**Technical Report for Almeda Site**

**Overview**

The analysis presented in this report revolves around the operations of the Alameda Site, with a primary focus on predicting the treated water flow and energy consumption at this location. The dataset at our disposal captures a myriad of metrics, including rainfall, the total volume of treated water leaving the plant, operational hours of the plant, raw water input, among other significant attributes that play a pivotal role in determining the patterns of treated water flow and energy consumption at the Alameda Site. The objective of this report is to connect the power of this historical data with advanced machine learning methodologies and time series forecasting techniques, to explore predictive models capable of accurately forecasting future values of Treated water flow and Energy Consumption. Therefore, this report aims to deliver valuable insights that can strengthen the efficiency and sustainability of the Alameda Site's operations. The predictions made for Treated Water Flow and Energy Consumption will serve as a roadmap for informed decision-making, ensuring optimal resource allocation and operational excellence.

**Data Preprocessing and Feature Engineering**

Data preparation is a foundational step in any data analysis project. Ensuring that the dataset is structured correctly, devoid of errors, and supplemented with relevant features is crucial for the success of subsequent modeling efforts. This section elaborates on the various processes undertaken to get the data into an optimal format for analysis.

**Data Importation**

To create a comprehensive dataset, we merged these files based on the 'Date' column. This merger ensured that we had a unified dataset, capturing all the necessary attributes in a single table, which is instrumental for a holistic analysis. After merging, it's essential to understand the nature of each column in our dataset. This understanding aids in further preprocessing tasks and ensures that we utilize the data effectively in the modeling phase. In the subsequent section, we'll provide an overview of these columns, detailing their data types and significance.

For the Alameda Site analysis, data was sourced from three distinct datasets. Each dataset was imported and examined to ensure it was in the appropriate format for subsequent analysis.

* **Predictors Dataset:** This dataset was extracted from a CSV file titled almeda\_data.csv. It encompasses various metrics that could influence the treated water flow and energy consumption at the Alameda Site. The dataset includes attributes such as temperature, clear sky indices, cloud type, dew point, humidity, solar zenith angle, wind direction, and speed, among others.
* **Energy Consumption Dataset**: This dataset was derived from an Excel file named Almeda Electrical Info 1.xlsx. It provides detailed insights into the energy consumption patterns at the Alameda Site. The dataset captures metrics like the end of the billing period, electrical consumption, peak usage, charges, and various demand metrics.
* **Water Flow Dataset:** The water flow data was imported from an Excel file named Almeda Flows.xlsx. This dataset is more concise, focusing on the date and the average flow of treated water.

For a more structured overview, the table below presents the columns and their respective data types for each dataset. This table provides a comprehensive overview of the columns and their respective data types for each dataset.

|  |  |  |
| --- | --- | --- |
| Dataset Name | Column Name | Data Type |
| Predictors | Year | int64 |
| Month | int64 |
| Day | int64 |
| Hour | int64 |
| Minute | int64 |
| Temperature | float64 |
| Clearsky DHI | float64 |
| Clearsky DNI | float64 |
| Clearsky GHI | float64 |
| Cloud Type | int64 |
| Dew Point | float64 |
| DHI | int64 |
| DNI | int64 |
| Fill Flag | int64 |
| GHI | int64 |
| Relative Humidity | float64 |
| Solar Zenith Angle | float64 |
| Surface Albedo | float64 |
| Pressure | int64 |
| Precipitable Water | float64 |
| Wind Direction | int64 |
| Wind Speed | float64 |
| Global Horizontal UV Irradiance (280-400nm) | float64 |
| Global Horizontal UV Irradiance (295-385nm) | float64 |
| Energy Consumption | end of billing period | datetime64[ns] |
| Electrical | float64 |
| On-Peak Usage (kWh) | float64 |
| Part-Peak Usage (kWh) | float64 |
| Off-Peak Usage (kWh) | float64 |
| Total Usage (kWh) | float64 |
| Charges (dollars) | float64 |
| Billing Demand (kW) | float64 |
| Created Demand (kW) | float64 |
| On-Peak Demand (kW) | float64 |
| Part-Peak Demand (kW) | float64 |
| Off-Peak Demand (kW) | float64 |
| Water Flow | Date | datetime64[ns] |
| BPW Flow Average (MGD) | float64 |

