressive process, and $\Delta y(t)$ can be defined as

$\Delta y(t) \equiv y(t) - y(t-1)$, $\Delta y(t-1) \equiv y(t-1) - y(t-2)$ (Corrius, 2018)

The null hypothesis states that the time-series in non-stationary ($\lambda = 0$). If this hypothesis is rejected, it shows that the next movement ($\Delta y(t)$) is not just a random value, and it depends on the current level $y(t-1)$. Thus, the time-series is stationary. In our case, the p-value is smaller than 0.05, and the time-series is stationary.

Before:

```
{'ADF Statistic': -1.977277519802678,
 'p-value': 0.2966241133017994,
 'Used Lag': 5,
 'Number of Observations': 1455,
 'Critical Values': {'1%': -3.4348523191002123,
  '5%': -2.8635284734563364,
  '10%': -2.567828646449617},
 'Information Criterion': 14781.54274843259}
```

```
{'ADF Statistic': -2.5974121424699415,
 'p-value': 0.09353330542838084,
 'Used Lag': 9,
 'Number of Observations': 1451,
 'Critical Values': {'1%': -3.4348647527922824,
  '5%': -2.863533960720434,
  '10%': -2.567831568508802},
 'Information Criterion': 15475.92929776032}
```

```
{'ADF Statistic': -3.9914120032297173,
 'p-value': 0.0014561541869830437,
 'Used Lag': 16,
 'Number of Observations': 1444,
 'Critical Values': {'1%': -3.434886677803751,
  '5%': -2.8635436366589673,
  '10%': -2.5678367211155533},
 'Information Criterion': 14684.892179778903}
```

After Differencing df1 and df2:


```
{'ADF Statistic': -23.205827424913217,
 'p-value': 0.0,
 'Used Lag': 4,
 'Number of Observations': 1455,
 'Critical Values': {'1%': -3.4348523191002123,
  '5%': -2.8635284734563364,
  '10%': -2.567828646449617},
 'Information Criterion': 14774.336351217129}
```

-

```
{'ADF Statistic': -16.39981253296912,
 'p-value': 2.644850947519408e-29,
 'Used Lag': 9,
 'Number of Observations': 1450,
 'Critical Values': {'1%': -3.4348678719530934,
  '5%': -2.863535337271721,
  '10%': -2.5678323015457787},
 'Information Criterion': 15471.19372295788}
```

-

Modeling

For each client group we identified, we chose representative client data to train and test the model. Specifically, we use client “MT_002” for cluster 1, client “MT_182” for cluster 2, and client “MT_001” for cluster 1, as figured from K-Means clustering.

We used the previously designed ARIMA, SARIMA, and Facebook Prophet models but on the newer set of clusters. In addition to that, we have also **employed two new models - Temporal Fusion Transformer (TFT) and Random Forest Regressor** - and evaluated model results using MAPE.

In the modeling phase, we have employed the following models -

- ARIMA
- SARIMA
- Facebook Prophet
- Random Forest Regressor
- Temporal Fusion Transformer (TFT)

on the electricity load dataset to forecast future energy consumption trends. Below, we outline the model description, approach, results and evaluation of the models used.

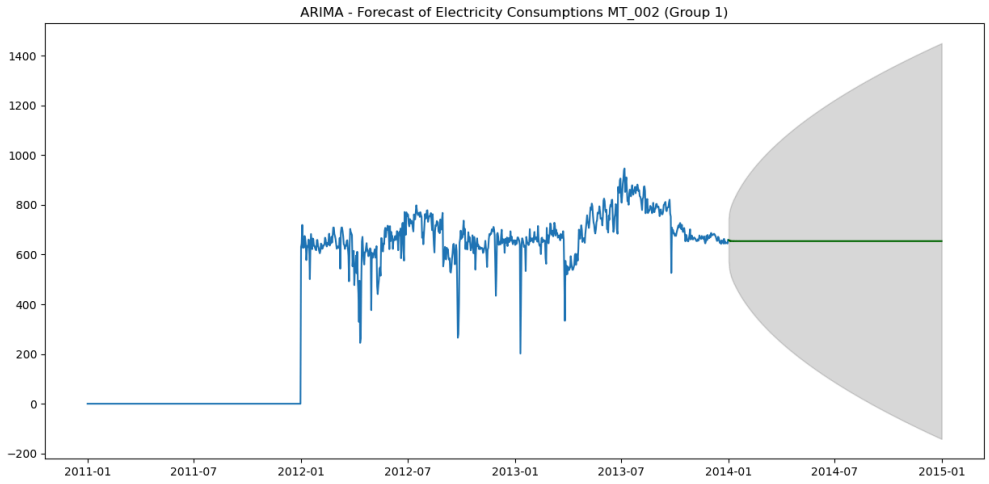
ARIMA (AutoRegressive Integrated Moving Average) Model:

- **Model Description:**
  The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series forecasting technique that captures linear relationships and patterns within a time series. It consists of three main components: autoregression (AR), differencing (I), and moving average (MA). The AR component models the relationship between an observation and a number of lagged observations, the differencing component removes trends from the time series, and the MA component models the relationship between an observation and the residual errors from a moving average model applied to lagged observations. ARIMA is well-suited for stationary time series data.

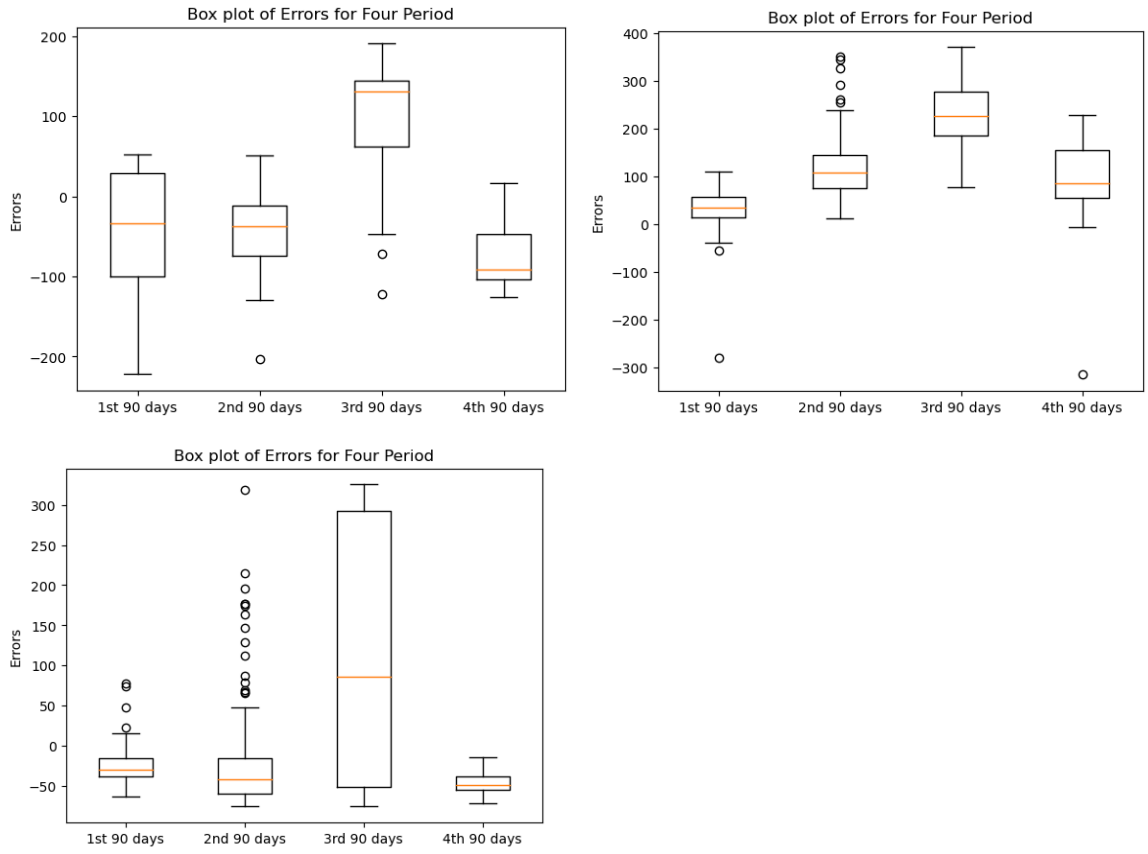
  Rationale behind using:
  - ARIMA is a simple yet effective model for forecasting time series data.
  - It can capture linear relationships and patterns in the data.
  - ARIMA models can be easily interpreted and understood.

