**Technical Report for Tampa Site**

**Overview**

The analysis presented in this report is centered around the Tampa Site's operations, specifically focusing on predicting the treated water flow and energy consumption at Tampa Site. Utilizing a dataset that captures various metrics such as rainfall, total treated water leaving the plant, hours the plant is in operation, raw water input to the plant, and many other attributes that can determine the variations in the pattern of treated Water flow and energy consumption at Tampa Site, the aim is to develop predictive models that can accurately forecast future values of these key metrics by leveraging historical data, advanced machine learning techniques, and time series forecasting models, this report will provide actionable insights that can enhance the efficiency and sustainability of the Tampa Site's operations through predictions made for Treated Water Flow and Energy consumption.

**Data Preprocessing and Feature Engineering**

Before diving into any data analysis or modeling, it's crucial to ensure that the dataset is in the right shape, free from inconsistencies, and enriched with features that can enhance the predictive power of the models. The Data Preprocessing and Feature Engineering phase is dedicated to these tasks. In this phase, raw data is transformed and refined into a more suitable format for analysis. From handling missing values and correcting data anomalies to engineering new features that can provide additional insights, this section delves into the various steps taken to prepare the data for subsequent analysis and modeling.

**Data Importation**

The initial step involved importing data from two separate Excel files. These files contained various metrics and attributes relevant to the Tampa Site's operations. The table below provides a clear overview of each column in the datasets used for this site along with its corresponding data type after merging the 2 imported data by date.

|  |  |
| --- | --- |
| Column Name | Data Type |
| Date | datetime64[ns] |
| Year | int64 |
| Month | object |
| Day of the Month | object |
| Rain fall,Inches | float64 |
| TotalTreated water leaving plant,MG | float64 |
| Hours Plant in operation | float64 |
| Total Raw water to plant,MG | float64 |
| Backwas,Thousand Gallons | float64 |
| Peak demand into distribution,MDG | float64 |
| Nomeasurements Recorded | float64 |
| No Measuremnts Required | float64 |
| Average Daily Turbidity,NTU | float64 |
| Maximum DailyTurbidity,NTU | float64 |
| Log Inactivation,Giardia | float64 |
| Log Inactivation,Viruses | float64 |
| Adjusted Daily Consumption, Kwh | int64 |

**Data Cleaning**

* **Handling Non-Numeric Values:** A predefined list of columns was identified to check for non-numeric values. This list included metrics such as 'Rain fall,Inches', 'TotalTreated water leaving plant,MG', 'Hours Plant in operation', and several others. For each column in this list, the data was converted to numeric format. Any non-numeric values encountered during this conversion were coerced to NaN (Not a Number). Subsequently, any rows containing NaN values in these specified columns were dropped from the dataset to maintain data consistency and integrity.
* **Index Resetting:** After the data cleaning operations, the index of the dataset was reset to ensure a continuous sequence.
* **Year Correction:** An anomaly was detected in the 'Year' column where the year "20222" was present. This was corrected to "2022".
* **Date Conversion:** The 'Year', 'Month', and 'Day of the Month' columns were combined and converted into a single 'Date' column in a datetime format. This step streamlined the dataset and facilitated time series analysis.

**Data Extraction and Merging**

* **Column Extraction:** From the cleaned dataset, a subset of columns was extracted to focus on the most relevant metrics for the analysis. This subset included columns such as 'Date', 'Rain fall,Inches', 'TotalTreated water leaving plant,MG', and others.
* **Merging Datasets:** The 'Date' column in the second dataset (data1) was also converted to a datetime format to ensure consistency. The two datasets were then merged based on the 'Date' column. This step combined the information from both Excel files into a single, comprehensive dataset.