**Data Cleaning and Feature Engineering**

Ensuring the data is accurate, consistent, and usable is a foundational step in any data analysis process. The data cleaning process for the Alameda Site involved several steps to refine and structure the data for subsequent analysis

* **Date Conversion:** For both others and other datasets, the 'Year', 'Month', and 'Day' columns were combined to form a new 'Date' column. This new column was converted into a datetime format, which is more suitable for time series analysis.
* **One-Hot Encoding:** The 'Month' column in the others dataset underwent one-hot encoding. This transformation is essential for converting categorical month values into a machine-readable format, enhancing the model's performance.
* **Data Type Conversion:** The data type of the 'Day' column in the others dataset was checked. If identified as an object type, it was converted to an integer type, ensuring data consistency.
* **Duplicate Handling:** Both the flow\_data and others datasets were scrutinized for duplicate 'Date' entries. The number of duplicate dates in each dataset was reported and subsequently removed, retaining only the first occurrence of each date.
* **Data Merging:** The flow\_data and others datasets were merged based on the 'Date' column, consolidating all relevant information into a single dataframe. The merged dataframe, flow\_data, contained [Number of rows in the merged dataframe] rows.
* **Missing Value Detection and Handling:** The merged flow\_data dataframe was examined for missing values across all columns. A detailed report was generated, highlighting the number of missing values in each column.
* Similarly, the energy\_data dataframe was inspected for missing values. Any rows with missing values were removed, ensuring data integrity. Post removal, the absence of missing values was re-verified for both dataframes.
* From the energy\_data dataframe, only the 'Total Usage (kWh)' and 'end of billing period' columns were extracted. The 'end of billing period' column was renamed to 'Date' for consistency.
* This refined energy\_data dataframe was then merged with the other dataframe based on the 'Date' column.
* The final lengths of the flow\_data and energy\_data dataframes were reported as 3111 and 1248 rows, respectively.

At the end of the data cleaning process, both the flow\_data and energy\_data dataframes were structured, clean, and primed for the next stages of exploratory data analysis and modeling. This thorough cleaning ensures that the subsequent analyses and models are built upon a reliable data foundation.

**Lagged Features**

In time series forecasting, one common technique to improve the predictive power of models is the creation of lagged features. These are essentially values from previous time steps that can provide context for a given prediction. For the Alameda Site data, lagged features were generated for both water flows and energy consumption:

* **Lagged Features for Water Flows:** Three lagged features were created for the 'BPW Flow Average (MGD)' column in the flow\_data dataframe. These represent the water flows from the previous three days. The new columns were named 'Water\_Flows\_Lag\_1', 'Water\_Flows\_Lag\_2', and 'Water\_Flows\_Lag\_3', where each number indicates the number of days before the current date. Due to the creation of these lagged features, the initial few rows, which didn't have sufficient historical data for lagging, contained NaN values. These rows were removed to maintain data consistency.
* **Lagged Features for Energy Consumption:** Similarly, three lagged features were created for the 'Total Usage (kWh)' column in the energy\_data dataframe, representing the energy consumption from the previous three days. The new columns were named 'Energy\_Lag\_1', 'Energy\_Lag\_2', and 'Energy\_Lag\_3'. As with the water flows data, rows with NaN values resulting from the lagging process were removed from the energy\_data dataframe.