- **Approach:**
  We applied the auto_arima function from the pmdarima library to automatically select the optimal parameters for the ARIMA model. We performed stepwise search to minimize the Akaike Information Criterion (AIC) and chose the best-fitted ARIMA model based on the lowest AIC value. Additionally, we added an exogenous variable representing the day index to the model to incorporate any daily patterns or seasonality.

- **Results:**
  The ARIMA model achieved a Total Mean Absolute Percentage Error (MAPE) of 11.94% on the test dataset. The MAPE for the four 90-day periods ranged from 8.25% to 14.47%. The box plot of errors showed relatively consistent performance across different time periods, with no significant outliers.



- **Performance Evaluation:**
  The ARIMA model performed reasonably well in forecasting electricity consumption for Cluster 1 (MT_002) and Cluster 2 (MT_182), with MAPEs of 11.94% and 10.06% respectively. However, for Cluster 3 (MT_001), the model exhibited poor performance with a very high MAPE of 101.34%. This suggests that the ARIMA model may not be suitable for capturing the complex patterns and dynamics present in Cluster 3's electricity consumption data. Further investigation is needed to understand the reasons for the model's poor performance and explore alternative modeling approaches.



SARIMA (Seasonal AutoRegressive Integrated Moving Average) Model:

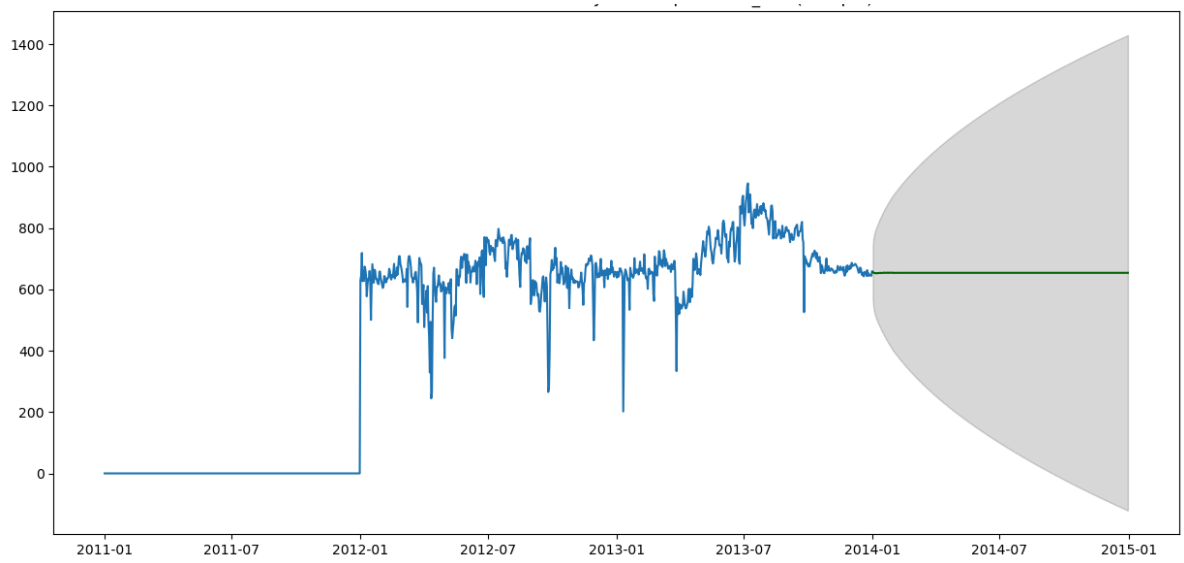
- **Model Description:**
  The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA model to account for seasonal patterns in time series data. In addition to the ARIMA components, SARIMA includes seasonal parameters that capture recurring patterns over fixed intervals. SARIMA is effective for modeling time series data with both trend and seasonality.

  Rationale behind using:
  - SARIMA can capture both linear and seasonal patterns in time series data.
  - It extends the capabilities of the ARIMA model to handle seasonal variations.
  - SARIMA models are relatively easy to interpret and implement.

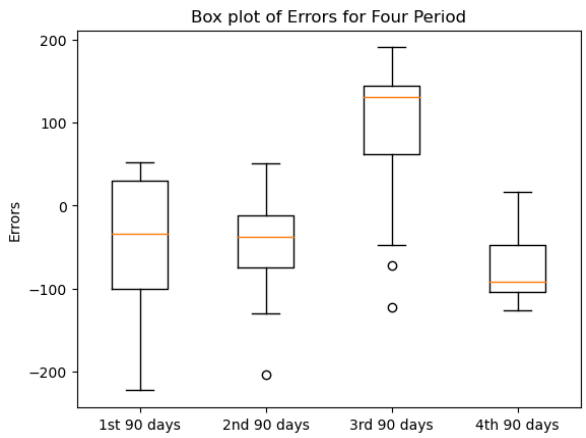
- **Approach:**
  Similar to the ARIMA model, we used the auto_arima function to automatically select the optimal parameters for the SARIMA model. We performed stepwise search to minimize the AIC and selected the SARIMA model with the lowest AIC value.

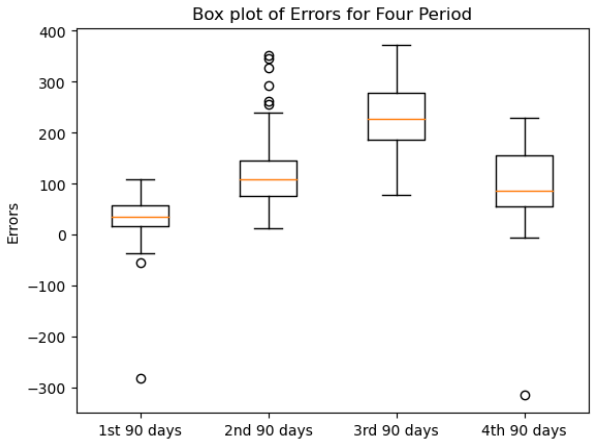
Additionally, we included an exogenous variable representing the day index to capture any daily patterns or seasonality.

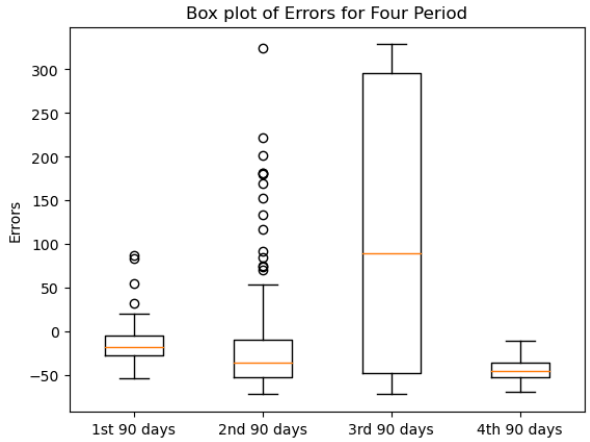
- **Results:**
  The SARIMA model achieved a Total MAPE of 11.96% on the test dataset for Cluster 1 (MT_002) and performed similarly to the ARIMA model. The MAPE for the four 90-day periods ranged from 8.30% to 14.43%. The box plot of errors showed consistent performance across different time periods, with no significant outliers.