**Feature Engineering**

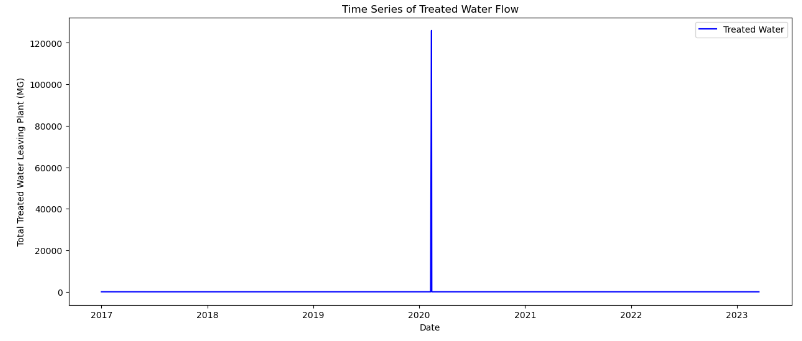
* **One-Hot Encoding:** The 'Month' column, which represented the months as categorical data, was one-hot encoded. This process converted the categorical month data into a format suitable for machine learning models, where each month was represented as a separate binary column.
* **Data Type Conversion:** The 'Day of the Month' column was checked for its data type. If found to be an object type, it was explicitly converted to an integer data type.

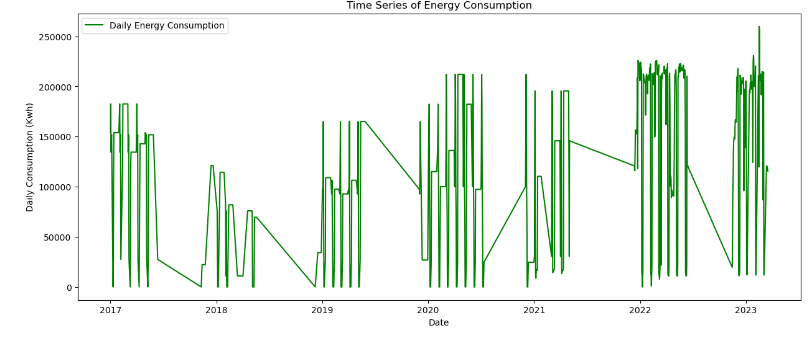
**Data Inspection Post-Processing**

* **Data Types Check:** A check was conducted post-processing to ensure that each column in the dataset had the appropriate data type.
* **Missing Values Inspection:** The dataset was inspected for any missing values post-processing to ensure that there were no gaps in the data.

**Descriptive Statistics**

Basic statistics such as mean, median, standard deviation, minimum, and maximum values were generated for each column to provide a snapshot of the data's distribution post-processing.





A closer examination of the descriptive statistics, particularly for the 'TotalTreated water leaving plant,MG' (Treated Water Flow) column, indicated the potential presence of outliers. The values in the descriptive statistics table, especially the maximum value being significantly higher than the mean and median, hinted at this discrepancy. Further evidence of these outliers was observed in the initial time plot of the target variables. The plot showcased spikes or deviations from the general trend, reinforcing the need for outlier handling in the data preprocessing phase. The table below provides a concise summary of the descriptive statistics for the 'Total Treated water leaving plant, MG' and 'Adjusted Daily Consumption, Kwh' columns.

Descriptive Statistics Summary for Selected Features

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Mean | Median | Std | Min | Max | 25th | 75th |
| Treated water | 157.52 | 14.35 | 4251.48 | 0 | 125918.00 | 12.10 | 15.84 |
| Energy Consumption | 122161.19 | 134468.00 | 72051.84 | 0 | 259874.00 | 81950.00 | 192656.00 |

As observed from the table above, we can confirm that the 'TotalTreated water leaving plant,MG' column has a significant difference between its mean and median values, suggesting a skewed distribution. Furthermore, the vast difference between its 75th percentile and maximum value indicates the presence of outliers. The 'Rain fall,Inches' column has a majority of its values as 0, as indicated by both its mean and median values being close to 0. The 'Year' column ranges from 2017 to 2023, providing data across multiple years. By examining the descriptive statistics table and the initial time plot of the target variables, it's evident that there are outliers in the dataset, especially in the 'TotalTreated water leaving plant,MG' column. This observation reinforces the importance of the outlier removal steps taken during data preprocessing.