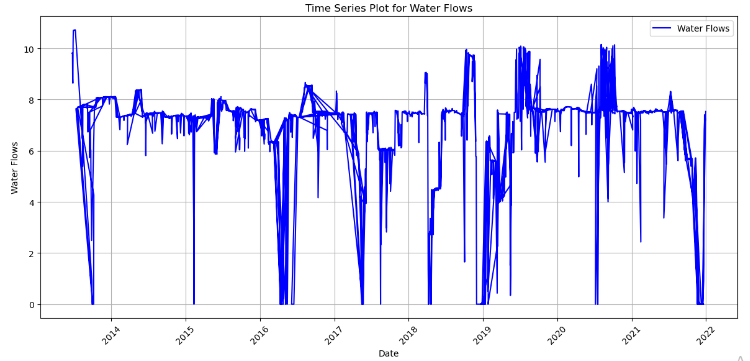
The introduction of these lagged features can enhance the models' ability to recognize patterns and dependencies in the data, potentially leading to more accurate predictions. By considering the values from previous days, the models can better understand the temporal dynamics and relationships inherent in the data.

**Exploratory Data Analysis**

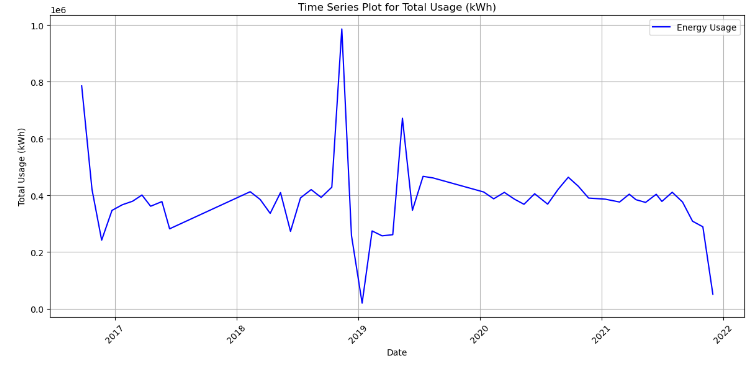
Exploratory Data Analysis (EDA) is crucial in understanding the underlying patterns, relationships, and structures within the dataset. This section examines the visual and statistical exploration of the dataset, focusing primarily on the target variables ' BPW Flow Average (MGD)' and ‘Total Usage, Kwh' columns.

**Time Series Analysis**

A time series plot was generated to visualize the flow of treated water over time. The x-axis represents the 'Date', while the y-axis denotes the BPW Flow Average (MGD)'. The plot provides a visual representation of how the treated water flow has varied over the dataset's timeframe.



Similarly, a time series plot was created to showcase the daily energy consumption. The x-axis represents the 'Date', and the y-axis indicates the ‘Total Usage, Kwh'. This visualization helps in understanding the energy consumption patterns and any anomalies or spikes in consumption.

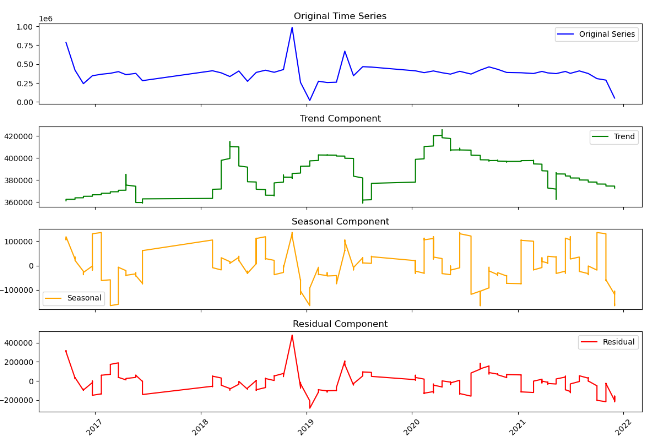


**Seasonal Decomposition**

To understand the underlying seasonal patterns, trends, and residuals in the treated water flow, a seasonal decomposition was performed. This decomposition breaks down the time series into its constituent components, providing insights into the periodic fluctuations, long-term trends, and any irregular patterns not accounted for by the seasonality or trend.



Similarly, a seasonal decomposition was conducted for the energy consumption data. This decomposition helps in discerning the cyclical patterns, overarching trends, and any anomalies in the energy consumption data.