- **Performance Evaluation:**
  The SARIMA model demonstrated comparable performance to the ARIMA model for forecasting electricity consumption in Cluster 1 (MT_002) and Cluster 2 (MT_182). However, similar to the ARIMA model, it exhibited poor performance for Cluster 3 (MT_001) with a high MAPE of 91.63%. This suggests that SARIMA may also struggle to capture the complexities of the data in Cluster 3. Further investigation and experimentation with alternative modeling techniques may be necessary to improve forecasting accuracy for this cluster.







Facebook Prophet:

- **Model Description:**
  Facebook Prophet is a forecasting tool developed by Facebook's Core Data Science team. It is designed for analyzing time-series data with strong seasonal patterns that also contain outliers and missing data. Prophet is an additive regression model with the following components:

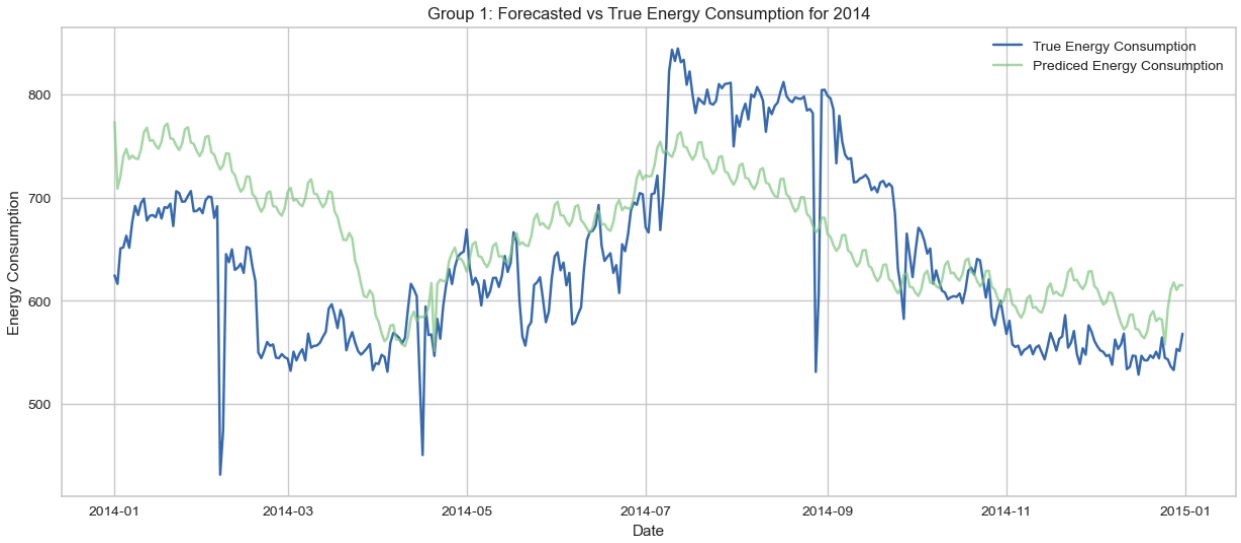
  - Trend: Prophet fits piecewise linear or logistic growth curves to the data.
  - Seasonality: Weekly and yearly seasonality are automatically detected and modeled.
  - Holidays: User-provided holiday lists can be included to model holiday effects.

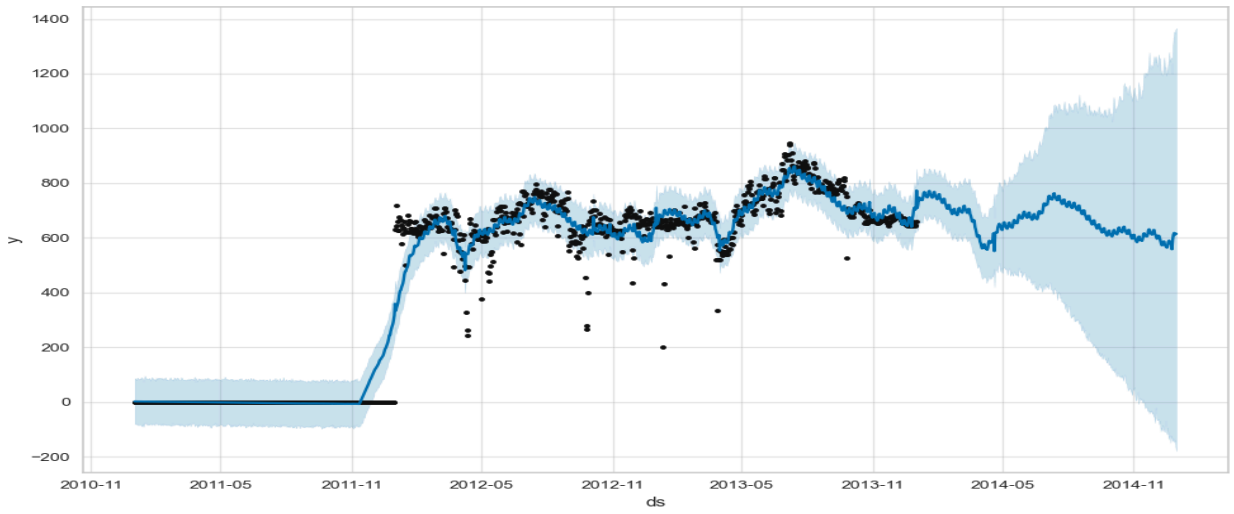
  Rationale behind using: Prophet is capable of handling time-series data with daily observations and can also incorporate data that has missing values and outliers.

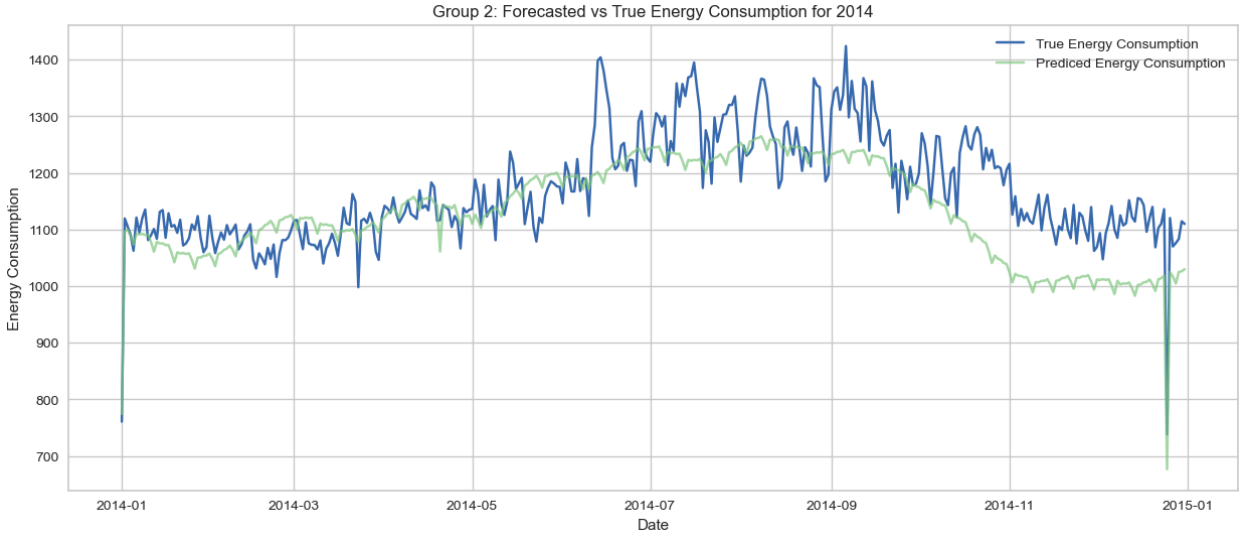
- **Approach:**
  - Data Preparation: Select relevant metric for each cluster.