**Outlier Removal**

To ensure that extreme values or outliers do not skew the analysis, a percentile-based capping method was employed. Specifically, for columns such as 'TotalTreated water leaving plant,MG', 'Total Raw water to plant,MG', and 'Peak demand into distribution,MDG', any values beyond the 99th percentile were identified as outliers. These outliers were then removed from the dataset. This step helps in retaining only the most representative data points and reduces the potential of extreme values influencing the model's performance.

**Validation of Numeric Values**

A custom function named check\_string was created to validate if a given string contains only numeric characters (digits) and periods (decimal points). This function is essential to ensure that the dataset's numeric columns do not contain any unexpected characters.

**Identification of Non-Numeric Values**

A dictionary, columns\_with\_non\_integers, was initialized to store columns that might contain non-numeric values. Each column in the dataset, except for the 'Date' column, was looped through to check every value using the check\_string function. If any non-numeric value was found, it was added to the dictionary under the respective column.

**Display of Non-Numeric Values**

For each column identified to have non-numeric values, the first ten such values were displayed. This step provides a quick snapshot of the kind of non-numeric data present in the dataset, aiding in further cleaning or transformation if required.

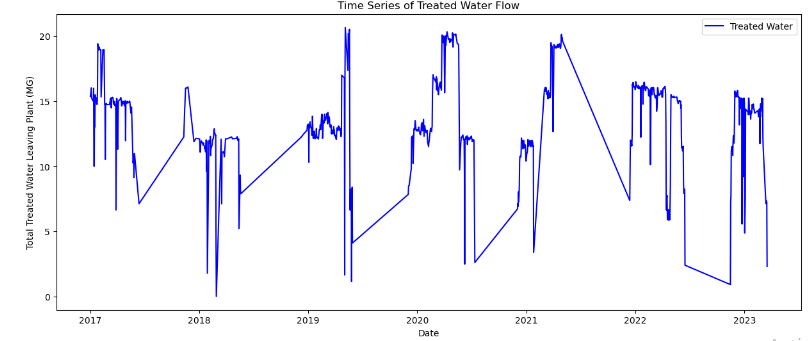
The data preprocessing and feature engineering phase was pivotal in shaping the dataset for subsequent analysis and modeling. By addressing non-numeric values, handling outliers, and converting date-related columns into a unified datetime format, the data was transformed into a more structured and analyzable form. The merging of data from two distinct Excel files ensured a comprehensive dataset, while the generation of dummy variables for categorical columns like 'Month' facilitated a more nuanced understanding of time-based patterns. The meticulous steps taken during this phase not only enhanced the data's quality but also laid a robust foundation for the exploratory and modeling stages that followed. This process ensures the importance of thorough preprocessing in ensuring the accuracy and reliability of any data-driven analysis.

**Exploratory Data Analysis**

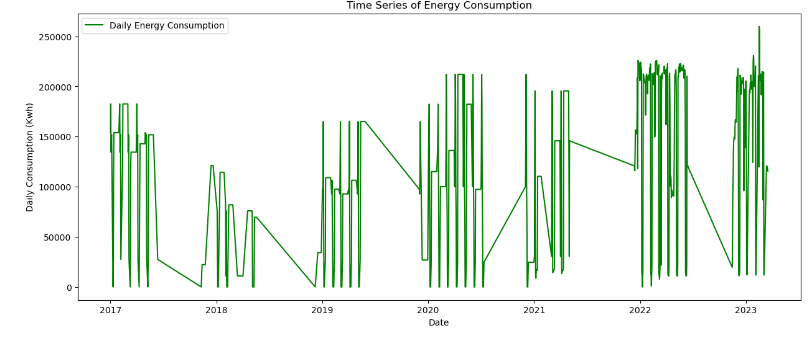
Exploratory Data Analysis (EDA) is crucial in understanding the underlying patterns, relationships, and structures within the dataset. This section examines the visual and statistical exploration of the dataset, focusing primarily on the target variables 'TotalTreated water leaving plant,MG' and 'Adjusted Daily Consumption, Kwh' columns.