**Stationarity Test**

In time series analysis, ensuring that a dataset is stationary is crucial before applying forecasting models. A stationary time series is one whose properties do not depend on the time at which the series is observed. This means that the series has constant mean and variance over time, and the covariance between two time periods is only a function of the difference between the two periods and not the actual time at which the series is observed. To test the stationarity of the datasets, the Augmented Dickey-Fuller (ADF) test was employed. The ADF test is a type of unit root test that aims to determine how strongly a time series is defined by a trend. The null hypothesis of the ADF test is that the time series is non-stationary. If the p-value obtained from the test is less than the significance level (e.g., 0.05), then the null hypothesis can be rejected and it can be concluded that the time series is stationary.

The ADF test was applied to both the 'BPW Flow Average (MGD)' column from the flow\_data dataframe (representing treated water) and the 'Total Usage (kWh)' column from the energy\_data dataframe (representing energy consumption) and the results are as summarized in the table below

|  |  |  |
| --- | --- | --- |
| Metric/Variable | Treated water | Energy Consumption |
| ADF Statistic | -6.0261 | -4.4513 |
| p-value | 1.45e-07 | 0.0002406 |
| Critical Value (1%) | -3.4325 | -3.4356 |
| Critical Value (5%) | -2.8625 | -2.8639 |
| Critical Value (10%) | -2.5673 | -2.5680 |
| Stationarity Conclusion | Stationary | Stationary |

As observed from the results summarized above, both treated water and energy consumption datasets, the p-values were significantly less than 0.05. This provides strong evidence against the null hypothesis, leading to the conclusion that both datasets are stationary. This is a positive outcome, as stationary data is a prerequisite for many time series forecasting techniques.

Summarily, the exploratory data analysis provided a comprehensive understanding of the dataset's characteristics and underlying patterns. Through visualizations, we observed the temporal trends in treated water flow and energy consumption, highlighting their seasonality and potential cyclic behaviors. The seasonal decomposition further emphasized these periodic components, separating the trend, seasonality, and residuals for both variables. Moreover, the Augmented Dickey-Fuller test confirmed the stationarity of the time series data, ensuring its readiness for subsequent modeling and forecasting tasks. Overall, the EDA has laid a solid foundation for the next steps in the data analysis pipeline, ensuring that the modeling phase is informed by a deep understanding of the data's structure and properties.

**Initial Training**

In our quest to understand the predictive power of our selected models, we embarked on an initial training phase. This phase was designed to explore the strength of the models when trained and tested on a 50:50 data partition. This means that half of the available data was used for training, while the other half was reserved for testing.

**Data Partitioning**

* **For the Treated Water dataset**
* **Training set size:** 50% of the total data
* **Test set size:** 50% of the total data
* **For the Energy Consumption dataset**
* Training set size: 50% of the total data
* Test set size: 50% of the total data

**Model Training**

For both datasets, three models were trained

* **Random Forest (RF):** A versatile ensemble learning method that can be used for both regression and classification tasks.
* **Gradient Boosting (GB):** An iterative technique that adjusts for the errors of the previous model.
* **SARIMA:** A traditional time series forecasting method that combines autoregressive (AR) and moving average (MA) models with the concept of differencing.

**Model Evaluation**

The models were evaluated using three metrics

* **Root Mean Squared Error (RMSE):** Measures the average magnitude of the errors between predicted and observed values.
* **Mean Absolute Error (MAE):** Represents the average of the absolute differences between predictions and actual values.
* **Error Rate:** Calculates the proportion of predictions that deviate from the actual values by more than a given threshold.

The results of the evaluations are summarized in the tables below

Treated Water Evaluation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Treated Water | | | Energy Consumption | | |
| Model | **RMSE** | **MAE** | **Error** | **RMSE** | **MAE** | **Error** |
| Random Forest | 1.4493 | 0.8297 | 0.4054 | 158007.67 | 87261.68 | 1.0000 |
| Gradient Boosting | 1.7119 | 1.1487 | 0.6216 | 167453.20 | 97316.21 | 1.0000 |
| SARIMA | 2.8473 | 2.5797 | 0.9345 | 744929.67 | 696573.44 | 0.4438 |
| Ensemble | 1.7864 | 1.4083 | 0.8636 | 276639.41 | 257280.24 | 1.0000 |

The report also includes visualizations that juxtapose the actual data against the predictions. These visualizations provide a clear, intuitive means for stakeholders to grasp the findings and implications of our analyses. The plots showcase the training data, validation data, test data, and predictions, offering a comprehensive view of the model's performance over time. The 50:50 data partitioning offered a balanced approach, ensuring that the models were exposed to a representative sample of the data during training. The evaluation metrics further illuminated the strengths and weaknesses of each model, guiding our subsequent modeling decisions.

