**Technical Report for San Antonio Site**

**Overview**

This report looks at the San Antonio Site's operations, mainly focusing on predicting how much treated water the site produces and how much energy it uses. We've used a dataset that includes information like rainfall, how much treated water the site sends out, how long the plant runs, and how much raw water comes in. These details help us understand the patterns of water treatment and energy use at the San Antonio Site. The goal is to use past data and modern prediction methods to make accurate forecasts. With these predictions, we hope to give useful advice that can help the San Antonio Site run more efficiently and sustainably. By knowing the future values for water treatment and energy use, this report aims to help the San Antonio Site make better decisions.

**Data Preprocessing and Feature Engineering**

Before diving deep into the analysis, it's essential to set the stage with clean and well-organized data. Making sure our data is in the right shape, free from mistakes, and enriched with useful details is key to making our later steps fruitful. In this section, we'll walk through the steps we took to tidy up and enhance our data, ensuring it's ready for the tasks ahead.

**Data Importation**

The foundation of any data analysis is the quality and structure of the dataset. For the San Antonio Site, we sourced our data from a single comprehensive CSV file named 'San Antonio.csv'. This dataset captures a range of metrics that can influence the treated water flow and energy consumption at the San Antonio Site.

Having all the necessary attributes in a single dataset simplifies the analysis process and ensures that we have a holistic view of the operations at the San Antonio Site. After importing the data, it's crucial to understand the nature and type of each column. This understanding not only aids in further preprocessing tasks but also ensures that we leverage the data effectively during the modeling phase. To provide a clearer perspective, the table below presents an overview of the columns in the dataset along with their respective data types

|  |  |
| --- | --- |
| Column Name | Data Type |
| DATE | object |
| Water Flows | float64 |
| RAW WATER TURBIDITY (NTU) | float64 |
| RO FEED TURBIDITY (NTU) | float64 |
| RAW WATER TDS (MG/L) | float64 |
| FINISHED WATER TURBIDITY (NTU) | float64 |
| FINISHED WATER TDS (MG/L) | float64 |
| Daily KWH | float64 |
| Daily COST | float64 |

This table offers a structured view of the dataset, detailing the columns and their data types. With this understanding, we can proceed with further data preprocessing, feature engineering, and modeling tasks to derive actionable insights for the San Antonio Site's operations.

**Data Cleaning**

Data cleaning is an essential step in the preprocessing phase, ensuring that the dataset is free from inconsistencies, missing values, and any other anomalies that might skew the analysis. Here's a breakdown of the cleaning process we applied to the San Antonio Site dataset:

**Date Conversion and Indexing**

* The 'DATE' column, which originally was of the object data type, was converted into a datetime format. This conversion facilitates time-based operations and analyses.
* After the conversion, the 'DATE' column was set as the index of the dataset. This restructuring is particularly useful for time series data, allowing for more intuitive data slicing and dicing based on time periods.
* The data was then sorted chronologically to ensure that all entries follow a logical time sequence.

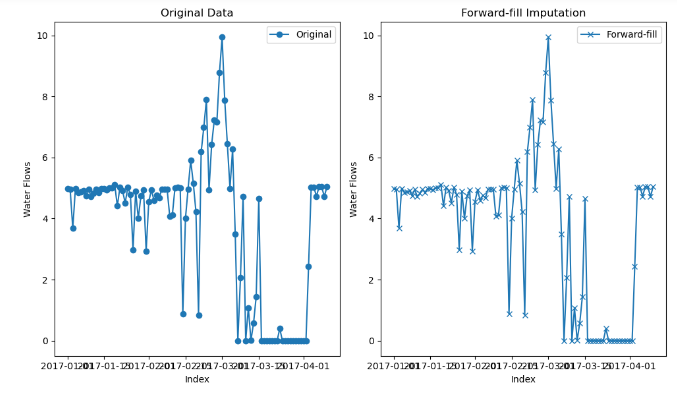
**Handling Missing Values**

Missing values can introduce bias or inaccuracies in the analysis. To address this, we first applied a forward-fill method, which replaces missing values with the last observed non-missing value. For any remaining missing values, especially at the beginning of the dataset where forward-fill might not be applicable, a backward-fill method was used. This method replaces missing values with the next observed non-missing value. As an additional step, polynomial interpolation of order 2 was attempted to estimate missing values based on the existing data. This method can provide a more nuanced estimation of missing values, especially if there's a trend or pattern in the data.