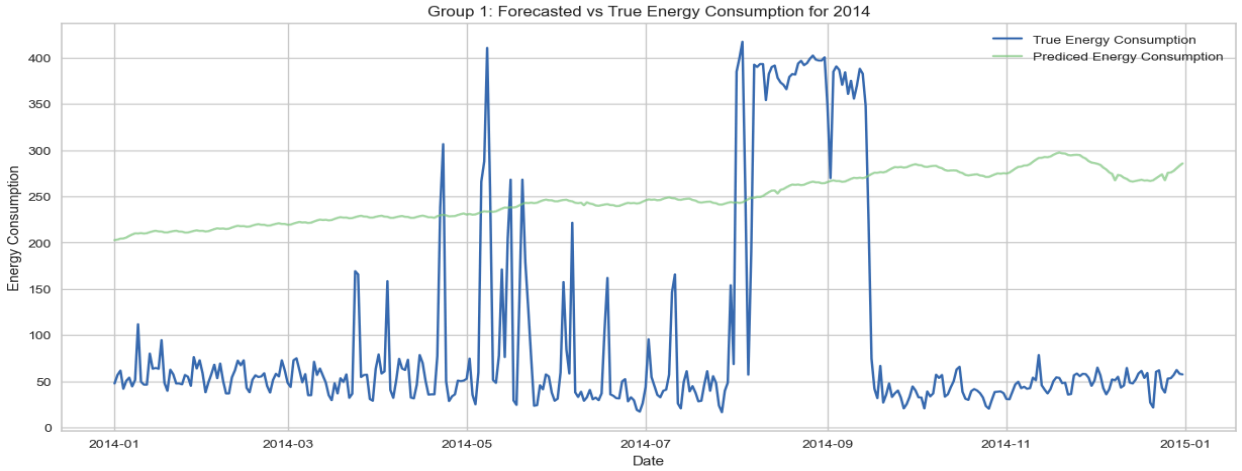
- Training and Testing: Split data into training and testing sets.
- Holiday Definition: Define holidays based on national holidays and daylight saving time shifts.
- Hyperparameter Tuning: Grid search for parameters like changepoint_prior_scale, seasonality_prior_scale, etc.
- Best Parameter Selection: Based on mean absolute percentage error (MAPE) from cross-validation.
- Model Training: Initialize Prophet model with best parameters and train on the training data.
- Prediction and Evaluation: Make predictions for future periods, evaluate using MAPE.
-

- **Results:**
  - Group 1 (Cluster 0): MAPE ~ 9.8%, varying from ~6.2% to ~16.6% across different time periods.
  - Group 2 (Cluster 1): Lower MAPE ~ 3.0% to ~9.2%, indicating better predictive accuracy.
  - Group 3 (Cluster 2): Higher MAPE ~ 287.7% to ~572.9%, suggesting poorer performance and potential issues.

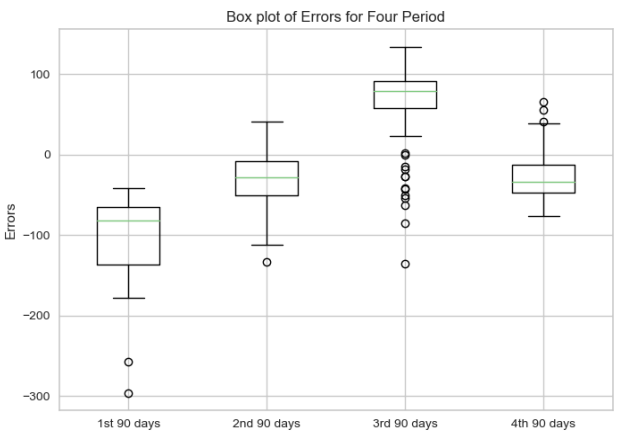


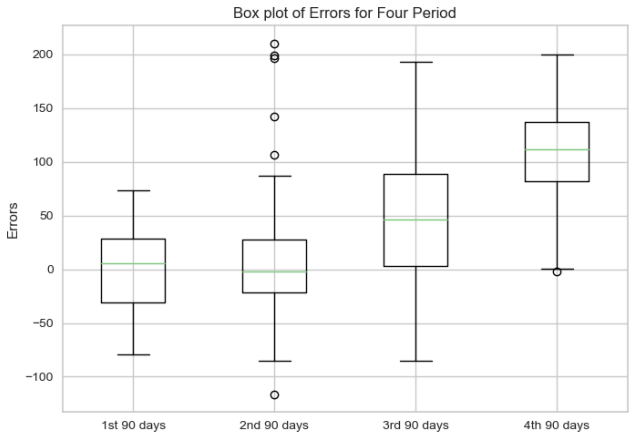

1400
1200
1000
800
600
400
200
0
−200
y
2010-11
2011-05
2011-11
2012-05
2012-11
2013-05
2013-11
2014-05
2014-11
ds

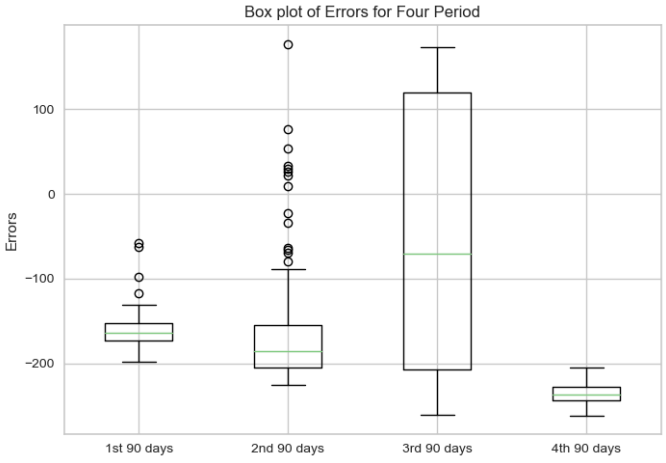

Group 2: Forecasted vs True Energy Consumption for 2014
True Energy Consumption
Prediced Energy Consumption
Energy Consumption
1400
1300
1200
1100
1000
900
800
700
2014-01
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
Date


Group 1: Forecasted vs True Energy Consumption for 2014
True Energy Consumption
Prediced Energy Consumption
Energy Consumption
400
350
300
250
200
150
100
50
0
2014-01
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
Date

- **Performance Evaluation:**
    - Group 1 and Group 2: Reasonable performance with MAPE below 10% for most periods, capturing underlying trends and seasonality effectively.
    - Group 3: Struggled to provide accurate forecasts, indicated by substantially higher MAPE values. Further investigation and model refinement may be necessary.