**Model Training**

In the model training phase, the primary goal is to use historical data to train algorithms to recognize patterns and relationships. This enables the model to make accurate predictions on unseen data. For the Alameda Site datasets, we employed a combination of time series and machine learning models to forecast treated water flow and energy consumption. The following details the steps taken during the model training process

**Data Splitting**

* **Treated Water Data:** The dataset was sorted chronologically and split into training (80%), and testing (20%) sets.
* **Energy Consumption Data:** Similarly, the energy data was sorted by date and divided into training (80%) and testing (20%) sets.
* Additionally, a more granular split was performed, dividing the data into training (60%), validation (20%), and testing (20%) sets for both treated water and energy consumption.

**SARIMA Model Training**

* **Treated Water:** A Seasonal Autoregressive Integrated Moving Average (SARIMA) model was trained on the treated water data. The model parameters were set to order=(1, 1, 1) and seasonal\_order=(1, 1, 1, 12).
* **Energy Consumption:** A similar SARIMA model was trained for the energy consumption data with the same parameters.

**Machine Learning Model Training**

* **Treated Water**
* **Random Forest:** A Random Forest Regressor was trained using the treated water data.
* **Gradient Boosting:** A Gradient Boosting Regressor was also trained on the same dataset.
* SARIMA: In addition to the machine learning models, a SARIMA model was trained for the treated water data.
* **Energy Consumption**
* **Random Forest:** A Random Forest Regressor was trained using the energy consumption data.
* **Gradient Boosting:** Similarly, a Gradient Boosting Regressor was trained for the energy consumption data.
* **SARIMA:** A SARIMA model was also trained for the energy consumption data, using the same parameters as before.

**Error Handling**

During the SARIMA model training for both treated water and energy consumption, error handling mechanisms were in place. If any issues arose during the training process, the error messages were captured and displayed, ensuring that the training process could continue without interruption.

The model training phase involved a combination of time series and machine learning models. The SARIMA model, known for its efficacy in handling time series data, was chosen for its ability to account for seasonality, trend, and noise in the data. On the other hand, machine learning models like Random Forest and Gradient Boosting were employed for their robustness and ability to capture complex non-linear relationships in the data. By leveraging these models, we aim to achieve accurate and reliable forecasts for both treated water flow and energy consumption at the Alameda Site. The subsequent sections will delve into model validation and performance evaluation, where the true efficacy of these trained models will be assessed against unseen data.

**Model Predictions**

After successfully training our models, we proceeded to evaluate their performance by making predictions on the test data. This section elaborates on the prediction process for both treated water flow and energy consumption using the Random Forest, Gradient Boosting, and SARIMA models.

**Treated Water Predictions**

* **Random Forest:** Using the trained Random Forest model, we predicted the treated water flow on the test data, resulting in the predictions labeled as rf\_pred\_flow.
* **Gradient Boosting:** The Gradient Boosting model was similarly applied to the test data, producing the predictions gb\_pred\_flow.
* **SARIMA:** The SARIMA model, tailored for time series forecasting, was used to predict the treated water flow for the test period. The prediction interval spanned from the end of the training data to the end of the test data, yielding the predictions sarima\_pred\_flow.
* **Ensemble Predictions**: To harness the collective power of each model and potentially enhance prediction accuracy, we adopted an ensemble method. The ensemble predictions for treated water, ensemble\_pred\_flow, were derived by averaging the predictions from the three individual models.