**Visualization of Data Cleaning**

To visually assess the impact of our cleaning methods, we plotted a subset of the 'Water Flows' data before and after applying the forward-fill method.

The original data was plotted alongside the forward-filled data for the first 100 rows to provide a clear comparison. This visualization helps in understanding how the forward-fill method fills gaps in the dataset and ensures continuity in the data sequence.

By the end of this cleaning process, the dataset was in a more structured and consistent state, ready for exploratory data analysis and modeling. The visualizations further confirmed the effectiveness of the cleaning methods applied, showcasing a seamless flow of data without any noticeable gaps or anomalies.  


**Feature Engineering**

Feature engineering involves transforming the raw data into a more usable format, making it easier to work with and enhancing its utility for analysis. For the San Antonio Site dataset, the following engineering steps were undertaken

**Date Conversion**

The 'DATE' column, which was initially in a string format, was converted into a datetime format. This conversion is crucial for any time series analysis as it allows for more intuitive date-based operations.

The format '%Y/%m/%d' was specified during the conversion to ensure that the date strings were interpreted correctly. After the conversion, a quick check confirmed that the 'DATE' column was indeed in the datetime format.

**Feature Extraction from Date**

* With the 'DATE' column now in the datetime format, we could easily extract additional time-related features that might be useful for our analysis.
* Three new columns were created: 'Day', 'Month', and 'Year'. These columns capture the day of the month, the month, and the year from the 'DATE' column, respectively. Such features can be invaluable when analyzing seasonal patterns, yearly trends, or monthly variations in the data.
* By the end of this engineering process, the dataset was enriched with additional time-related features, making it more versatile for subsequent analyses. These new features can provide deeper insights into the time-based patterns and trends present in the data, enhancing the overall quality of the analysis.

**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 San Antonio Site data, lagged features were generated for both water flows and energy consumption:

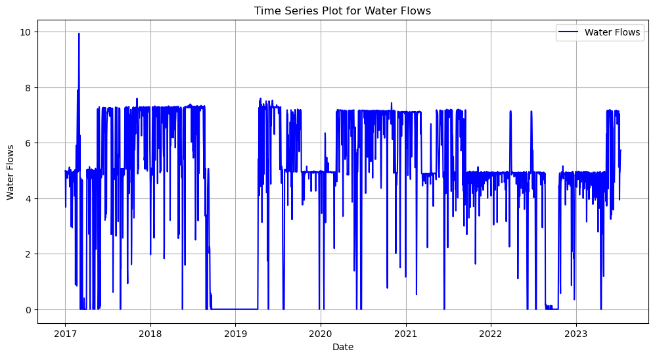
* **Lagged Features for Water Flows:** Three lagged features were created for the ' Water Flows' column in the dataframe. These represent the water flows from the previous three days. The new columns were named 'Water\_Flows\_Lag\_1', 'Water\_Flows\_Lag\_2', and 'Water\_Flows\_Lag\_3', where each number 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 Daily kWh' column in the 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.

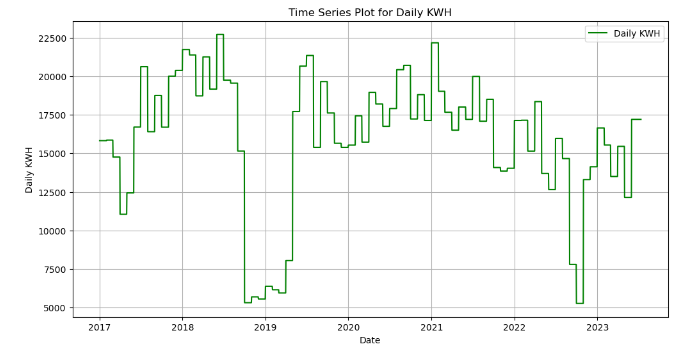
**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves visualizing and understanding the patterns, relationships, anomalies, and structures within the data. For the San Antonio Site dataset, EDA provides insights into the trends and patterns of key metrics, aiding in the subsequent modeling phase.