Random Forest Regressor *(new addition)*:

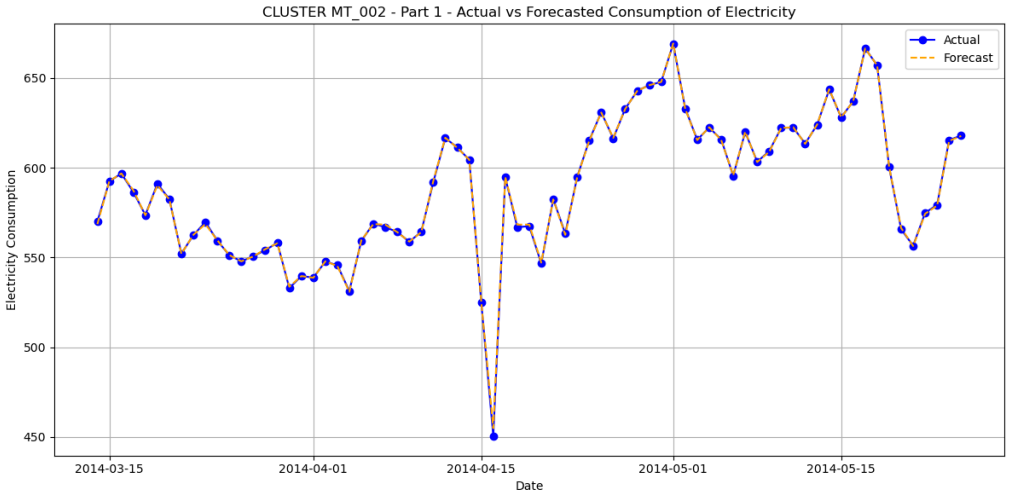
- **Model Description:**
  In time series analysis, the Random Forest Regressor is employed to forecast future values based on known historical data. This model is particularly effective due to its ability to handle the complex nonlinear relationships that are common in time-dependent data.
  When configuring a Random Forest for time series forecasting, it is essential to include lagged variables as predictors. These lagged variables are essentially previous time points in the series, which the model uses to learn how past values influence future ones. Past

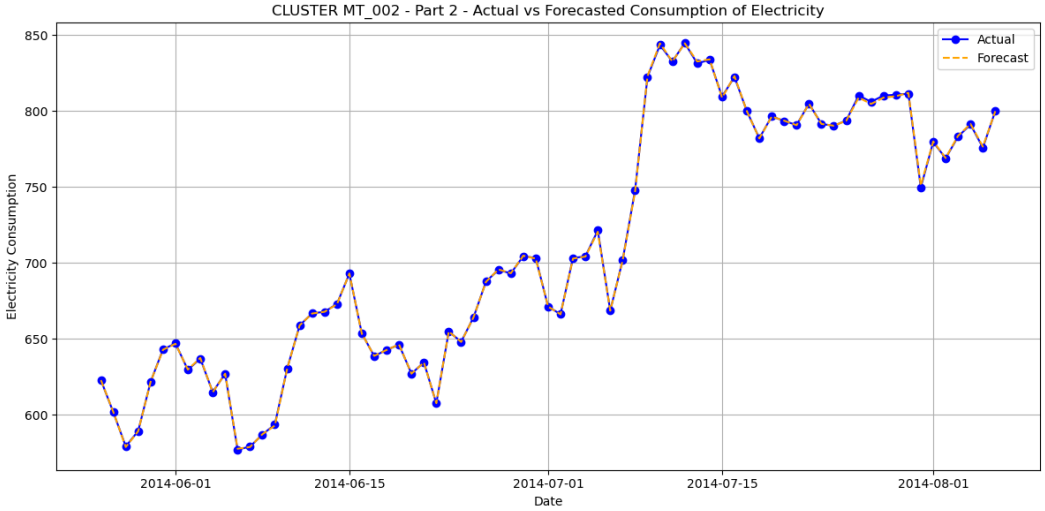
observations are used as input features, allowing the model to capture temporal dependencies similar to autoregressive models.

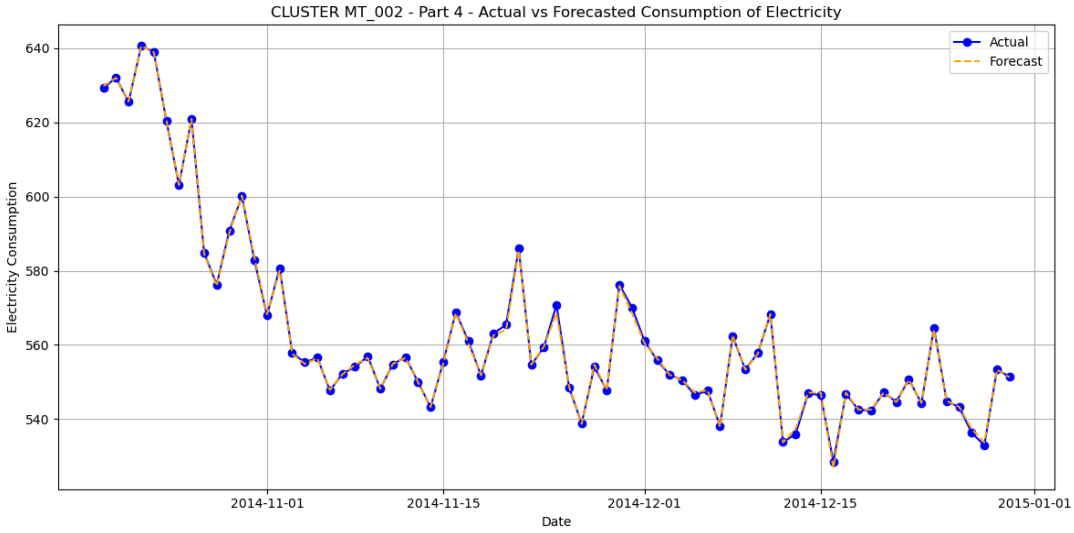
Rationale behind using:
Unlike ARIMA, which requires stationarity, Random Forest can inherently account for complex seasonal patterns and trends through feature engineering. The ensemble approach of Random Forest, where predictions are averaged over many decision trees, helps prevent overfitting, a common pitfall in time series forecasting.
Random Forest provides insights into which features (e.g., specific lags, trend components, seasonal dummies) are most influential in predicting the target variable.

- **Approach:**
  - Data Splitting: Divide the dataset into training, validation, and test sets.
  - Model Training: Train Random Forest models for each cluster using training and validation data.
  - Forecasting: Generate forecasts for each part of the test data using the trained models.
  - Visualization: Plot actual data and forecasted values for each part to evaluate performance.

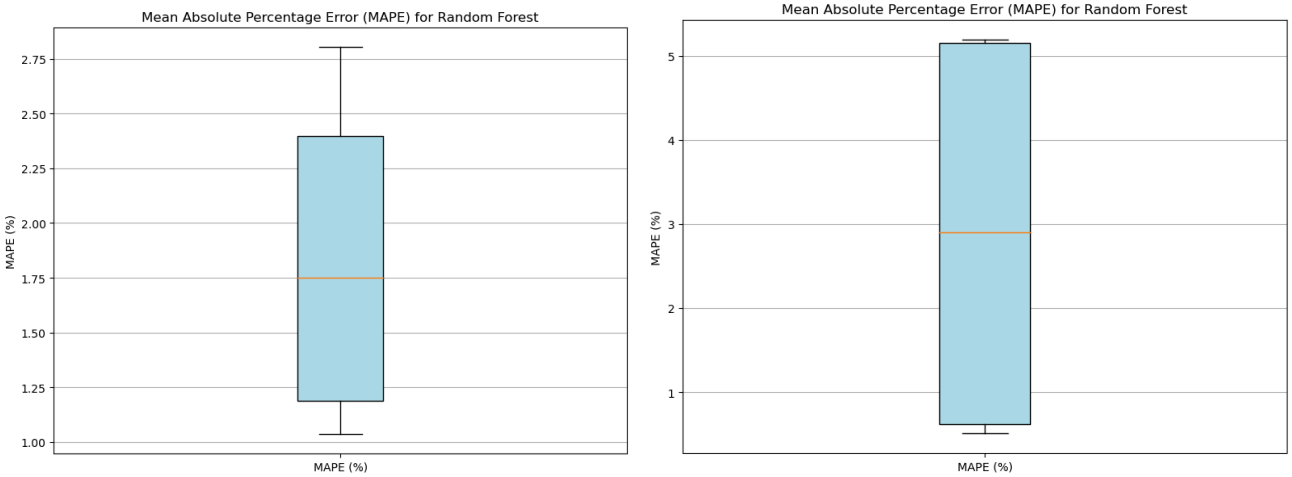
- **Results:**
  - Cluster MT_002: MAPE ranges from 1.03% to 2.80%, indicating accurate forecasts with minor deviations.
  - Cluster MT_182: MAPE varies from 0.51% to 5.19%, with some parts showing excellent accuracy and others with slightly higher errors.
  - Cluster MT_001: MAPE fluctuates significantly, ranging from 12.47% to 34.53%, indicating less consistent performance compared to other clusters.