**Energy Consumption Predictions**

* **Random Forest:** For predicting energy consumption, the Random Forest model was applied to the test data, with the results stored in rf\_pred\_energy.
* **Gradient Boosting:** Using the Gradient Boosting model on the test data, we obtained the predictions gb\_pred\_energy.
* **SARIMA:** The SARIMA model provided predictions for energy consumption over the specified test interval, resulting in sarima\_pred\_energy.
* **Ensemble Predictions:** As with treated water, an ensemble approach was used for energy consumption predictions. The ensemble predictions, ensemble\_pred\_energy, were computed by averaging the outputs from the three models.

Predicting on the test data is a pivotal step in assessing the efficacy of our trained models. By employing an ensemble approach, we aim to benefit from the unique strengths of each model, potentially improving the overall prediction accuracy. Subsequent sections will delve into the accuracy assessment of these predictions against the actual test values, offering insights into the models' effectiveness and the advantages of the ensemble strategy.

**Model Evaluation**

After making predictions on the test data, it's crucial to evaluate the performance of our models. This helps in understanding the accuracy of our predictions and provides insights into areas of improvement. In this section, we evaluate the performance of our models using three metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Error Rate.

* **Root Mean Squared Error (RMSE):** Measures the average magnitude of the errors between predicted and observed values. A lower RMSE indicates a better fit to the data.
* **Mean Absolute Error (MAE):** Represents the average of the absolute differences between the predicted and actual values. It provides a linear penalty for each unit of difference.
* **Error Rate:** Calculates the proportion of predictions that deviate from the true values by more than a specified threshold. A lower error rate indicates more accurate predictions.

These evaluations provide valuable insights into the strengths and weaknesses of each model, guiding future model refinement and selection processes. The table below summarizes the evaluation results for both treated water flow and energy consumption predictions

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Metrics | Treated Water Validation | Energy Consumption Validation |
| RF | RMSE | 1.5103 | 75771.8950 |
|  | MAE | 0.8519 | 41573.4852 |
|  | Error Rate | 38.84% | 100.00% |
| GB | RMSE | 1.4194 | 82364.0420 |
|  | MAE | 0.7239 | 37346.0781 |
|  | Error Rate | 28.89% | 100.00% |
| SARIMA | RMSE | 1.9577 | 164775.4875 |
|  | MAE | 1.0023 | 116407.8183 |
|  | Error Rate | 0.32% | 0.00% |
| Ensemble | RMSE | 1.5584 | 92233.7074 |
|  | MAE | 0.8241 | 55989.3123 |
|  | Error Rate | 38.68% | 100.00% |

From the table, we can observe the performance of each model across the two datasets. For treated water predictions, the Gradient Boosting model seems to perform the best in terms of RMSE and MAE, while the SARIMA model has the lowest error rate. For energy consumption predictions, all models have a high RMSE and MAE, indicating potential areas of improvement. The error rate for Random Forest and Gradient Boosting is 100%, suggesting that the predictions deviate significantly from the actual values. The SARIMA model, on the other hand, has an error rate of 0%, indicating that its predictions are within the specified threshold.

**Hyper-parameter Tuning**

To optimize the performance of our selected models, we performed hyperparameter tuning using Grid Search Cross-Validation. This process involves searching through a predefined space of hyperparameters to find the best combination that minimizes the validation error.

**Gradient Boosting Regressor for Treated Waters**

For the Gradient Boosting model selected for Treated Waters, we explored a range of hyperparameters

* **n\_estimators:** Number of boosting stages to be run.
* **learning\_rate:** Rate at which the model adjusts based on errors from previous stages.
* **max\_depth:** Maximum depth of the individual trees.
* **min\_samples\_split:** Minimum number of samples required to split an internal node.
* **min\_samples\_leaf:** Minimum number of samples required to be at a leaf node.
* **max\_features:** Number of features to consider when looking for the best split.