**Time Series Plots**



A time series plot for 'Water Flows' as displayed above was generated to visualize the trend and patterns over time. The plot provides a clear representation of how the water flows have changed over the dataset's duration. The x-axis represents the date, while the y-axis shows the water flows. The data points are connected with a blue line, providing a continuous view of the water flow trends. From the plot, we can observe any seasonality, trends, or anomalies in the water flows, which can be crucial for understanding the underlying patterns and making accurate predictions.



Similarly, a time series plot for 'Daily KWH' (Daily Kilowatt-Hours) was created as displayed above. This plot visualizes the energy consumption patterns over time. The x-axis represents the date, and the y-axis indicates the daily energy consumption in KWH. The data points are connected with a green line, offering a continuous view of the energy consumption trends. This plot helps in understanding the energy consumption patterns, identifying any regularities or irregularities, and observing how energy consumption has evolved over time.

**Seasonal Decomposition**

Seasonal decomposition is a technique used to break down a time series into its constituent components, namely trend, seasonality, and residuals. This decomposition helps in understanding the underlying patterns in the data, such as long-term trends and recurring seasonal effects. For the San Antonio Site dataset, seasonal decomposition was applied to both the treated water flows and energy consumption metrics.

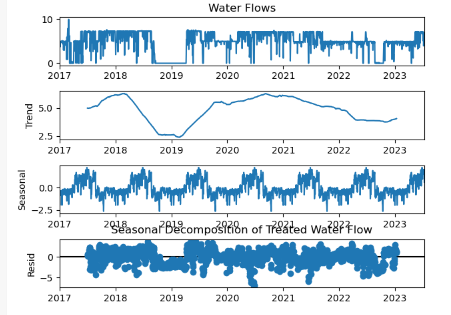
**Seasonal Decomposition of Treated Water Flow**

The treated water flow data was decomposed into its trend, seasonal, and residual components using a period of 365 days, representing a yearly cycle.

**The decomposition plot showcases**

* **Trend:** The long-term movement in the data, highlighting any upward or downward trajectory over time.
* **Seasonality:** The recurring patterns or cycles observed in the data, which repeat over the specified period (in this case, annually).
* **Residual:** The noise or random fluctuations in the data after removing the trend and seasonal components.

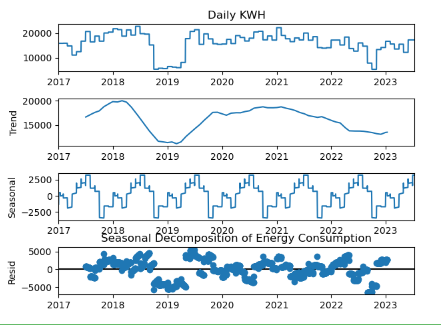
The seasonal decomposition plot as revealed below provides insights into the yearly patterns of treated water flow, helping identify any consistent behaviors or anomalies.



**Seasonal Decomposition of Energy Consumption**

Similarly, the energy consumption data ('Daily KWH') was decomposed using a period of 365 days.

The decomposition plot for energy consumption also presents the trend, seasonal, and residual components, offering a comprehensive view of the energy consumption patterns over time.



This decomposition aids in understanding how energy consumption at the San Antonio Site varies throughout the year and if there are any consistent seasonal effects or anomalies.

Both of these seasonal decomposition plots are included in the report, offering a detailed view of the inherent patterns in the treated water flows and energy consumption at the San Antonio Site. Understanding these patterns is crucial for effective forecasting and making informed decisions based on the data's behavior.

**Stationarity Test**

One of the fundamental assumptions in time series forecasting is that the data is stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed. In simpler terms, it means that the series has constant mean and variance over time. To confirm the stationarity of our data, we employed the Augmented Dickey-Fuller (ADF) test, a formal statistical test for stationarity. The ADF test provides us with an ADF statistic, which we compare against critical values at different confidence levels. A lower ADF statistic compared to the critical values and a p-value less than 0.05 indicates that we can reject the null hypothesis and conclude that the series is stationary.