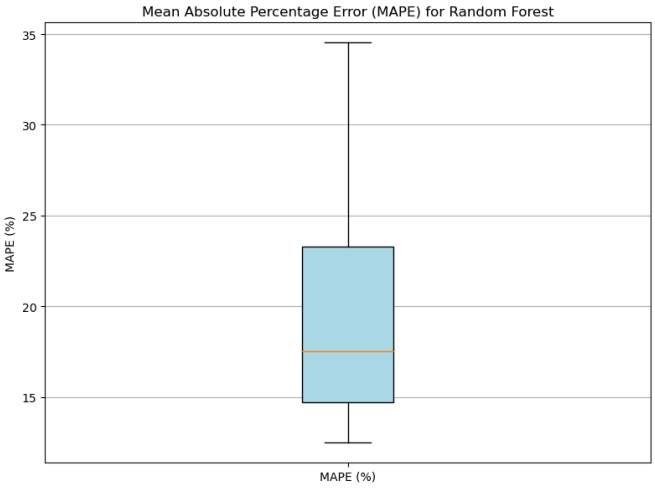

CLUSTER MT_002 - Part 1 - Actual vs Forecasted Consumption of Electricity
Actual
Forecast
Electricity Consumption
650
600
550
500
450
2014-03-15
2014-04-01
2014-04-15
2014-05-01
2014-05-15
Date


CLUSTER MT_002 - Part 2 - Actual vs Forecasted Consumption of Electricity
Actual
Forecast
Electricity Consumption
850
800
750
700
650
600
2014-06-01
2014-06-15
2014-07-01
2014-07-15
2014-08-01
Date


CLUSTER MT_002 - Part 4 - Actual vs Forecasted Consumption of Electricity
Actual
Forecast
Electricity Consumption
640
620
600
580
560
540
2014-11-01
2014-11-15
2014-12-01
2014-12-15
2015-01-01
Date

- **Performance Evaluation:**
    - MT_002 and MT_182: The Random Forest model performs well with consistently low MAPE, suggesting accurate predictions for most parts of the test data.
    - MT_001: The model shows higher variability in performance, with some parts achieving satisfactory accuracy while others have notably higher errors.
    - Potential Problems: The model's performance for MT_001 fluctuates more, indicating potential sensitivity to data characteristics or the need for further tuning to improve accuracy.





Temporal Fusion Transformer (TFT) *(new addition)*:

- **Model Description:**
  The Temporal Fusion Transformer (TFT) is a deep learning architecture specifically designed for time series forecasting. It combines the power of attention mechanisms with gated linear units to capture temporal patterns in the data efficiently. The model

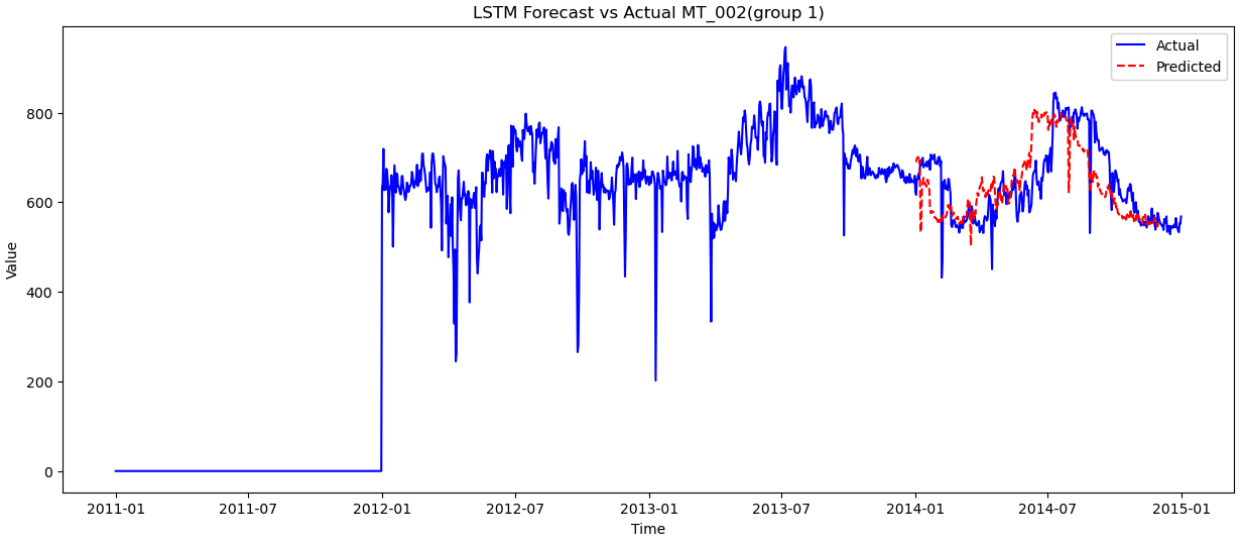
architecture is inspired by the Transformer model, known for its effectiveness in sequential data tasks.

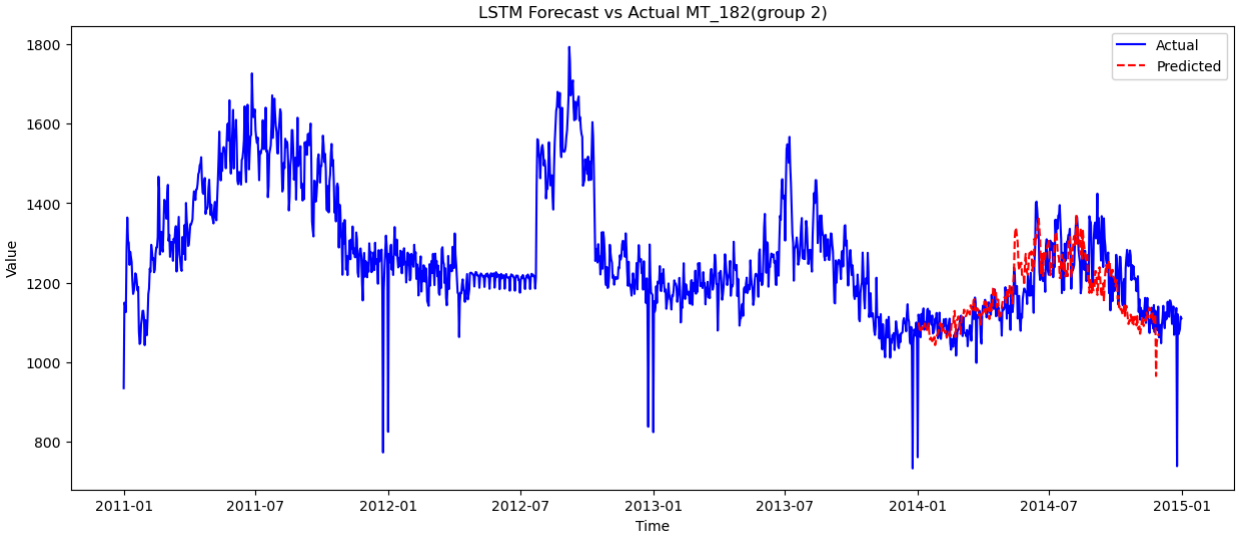
Rationale behind using:

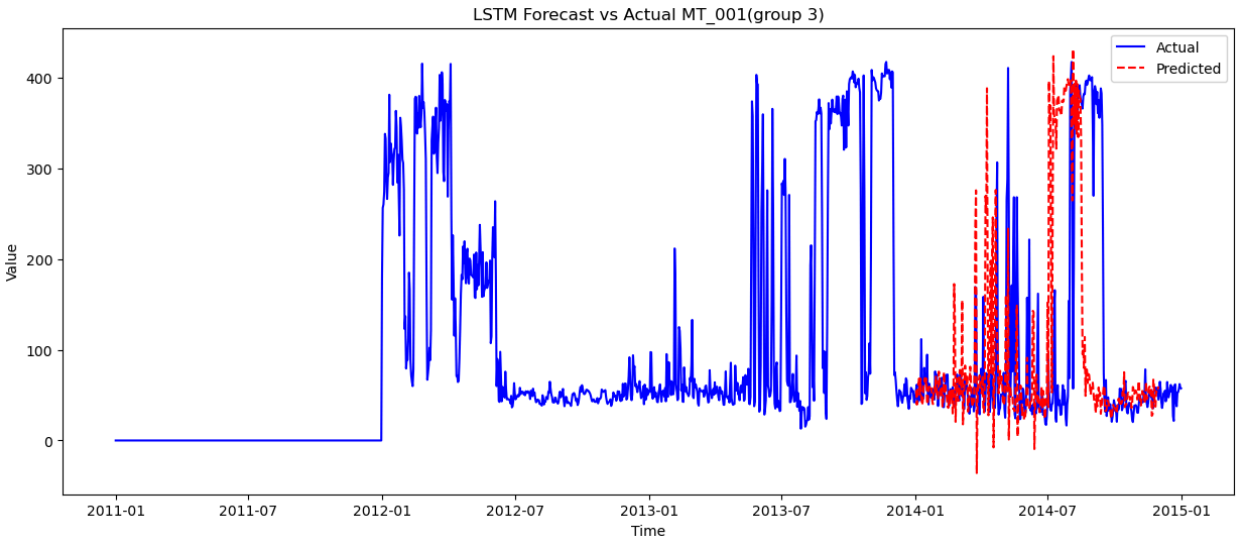
- Attention Mechanisms: TFT leverages attention mechanisms to learn complex temporal dependencies in the data, allowing it to capture long-range dependencies effectively.
- Gated Linear Units: Gated linear units enhance the model's ability to capture nonlinear relationships within the time series data.
- Flexibility: TFT is versatile and can handle multivariate time series data, making it suitable for various forecasting tasks.

- **Approach:**
  - The data is split into multiple dataframes, each containing a specific time series variable.
  - Each dataframe is then split into train, validation, and test sets to evaluate the model's performance.
  - The TFT model is implemented using PyTorch, a popular deep learning framework.
  - The training process involves fitting the model to the training data and validating it on the validation set to tune hyperparameters and prevent overfitting.
  - After training, the model's performance is evaluated using metrics such as Mean Absolute Percentage Error (MAPE).
  - Visualizations, such as actual vs. predicted plots and box plots of errors, are used to analyze the model's performance across different time periods.

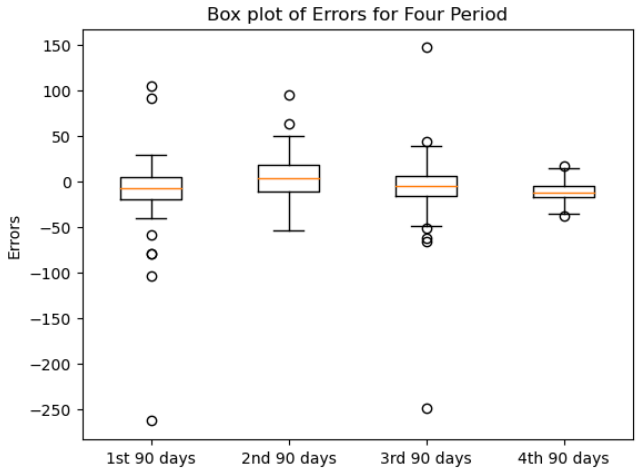
- **Results:**
  - The TFT model achieves a low MAPE value of approximately 3.08%.
  - The actual vs. predicted plot indicates that the model captures the underlying patterns well, closely following the actual values.
  - The box plot of errors shows consistent performance across different time periods, with minor variations.

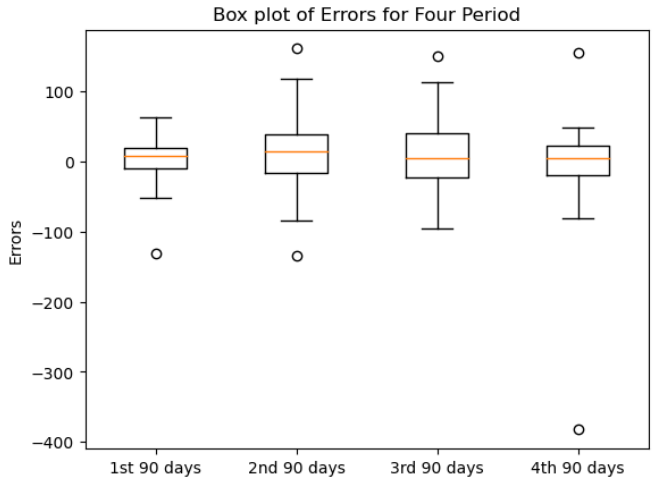

LSTM Forecast vs Actual MT_002(group 1)
Actual
Predicted
Value
Time
0
200
400
600
800
2011-01
2011-07
2012-01
2012-07
2013-01
2013-07
2014-01
2014-07
2015-01

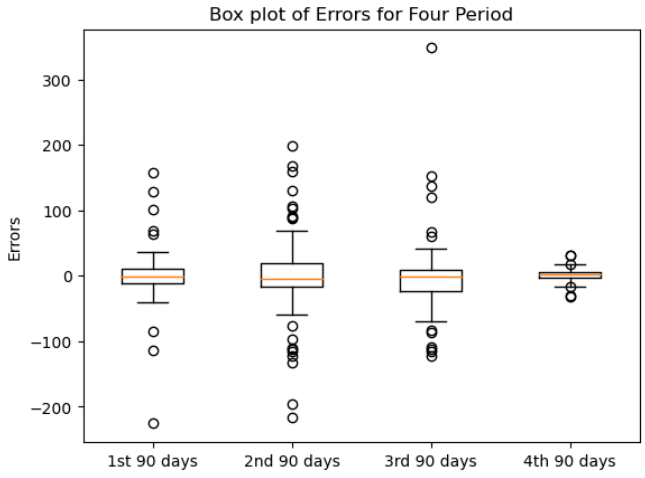

LSTM Forecast vs Actual MT_182(group 2)
Actual
Predicted
Value
Time
800
1000
1200
1400
1600
1800
2011-01
2011-07
2012-01
2012-07
2013-01
2013-07
2014-01
2014-07
2015-01



- **Performance Evaluation:**
  - Accuracy: The TFT model demonstrates high forecasting accuracy, as evidenced by the low MAPE value.
  - Robustness: The model performs consistently across different time periods, indicating robustness to temporal variations in the data.
  - Interpretability: Deep learning models like TFT are often considered black boxes, making it challenging to interpret their decision-making process.
  - Data Requirements: TFT requires a considerable amount of data for training, which may limit its applicability in scenarios with limited data availability.