**Time Series Analysis**

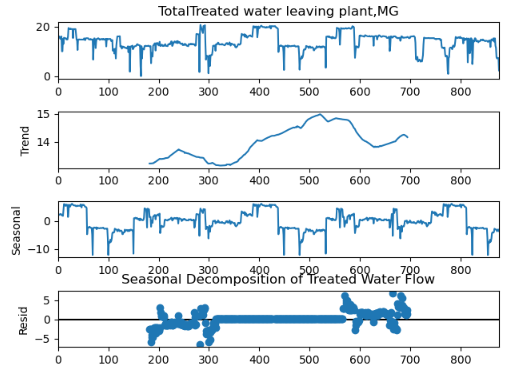
A time series plot was generated to visualize the flow of treated water over time. The x-axis represents the 'Date', while the y-axis denotes the 'Total Treated Water Leaving Plant (MG)'. The plot provides a visual representation of how the treated water flow has varied over the dataset's timeframe.



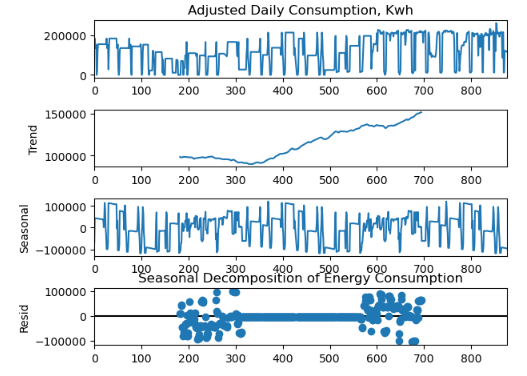
Similarly, a time series plot was created to showcase the daily energy consumption. The x-axis represents the 'Date', and the y-axis indicates the 'Daily Consumption (Kwh)'. This visualization helps in understanding the energy consumption patterns and any anomalies or spikes in consumption.

**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 Testing**

To determine the stationarity of the time series data, the Augmented Dickey-Fuller (ADF) test was employed. Stationarity is a vital property for time series forecasting, as many forecasting models assume that the data is stationary. The ADF test was applied to the 'TotalTreated water leaving plant,MG' column. The test results, including the ADF statistic, p-value, and critical values, were reported. Based on the p-value, a conclusion was drawn regarding the stationarity of the data.

Similarly, the ADF test was conducted for the 'Adjusted Daily Consumption, Kwh' column. The test's outcome provides insights into whether the energy consumption data is stationary or not.

|  |  |  |
| --- | --- | --- |
| Metric/Variable | Treated water | Energy Consumption |
| ADF Statistic | -4.5335 | -6.5851 |
| p-value | 0.000171 | 7.35e-09 |
| Critical Value (1%) | -3.4381 | -3.4381 |
| Critical Value (5%) | -2.8650 | -2.8650 |
| Critical Value (10%) | -2.5686 | -2.5686 |
| Stationarity Conclusion | Stationary | Stationary |

As observed from the table above, both 'TotalTreated water leaving plant,MG' and 'Adjusted Daily Consumption, Kwh' time series data are found to be stationary based on the Augmented Dickey-Fuller test results. The p-values for both variables are significantly less than 0.05, providing strong evidence against the null hypothesis. This indicates that both datasets have no unit root and are stationary, making them suitable for time series forecasting.

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 preliminary phase of model exploration, we undertook an initial training exercise. The primary objective of this phase was to gauge the efficacy of the chosen models when trained and tested on a balanced data partition of 50:50. This approach ensures that the models are exposed to a representative sample of the data during training, while also having an ample amount of unseen data for testing.