Based on the ADF test, the results for both the treated water flows and energy consumption metrics are as follows:

|  |  |  |
| --- | --- | --- |
| Metric/Variable | Treated water | Energy Consumption |
| ADF Statistic | -4.4284 | -3.2301 |
| p-value | 0.0002643 | 0.01831 |
| Critical Value (1%) | -3.4331 | -3.4331 |
| Critical Value (5%) | -2.8628 | -2.8628 |
| Critical Value (10%) | -2.5674 | -2.5674 |
| Stationarity Conclusion | Stationary | Stationary |

From the table above, it's evident that the ADF statistic of -4.4284 is less than the critical values at all confidence levels, and the p-value is significantly less than 0.05. This provides strong evidence against the null hypothesis, indicating that the treated water flow data is stationary. For Energy Consumption: The ADF statistic of -3.2301 is also less than the critical values at all confidence levels, and the p-value is less than 0.05. This suggests that the energy consumption data is also stationary.

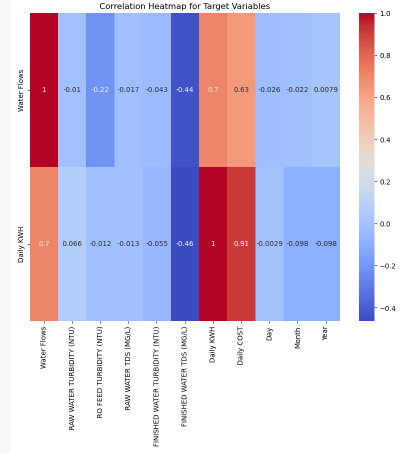
In conclusion, both the treated water flows and energy consumption metrics are stationary, making them suitable for time series forecasting without the need for further differencing or transformations.

**Correlation Matrix Analysis**

Understanding the relationships between different variables in a dataset is crucial for feature selection and model building. Correlation analysis provides insights into how variables change in relation to each other, which can be instrumental in predicting one variable based on another. In this section, we delve into the correlation matrix to understand the relationships between our target variables, 'Water Flows' and 'Daily KWH', and other features in the dataset. As observed from the results,

* **Water Flows and Daily KWH:** There's a strong positive correlation of approximately 0.7046 between 'Water Flows' and 'Daily KWH'. This suggests that as the water flow increases, the energy consumption (KWH) tends to increase as well.
* **Water Flows and Finished Water TDS (MG/L):** There's a notable negative correlation of approximately -0.4390. This indicates that as the TDS (Total Dissolved Solids) in the finished water increases, the water flow tends to decrease.
* **Daily KWH and Finished Water TDS (MG/L):** A negative correlation of -0.4625 is observed between these two variables, suggesting that as the TDS in the finished water increases, the energy consumption tends to decrease.
* **Water Flows and RO Feed Turbidity (NTU):** A negative correlation of -0.2150 suggests that as the turbidity of the RO feed increases, the water flow tends to decrease.
* **Daily KWH and Daily COST:** A strong positive correlation of 0.9121 indicates that as the energy consumption increases, the cost associated with it also increases.
* **Month and Daily KWH:** A negative correlation of -0.0975 suggests a slight decrease in energy consumption as the months progress.

The heatmap for the target variables visually represents these correlations, making it easier to discern the strength and direction of relationships between variables. It's important to note that while correlation indicates a relationship, it does not imply causation. Further analysis, such as regression or causal inference, would be required to understand the underlying mechanisms driving these relationships.



Summarily, the correlation matrix provides valuable insights into the relationships between different features and our target variables. These insights can guide feature selection and model building in subsequent stages of the analysis.

Initial Model Training

The primary objective of the initial training phase was to assess the performance of the selected models when trained and tested on a balanced data partition of 50:50. This approach was adopted to ensure that the models were trained on a representative sample of the data and validated on an equally substantial unseen dataset.

**Data Partitioning**

The dataset was split into two equal parts

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

**Model Training**

For both Water Flows and Daily KWH (Energy Consumption), three distinct models were trained:

* **Random Forest (RF):** An ensemble learning method suitable for both regression and classification tasks.
* **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 model predictions were compared against the actual test data values. The evaluation metrics used were:

* **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 observations.
* **Error Rate:** Calculates the proportion of predictions that deviate from the actual values beyond a specified threshold.