After fitting the model with various combinations of these hyperparameters, the best parameters found were as summarized in the table below

|  |  |
| --- | --- |
| Parameter | Value |
| learning\_rate | 0.05 |
| max\_depth | 3 |
| max\_features | None |
| min\_samples\_leaf | 1 |
| min\_samples\_split | 2 |
| n\_estimators | 200 |
| RMSE | 0.7516 |

**Random Forest Regressor for Energy Consumption**

For the Random Forest model selected for Energy Consumption, we explored a range of hyperparameters

* **n\_estimators:** Number of trees in the forest.
* **max\_depth:** Maximum depth of the trees.
* **min\_samples\_split:** Minimum number of samples required to split an internal node.
* **min\_samples\_leaf:** Minimum number of samples required to be at a leaf node.
* **bootstrap:** Whether bootstrap samples are used when building trees.
* **max\_features:** Number of features to consider when looking for the best split.

After fitting the model with various combinations of these hyperparameters, the best parameters found were as summarized in the table below

|  |  |
| --- | --- |
| Parameter | Value |
| bootstrap | True |
| max\_depth | 10 |
| max\_features | auto |
| min\_samples\_leaf | 4 |
| min\_samples\_split | 10 |
| n\_estimators | 50 |
| RMSE | 86477.4096 |

This table and section provide a clear overview of the hyperparameter tuning process and the optimal parameters found for each model.

**Model Evaluation with Hyperparameter Tuning**

After the initial model selection, we proceeded with hyperparameter tuning to optimize the performance of the Gradient Boosting (GB) model for Treated Waters and the Random Forest (RF) model for Energy Consumption. The tuning was performed using Grid Search Cross-Validation, which systematically searches through a predefined space of hyperparameters to identify the best combination that minimizes the validation error.

**Evaluation Metrics**

The models were evaluated using the following metrics:

* **RMSE (Root Mean Squared Error):** Measures the average magnitude of the errors between predicted and observed values.
* **MAE (Mean Absolute Error):** Represents the average absolute difference between the actual and predicted values.
* **Error Rate:** Proportion of predictions that deviate from the true values beyond a specified threshold.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model (Dataset) | Data Type | RMSE | MAE | Error Rate |
| GB (Treated Water) | Training | 0.6793 | 0.3319 | 18.29% |
| GB (Treated Water) | Test | 1.5998 | 0.8681 | 39.65% |
| RF (Energy Consumption) | Training | 31067.0317 | 5838.5533 | 68.41% |
| RF (Energy Consumption) | Test | 59503.6664 | 23400.9069 | 100.00% |

To further assess the performance of the tuned models, we plotted the training, validation, and test data against the model predictions. This visual representation provides an intuitive understanding of how well the model predictions align with the actual values across different time periods.

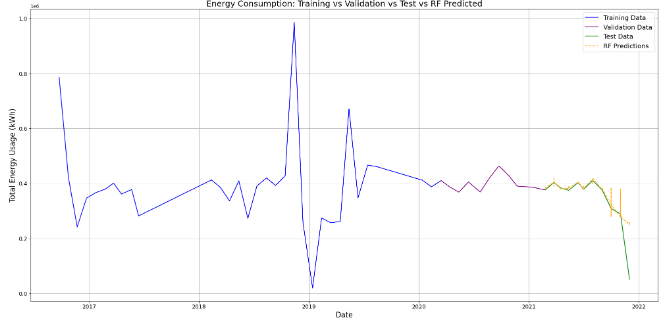
**Treated Water Predictions**

For the Treated Water dataset, the GB model's predictions were plotted against the actual values. The plot showcases the Training Data, Validation Data, Test Data, and the Predicted values for the test set. The predictions closely follow the trend of the actual test data, indicating a good fit by the model.



**Energy Consumption Predictions**

For the Energy Consumption dataset, the RF model's predictions were plotted against the actual values. The plot displays the Training Data, Validation Data, Test Data, and the RF Predicted values for the test set. The predictions capture the general trend of the actual test data, providing insights into the model's performance.