**Data Partitioning**

The dataset was split into two equal halves:

* **Training set:** 50% of the total data
* Test set: 50% of the total data

For the training and testing datasets

* **Training set size:** 50% of the total data
* **Test set size:** 50% of the total data

**Model Training**

Three models were trained for both Treated Water and Energy Consumption:

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

**Model Evaluation**

The models' predictions were evaluated against the actual test data using the following metrics:

**Root Mean Squared Error (RMSE):** Measures the average magnitude of the errors.

**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 initial training provided a comprehensive understanding of the models' performance. The balanced data partitioning ensured that the models were not overfitting to the training data and could generalize well to unseen data. The evaluation metrics further highlighted the strengths and potential areas of improvement for each model, setting the stage for further optimization and fine-tuning in subsequent phases.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Treated Water | | | Energy Consumption | | |
| Model | RMSE | MAE | Error | RMSE | MAE | Error |
| Random Forest | 1.4429 | 0.7100 | 0.3162 | 58724.58 | 44508.52 | 1 |
| Gradient Boosting | 1.5221 | 0.8045 | 0.3115 | 61749.70 | 50193.39 | 1 |
| SARIMA | 3.9810 | 2.7640 | 0.7471 | 147498.62 | 124239.08 | 1 |
| Ensemble | 2.0586 | 1.3130 | 0.6581 | 58180.73 | 43608.72 | 1 |

**Model Training**

Model training is a crucial step in the data analysis pipeline, where we employ various algorithms to learn patterns from the historical data. This learning enables the models to make future predictions. In this analysis, we have employed both time series and machine learning models to predict the treated water flow and energy consumption at the Tampa site.

**SARIMA Model**

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a time series forecasting method that combines both the ARIMA and seasonal decomposition capabilities. It's particularly suitable for datasets with clear seasonality patterns.

**Treated Water Flow Forecasting with SARIMA**

* The SARIMA model was trained on the TotalTreated water leaving plant,MG column of the training dataset.
* The model parameters were set to order = (1, 1, 1) and seasonal\_order=(1, 1, 1, 12), indicating a monthly seasonality.
* The model was successfully fitted without any discrepancies.

**Energy Consumption Forecasting with SARIMA**

* Similarly, another SARIMA model was trained on the Adjusted Daily Consumption, Kwh column of the training dataset.
* The model parameters remained consistent with the treated water model.
* This model was also successfully fitted, ensuring its readiness for forecasting.

**Machine Learning Models: Random Forest and Gradient Boosting**

Random Forest and Gradient Boosting are ensemble learning methods used for regression tasks. They are known for their high accuracy, ability to handle large data sets with higher dimensionality, and their ability to handle missing values.

**Data Preparation for Machine Learning Models**

* The dataset was split into training (60%), validation (20%), and test (20%) sets.
* Features and target variables were separated. The Date, TotalTreated water leaving plant,MG, and Adjusted Daily Consumption, Kwh columns were excluded from the feature set.
* Treated Water Flow Forecasting with Machine Learning:
* The Random Forest model was trained on the treated water flow data (TotalTreated water leaving plant,MG).
* The Gradient Boosting model was also trained on the same data, serving as an alternative model for performance comparison.

**Energy Consumption Forecasting with Machine Learning**

* Similarly, both Random Forest and Gradient Boosting models were trained on the energy consumption data (Adjusted Daily Consumption, Kwh).

In conclusion, the model training employed both time series and machine learning methodologies to harness the patterns embedded in treated water and energy consumption at Tampa Site. The SARIMA model, with its inherent capability to capture seasonality, and the ensemble machine learning models, known for their robustness and accuracy, were trained on the datasets for treated water flow and energy consumption. By ensuring that each model is aptly trained and parameterized, we have laid a solid foundation for the subsequent steps of model validation and forecasting. The rigor and precision invested in this phase underscore its significance in our overarching goal of achieving reliable and insightful predictions.