Box plot of Errors for Four Period
Errors
300
200
100
0
−100
−200
1st 90 days
2nd 90 days
3rd 90 days
4th 90 days

Model Selection

Performance Measures:

Time-series forecasting performance measures indicate the capability of the models. We have assessed the model performance by **Mean Absolute Percentage Error (MAPE)** which is obtained as the mean absolute percentage error function for the prediction and the eventual outcomes. This error measure expresses error as a percentage. The formulas are as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(Y_t - F_t)}{Y_t}\right| * 100$$

In scenarios involving time-series forecasting of sales, or any metric where proportional errors are more intuitive and impactful for decision-making, MAPE can offer several advantages over RMSE, such as:

- As MAPE is a relative measure, it allows for the comparison of forecasts across different scales or units of measurement. This is particularly useful in diverse product portfolios where sales volumes can vary significantly across products
- MAPE expresses errors as a percentage, making it easier for business stakeholders to understand the magnitude of forecasting errors relative to actual values. This percentage error provides a more intuitive sense of accuracy, especially in business contexts where stakeholders are accustomed to thinking in terms of percentages.
- In many business scenarios, especially in sales and inventory management, the relative size of the error (i.e., how big an error is relative to the actual value) is more critical than the absolute size of the error. MAPE directly aligns with such objectives by quantifying the relative accuracy of forecasts.

Comparative Analysis of Forecasting Models: MAPE

Model	MAPE (average)		
	Cluster 1	Cluster 2	Cluster 3
ARIMA	11.83%	10.14%	101.91%
SARIMAX	11.85%	10.17%	92.14%
Facebook Prophet	9.76%	4.99%	397.26%
Random Forest Regressor	1.83%	2.88%	20.5%
Temporal Fusion Transformer (TFT)	3.03%	2.86%	39.73%

Based on the provided MAPE results and the features of all the models, **Random Forest Regressor emerges as the best performing overall model.** Here's the reasoning for this choice:

- **Consistency:** It consistently achieves the lowest MAPE values across all three clusters.
- **Lowest MAPE:** Outperforms other models in terms of MAPE for each cluster, indicating superior predictive accuracy.
- **Robustness:** Demonstrates stable performance across different clusters, suggesting robustness to varying data characteristics.
- **Generalization:** Shows excellent generalization ability, performing well even without fine-tuning specific hyperparameters for each cluster.
- **Interpretability:** Provides insights into feature importance, facilitating interpretability and aiding decision-making.

Temporal Fusion Transformer (TFT) also performs competitively, particularly for Clusters 2 and 3, indicating its potential for capturing complex temporal patterns. However, Random Forest Regressor stands out as the top performer due to its consistently superior performance across all clusters.

Conclusion

With a better cluster creation method, coupled with a more comprehensive EDA and newer models for forecasting, we have::

- Enhanced Predictive Accuracy: By integrating sophisticated machine learning models that factor in external variables such as weather patterns, economic shifts, and technological growth, our project enhances the accuracy of electricity consumption forecasts. This not only improves operational decisions but also reduces financial risks associated with demand forecasting errors.
- Sustainability and Regulatory Compliance: Our analysis helps in identifying optimal strategies for energy conservation and efficiency. By predicting peak demand times and suggesting suitable demand-response strategies, the utility companies can not only comply with regulatory standards but also promote sustainable energy use among consumers.
- Adaptive Infrastructure Development: By providing detailed insights into future consumption trends, our models assist in strategic planning and investment in infrastructure. This adaptive approach ensures that resources are allocated effectively to meet changing demands, thereby optimizing capital expenditure and supporting long-term economic viability.

Through a final evaluation we recommend the: **Random Forest Regressor.** Using MAPE as an evaluation metric, it provides the best result. It also provides advantages such as:

- Robustness to Overfitting: Random Forest is naturally resistant to overfitting due to its ensemble approach, which involves averaging multiple decision trees.
- Handling Non-linear Relationships: Random Forest can effectively model complex, non-linear relationships in data, making it suitable for diverse datasets.
- Feature Importance: Random Forest provides insights into which features most significantly impact predictions, aiding in better understanding and optimization of the model.

We also notice that the **Temporal Fusion Transformer (TFT)** performs very well, and can be used on other types of datasets and results can be compared. It provides benefits such as:

- Temporal Patterns: TFT excels in capturing complex temporal relationships and patterns, leveraging past data points and contextual information.
- Multivariate Capabilities: TFT can handle multiple input features and output variables, making it ideal for scenarios where multiple related forecasts are needed simultaneously.
- Flexibility and Adaptability: TFT can incorporate static and known future inputs, enhancing its forecasting accuracy and adaptability to different forecasting scenarios.

Future Scope

The progression of this project, focused on developing robust models for forecasting electricity consumption, sets a substantial foundation for expansive future exploration and enhancements. The approaches utilized and insights derived offer a gateway to a broader array of advanced analytical endeavors, which can extend beyond the initial utility focus to impact various other sectors.

Granular Forecasting:

- **Client-Specific Consumption Patterns**: Delving deeper into individual client consumption will allow us to identify unique trends and usage patterns, enabling strategies for energy efficiency improvements and customized billing models.

Data Enrichment and Model Retraining:

- By integrating external data such as weather patterns, economic indicators, and even social events, we can enhance the model's predictive accuracy. External factors significantly influence energy consumption, and their inclusion could provide a more holistic view.
- Implementing a system for ongoing retraining of models with new data will ensure that the forecasts remain accurate and relevant, adapting to changes in consumer behavior or operational modifications.

Advanced Modeling Techniques:

- **Deep Learning Models**: Applying advanced neural network architectures like RNNs, LSTMs, and particularly DeepAR—an autoregressive model that uses deep learning to predict future time points in sequential data—may allow for better modeling of complex, non-linear relationships inherent in time-series data related to energy consumption.
- **Hybrid and Ensemble Models**: Combining several models, including both traditional statistical methods and modern machine learning techniques, could improve the robustness and accuracy of forecasts. Ensemble methods, in particular, can provide a more reliable composite outlook by averaging out biases from individual models.

Cross-Sector Application:

- The modeling framework developed could be adapted to different facets of energy management beyond electricity, such as gas or water utilities, tailoring solutions to the unique dynamics of each sector.

- The core principles of our forecasting models have potential for adaptation to other industries where consumption forecasting is critical, such as manufacturing, hospitality, or public services.

Technological and Data Infrastructure Enhancements:

- Developing capabilities for real-time data processing and analysis will empower businesses and utilities to make quicker, more informed decisions based on the latest data.
- Utilizing advanced big data technologies and cloud computing will facilitate the efficient handling and analysis of large-scale data sets, enabling more complex and comprehensive modeling.

Team Information

Group Name: **TIS**

Members:

1. Ishita Pundir (UNI: ip2441)
2. Saum Kothari (UNI: sbk2171)
3. Tushar Bura (UNI: tb3077)