The evaluation results are summarized in the tables below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Treated Water | | | Energy Consumption | | |
| Model | RMSE | MAE | Error | RMSE | MAE | Error |
| Random Forest | 1.0865 | 0.7110 | 0.4291 | 1897.62 | 1316.99 | 0.9992 |
| Gradient Boosting | 1.0301 | 0.6524 | 0.4047 | 1857.01 | 1377.10 | 0.9992 |
| SARIMA | 2.1492 | 1.6572 | 0.9370 | 8124.88 | 6688.84 | 1.0000 |
| Ensemble | 1.3078 | 0.9339 | 0.6406 | 3247.77 | 2615.87 | 0.9992 |

The initial training phase provided valuable insights into the performance of the models. The balanced data partitioning ensured that the models were not biased towards the training data and could generalize well to new data. The evaluation metrics further highlighted the strengths and potential areas of improvement for each model, setting the stage for subsequent optimization and refinement.

**Model Training**

The primary objective of this analysis is to predict the treated water flow and energy consumption at the San Antonio Site. To achieve this, we employed a combination of time series forecasting models and machine learning regression models. This section provides a detailed overview of the model training process for both the treated water flow and energy consumption predictions.

**Data Splitting**

Before training the models, the dataset was split into training, validation, and test sets. The data was sorted chronologically, and the following proportions were used:

* Training set: 60% of the data equals 1428 data points
* Validation set: 20% of the data equals 476 data points
* Test set: 20% of the data equals 477 data points

This split ensures that the models are trained on a majority of the data, validated on a separate set, and finally tested on unseen data to evaluate their performance.

**Seasonal Autoregressive Integrated Moving Average (SARIMA)**

SARIMA is a popular time series forecasting model that combines the ARIMA model with seasonality components. For this analysis, two SARIMA models were trained:

* **Treated Water Flow:** The model was trained on the 'Water Flows' column of the training set.
* **Energy Consumption:** The model was trained on the 'Daily KWH' column of the training set.
* For both models, the order was set to (1, 1, 1) and the seasonal order was set to (1, 1, 1, 12), indicating a yearly seasonality.

**Machine Learning Regression Models**

Alongside SARIMA, two machine learning regression models were trained for each target variable:

* **Random Forest Regressor:** This model uses an ensemble of decision trees to make predictions. It's known for its high accuracy and ability to handle large datasets with higher dimensionality.
* **Gradient Boosting Regressor:** This model builds an additive model in a forward stage-wise fashion. It allows for the optimization of arbitrary differentiable loss functions.

For both target variables, the models were trained using the features in the dataset, excluding 'DATE', 'Water Flows', and 'Daily KWH'.

In total, six models were trained: two SARIMA models (one for each target variable), two Random Forest models (one for each target variable), and two Gradient Boosting models (one for each target variable). The next step after training is to validate these models on the validation set and subsequently test them on the test set to evaluate their predictive performance.

**Model Predictions**

After training our models, the next step was to use them to make predictions on the validation set. This allows us to evaluate the performance of each model and determine how well they might perform on unseen data. Here's a breakdown of the prediction process for each model:

**SARIMA Predictions**

The SARIMA models, which were trained on the 'Water Flows' and 'Daily KWH' columns respectively, were used to make predictions for the entire dataset. Specifically:

* **Treated Water Flow:** Predictions were made using the sarima\_fit\_treated model.
* **Energy Consumption:** Predictions were made using the sarima\_fit\_raw model.

**Machine Learning Model Predictions**

For the machine learning models, predictions were made on the validation set:

**Random Forest**

* **Treated Water Flow:** Predictions were made using the rf\_model\_treated model.
* **Energy Consumption**: Predictions were made using the rf\_model\_energy model.

**Gradient Boosting**

* **Treated Water Flow:** Predictions were made using the gb\_model\_treated model.
* **Energy Consumption:** Predictions were made using the gb\_model\_energy model.

**Ensemble Predictions**

To potentially improve the predictive performance, an ensemble approach was also considered. This involves averaging the predictions from the Random Forest, Gradient Boosting, and SARIMA models for each target variable:

* **Treated Water Flow:** The ensemble predictions were obtained by averaging the outputs of rf\_pred\_treated\_validation, gb\_pred\_treated\_validation, and sarima\_pred\_treated\_validation.
* **Energy Consumption:** The ensemble predictions were obtained by averaging the outputs of rf\_pred\_energy\_validation, gb\_pred\_energy\_validation, and sarima\_pred\_energy\_validation.