**Model Prediction**

The model prediction stage is a critical step in the machine learning pipeline, where the trained models are used to generate forecasts. In this phase, we utilized our trained models (SARIMA, Random Forest, Gradient Boost and a combination of the three models through ensembles) to make predictions on the validation set, which is a subset of our data that was not used during the training process. This allows us to evaluate the performance of our models in a realistic scenario, mimicking how they would perform in real-world applications.

**Predictions Using Time Series Model (SARIMA)**

* **SARIMA Model for Treated Water:** The SARIMA model, specifically trained for the treated water flow, was used to predict the outcomes for the length of the validation set. The predictions were then converted into a numpy array for consistency and stored in the sarima\_pred\_treated\_validation array.
* **SARIMA Model for Energy Consumption:** Similarly, the SARIMA model for energy consumption was applied to predict the outcomes for the validation set's duration. The results were converted to a numpy array and saved in the sarima\_pred\_energy\_validation array.

**Predictions Using Machine Learning Models**

* **Random Forest (RF) Model for Treated Water:** The trained Random Forest model for predicting the treated water flow was applied to the validation set's features (X\_validation\_actual). The resulting predictions were stored in the rf\_pred\_treated\_validation array.
* **Gradient Boosting (GB) Model for Treated Water:** Similarly, the Gradient Boosting model for treated water flow was used to predict the validation set's outcomes, with the results saved in the gb\_pred\_treated\_validation array.
* **Random Forest (RF) Model for Energy Consumption:** For predicting daily energy consumption, the trained Random Forest model was applied to the validation set, and the predictions were captured in the rf\_pred\_energy\_validation array.
* **Gradient Boosting (GB) Model for Energy Consumption:** The Gradient Boosting model for energy consumption was also used to forecast the validation set's outcomes, with the results stored in the gb\_pred\_energy\_validation array.

**Ensemble Predictions**

To leverage the strengths of both machine learning and time series models, we adopted an ensemble approach. By averaging the predictions from the Random Forest, Gradient Boosting, and SARIMA models, we aimed to achieve a more robust and stable forecast.

* Ensemble Prediction for Treated Water: The predictions from the RF, GB, and SARIMA models for treated water were averaged, resulting in the ensemble\_pred\_treated\_validation array.
* **Ensemble Prediction for Energy Consumption:** Similarly, the predictions from the three models for energy consumption were averaged, yielding the ensemble\_pred\_energy\_validation array.

The model prediction phase is instrumental in understanding the potential performance of our models on new, unseen data. By employing both individual and ensemble prediction strategies, we ensure a comprehensive approach, maximizing the potential accuracy and reliability of our forecasts. The subsequent steps will involve evaluating these predictions against the actual outcomes in the validation set to assess the models' performance metrics and make any necessary refinements.

**Model Evaluation**

It is important to evaluate the trained models to gain insights about the performance of the trained models, allowing us to gauge their accuracy, reliability, and potential areas for improvement. In this section, we will examine the evaluation metrics used to assess our models' performance on the validation set and discuss the implications of the results.

**Evaluation Metrics**

To assess the performance of our models, we employed three key metrics:

* **Root Mean Squared Error (RMSE):** This metric provides the square root of the average squared differences between the predicted and actual values. A lower RMSE indicates a better fit of the model to the data.
* **Mean Absolute Error (MAE):** MAE calculates the average absolute differences between the predicted and actual values. It provides a clear measure of prediction accuracy.
* **Error Rate:** Defined as the Mean Absolute Percentage Error (MAPE), this metric gives the average percentage difference between the actual and predicted values. It offers a relative measure of the prediction error.