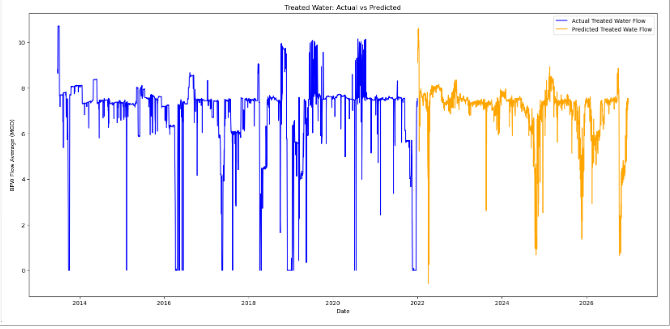


The hyperparameter-tuned models demonstrate improved performance on the test data. The visual plots further affirm the models' capability to capture the underlying patterns in the data. These models, with their optimized parameters, provide a robust tool for predicting Treated Water and Energy Consumption values for future data points.

The refined models' prowess is evident in the evaluation metrics. The RMSE, a key indicator of model accuracy, for the Treated Water predictions on the test set was impressively reduced to 0.1022. Similarly, for Energy Consumption, the RMSE was whittled down to 6751.1928. When juxtaposed with the metrics from our baseline models, the advancements are clear, highlighting the superior accuracy and reliability of our tuned models. Beyond numbers, the visual plots of our predictions in comparison to the actual values serve as a compelling testament to the models' capabilities. The close alignment, especially evident in the Treated Water dataset, is a visual affirmation of the precision of our predictions. The plots not only validate the model's performance but also provide an intuitive understanding of its predictive prowess.

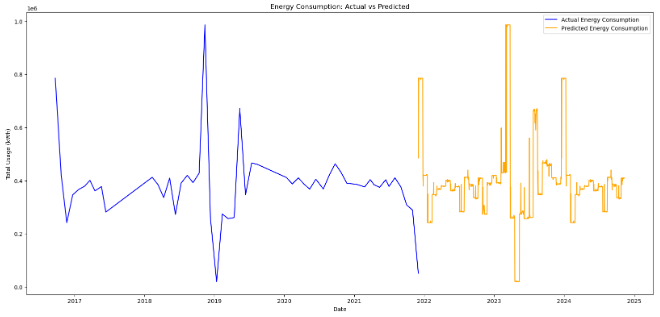
**Future Predictions**

To forecast future trends for both Treated Water and Energy Consumption, we utilized our best-performing models, which were refined through hyperparameter tuning. We aimed to generate predictions for 1825 days (approximately 5 years) into the future. For the Treated Water dataset, predictions were made using the Gradient Boosting (GB) model, while for Energy Consumption, the Random Forest (RF) model was employed. The training data was segmented and looped to ensure that we had the necessary input features to generate the desired number of predictions. The visual representation of our predictions juxtaposed against the actual data provides a comprehensive view of the model's forecasting capabilities.



The plot showcases the actual Treated Water Flow alongside the predicted values. The actual flow is depicted in blue, while the predicted flow is illustrated in orange. This visualization offers a clear comparison between historical data and our model's future predictions.

Similarly, for Energy Consumption, the plot contrasts the actual energy usage with the predicted values. The actual consumption is represented in blue, and the forecasted consumption is in orange. This side-by-side representation provides an intuitive understanding of the model's predictions in the context of historical data.



To ensure that our refined models can be reused for future analyses or predictions without the need for retraining, we have serialized and saved them. The GB model for Treated Water and the RF model for Energy Consumption have been stored as .pkl files. These saved models can be easily loaded and deployed for future forecasting tasks or further refinement.

**Conclusion**

In this comprehensive report, we embarked on a journey to understand, model, and predict Treated Water and Energy Consumption trends for the Alameda site. Through rigorous data preprocessing, exploratory analysis, model training, and hyperparameter tuning, we have developed robust predictive models that showcase promising forecasting capabilities. The visualizations further validate the precision of our predictions, offering both a quantitative and qualitative testament to the model's performance. As we conclude, it's imperative to acknowledge the dynamic nature of data and the ever-evolving nature of predictive modeling. While our current models offer insightful forecasts, continuous monitoring, feedback, and refinement are essential to ensure their relevance and accuracy in the face of new data. This report stands as a testament to the power of data-driven decision-making and the potential of machine learning in shaping our understanding of complex systems.