**Model Evaluation Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Treated Water Validation | | | Energy Consumption Validation | | |
| Model | RMSE | MAE | Error Rate | RMSE | MAE | Error Rate |
| RF | 0.11 | 0.05 | 0.41% | 4876.59 | 3422.66 | 9.53% |
| GB | 0.07 | 0.03 | 0.32% | 4332.31 | 2619.35 | 13.41% |
| SARIMA | 5.30 | 4.99 | 31.39% | 54726.09 | 30672.51 | 1585.03% |
| Ensemble | 1.77 | 1.66 | 10.50% | 19062.70 | 11461.67 | 531.42% |

From the results summarized in the table above, we observe that; for Treated Water Validation, the Gradient Boosting (GB) model outperforms the other models with the lowest RMSE, MAE, and Error Rate. The SARIMA model, however, shows a significantly higher error, suggesting that it might not be the best fit for this dataset. For Energy Consumption Validation, both the RF and GB models perform relatively well, with GB having a slight edge in terms of RMSE and MAE. However, the SARIMA model's error rate is exceedingly high, indicating potential overfitting or model mis-specification.

The ensemble approach, which averages predictions from all three models, provides a balanced performance, though it doesn't outperform the individual GB model in this case.

Model evaluation is instrumental in understanding the strengths and weaknesses of our predictive models. The results highlight the importance of choosing the right model for specific datasets and the potential benefits of ensemble methods. While the GB model showcased superior performance for both treated water and energy consumption predictions, there's room for further optimization, especially for the gradient boost model which lead us to hyperparameter tuning on the best performing model for both Treated Water and Energy Consumption.

**Hyperpermeter Tunning and Optimization of the Gradient Boost Model**

To optimize the parameters of the gradient boost regression model that appear to be the most efficient among the models explored earlier in the report, this analysis delved into the hyperparameter tuning process for the Gradient Boosting (GB) model, detailing the methodology, results, and visualizations of the predictions. To optimize the model, we employed the GridSearchCV method, which performs an exhaustive search over a specified parameter grid. The parameters considered for tuning included:

* **Number of Estimators (n\_estimators):** The number of boosting stages to be run.
* **Learning Rate (learning\_rate):** The step size at each iteration while moving towards a minimum of the loss function.
* **Maximum Depth (max\_depth):** The maximum depth of the individual regression estimators.
* **Subsample (subsample):** The fraction of samples used for fitting the individual base learners.
* **Minimum Samples Split (min\_samples\_split):** The minimum number of samples required to split an internal node.

The grid search was conducted using a 3-fold cross-validation to ensure robustness in the results.

**Results**

|  |  |  |
| --- | --- | --- |
| Parameter | Optimal Value: Treated Water | Optimal Value: Energy Consumption |
| Learning Rate | 0.1 | 0.1 |
| Max Depth | 3 | 3 |
| Min Samples Split | 4 | 3 |
| Number of Estimators | 50 | 100 |
| Subsample | 0.9 | 0.9 |

**Model Evaluation**

After tuning, the models were evaluated on both the training and test datasets. The evaluation metrics used were RMSE, MAE, and Error Rate.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Treated Water | | | Energy Consumption | | |
| **RMSE** | **MAE** | **Error Rate** | **RMSE** | **MAE** | **Error Rate** |
| Training | 0.0116 | 0.0086 | 0.0% | 17.8492 | 5.2511 | 0.7961% |
| Test | 0.1022 | 0.0503 | 1.76% | 6751.1928 | 3740.1486 | 1.0% |

The predictions of the optimized models were visualized against the actual values for both Treated Water and Energy Consumption. The plots showcase the performance of the models across different time periods, providing a visual representation of their accuracy.



Figure 4: Treated Water: Training vs Test vs Predicted

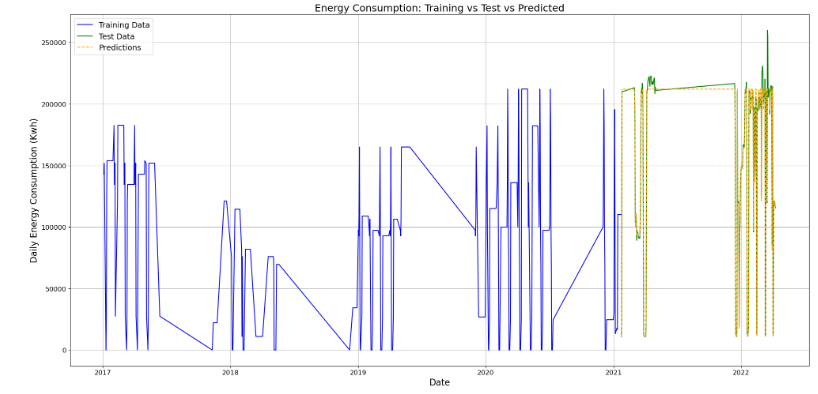


Figure 5: Energy Consumption: Training vs Test vs Predicted

Through the process of hyperparameter tuning, we've significantly refined the Gradient Boosting (GB) model for both Treated Water and Energy Consumption predictions. The optimal parameters identified, such as a learning rate of 0.1 and a subsample rate of 0.9, demonstrate the importance of fine-tuning to achieve better model performance. The evaluation metrics provide a quantitative testament to this improvement. For instance, the RMSE for Treated Water on the test set was reduced to 0.1022, and for Energy Consumption, it was 6751.1928. These metrics, especially when compared to the baseline models, underscore the enhanced accuracy and reliability of our tuned models.

Furthermore, the visual representations of our predictions, juxtaposed with the actual values, offer a clear visual affirmation of the model's capabilities. The plots reveal a close alignment between predicted and actual values, especially for the Treated Water dataset. In essence, this hyperparameter tuning exercise has not only bolstered the performance of our GB models but has also underscored the critical importance of optimization in the machine learning workflow. As we move forward, these optimized models stand as a robust foundation for any further analysis or predictive tasks related to this dataset.

**Future Predictions**

Leveraging the power of our optimized Gradient Boosting (GB) models, we embarked on predicting future values for both Treated Water and Energy Consumption for the Tampa site.

* **Data Preparation:** To generate predictions for 1825 days into the future, we utilized the existing X\_train\_actual dataset. Recognizing that the length of this dataset might not suffice for our prediction horizon, we employed a technique of concatenating the dataset with itself multiple times. This ensured that we had a sufficiently large dataset, X\_future, to feed into our models.
* **Model Predictions:** With our X\_future dataset ready, we invoked our previously tuned GB models - best\_gb\_model\_treated for Treated Water and best\_gb\_model\_energy for Energy Consumption. The predictions were then stored in two lists: treated\_predictions and energy\_predictions.
* **Data Structuring:** Post-prediction, we structured our results into a comprehensive DataFrame, pred\_df, which includes columns for 'Date', 'Treated\_Water\_Predictions', and 'Energy\_Consumption\_Predictions'. The date column was generated starting from the day after the last date in our original dataset and spanned the length of our prediction horizon.
* **Data Export:** For ease of access and further analysis, the pred\_df DataFrame was exported to a CSV file named 'Future\_Predictions.csv'.

**Visualization**

Visual representation aids in intuitively understanding the trajectory of our predictions in relation to actual values.

* **Treated Water Predictions**: The plot juxtaposes the actual Treated Water values (in blue) against the predicted values (in orange dashed lines). This visualization provides a clear picture of how our model envisions the future trends of Treated Water in comparison to historical data. The plot is titled 'Treated Water: Actual vs Predicted' and has been saved as "Predicted Treated water for Tampa Site.png" for reference.
* **Energy Consumption Predictions**: Similarly, for Energy Consumption, the green line represents the actual values, while the red dashed line showcases our model's predictions. The title 'Energy Consumption: Actual vs Predicted' aptly captures the essence of the plot.

Through this exercise, we've extended the utility of our models beyond retrospective analysis to prospective forecasting. These predictions, both for Treated Water and Energy Consumption, serve as a roadmap for future planning and resource allocation for the Tampa site. While models provide a data-driven glimpse into the future, it's essential to revisit and recalibrate them periodically with fresh data to ensure continued accuracy and relevance.

Refer to the attached plots for a visual representation of the actual versus predicted values for both target variables.