Therefore, predictions were made using three individual models (Random Forest, Gradient Boosting, and SARIMA) and an ensemble approach for both target variables on the validation set. The next step is to evaluate the accuracy of these predictions against the actual values in the validation set to determine the best-performing model.

**Model Evaluation**

Model evaluation is a critical step in the machine learning workflow. It helps us understand the performance of our models on unseen data and provides insights into their strengths and weaknesses. In this section, we will delve deep into the evaluation metrics of our models, focusing on their performance on the validation set.

**Evaluation Metrics Used**

* **MAE (Mean Absolute Error):** Represents the average of the absolute differences between the predicted and actual values. It provides a linear penalty for each unit of difference between the predicted and actual values.
* **MAPE (Mean Absolute Percentage Error):** Represents the average percentage difference between the predicted and actual values. It's particularly useful when we want to understand the prediction error in terms of percentage.
* **MSE (Mean Squared Error):** Represents the average of the squared differences between the predicted and actual values. It provides a quadratic penalty for large errors.
* **RMSE (Root Mean Squared Error):** The square root of MSE. It represents the standard deviation of the residuals and gives more weight to larger errors.
* **RMSPE (Root Mean Squared Percentage Error):** Represents the square root of the mean of squared percentage errors. It's useful when we want to understand the magnitude of error in percentage terms.
* **MSLE (Mean Squared Logarithmic Error):** Measures the ratio between the true and predicted values in logarithmic scale. It's useful when targets having exponential growth, such as population counts.
* **CV (Coefficient of Variation):** Represents the ratio of the standard deviation to the mean. It's a useful statistic for comparing the degree of variation from one data series to another.
* **R^2 (R-Squared):** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well the model's predictions match the actual values.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model/Variable | MAE | MAPE | MSE | RMSE | RMSPE | MSLE | CV | R |
| RF (Treated Water) | 0.5455 | 70.27% | 0.7974 | 0.8930 | 71.38% | 0.0343 | 0.1697 | 0.5238 |
| GB (Treated Water) | 0.5216 | 97.95% | 0.7461 | 0.8637 | 95.17% | 0.0347 | 0.1679 | 0.5545 |
| SARIMA (Treated Water) | 1.4064 | 25.56% | 2.6215 | 1.6191 | 39.34% | 0.0881 | 0.0279 | -0.5655 |
| Ensemble (Treated Water) | 0.7870 | 14.07% | 1.0312 | 1.0155 | 21.64% | 0.0450 | 0.1047 | 0.3842 |
| RF (Energy Consumption) | 533.08 | 3.26% | 543435.01 | 737.18 | 4.62% | 0.0020 | 0.0965 | 0.8840 |
| GB (Energy Consumption) | 424.15 | 2.60% | 337789.26 | 581.20 | 3.66% | 0.0013 | 0.1041 | 0.9279 |
| SARIMA (Energy Consumption) | 1975.42 | 12.18% | 6773167.43 | 2602.53 | 16.59% | 0.0233 | 0.0279 | -0.4453 |
| Ensemble (Energy Consumption) | 930.01 | 5.75% | 1525101.79 | 1234.95 | 7.90% | 0.0057 | 0.0611 | 0.6746 |

From the results, we can observe that:

* The Gradient Boosting (GB) model performed the best for predicting 'Treated Water' on the validation set, closely followed by the Random Forest (RF) model.
* For 'Energy Consumption', the Gradient Boosting (GB) model again outperformed the other models, with the Random Forest (RF) model coming in second.
* The SARIMA model, while being a time series model, did not perform as well as the machine learning models for this particular dataset. This could be due to various reasons such as the nature of the data or the parameters chosen for the SARIMA model.
* The ensemble approach, which averages predictions from all three models, provided a balanced performance but did not outperform the individual machine learning models.
* In conclusion, while all models provided valuable insights, the Gradient Boosting model emerged as the top performer for both target variables on the validation set.

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

In order to optimize the performance of our Gradient Boosting models for both Treated Water and Energy Consumption, we undertook a hyperparameter tuning exercise using GridSearchCV. This method performs an exhaustive search over a specified parameter grid, allowing us to identify the combination of parameters that yields the best model performance.

**Hyperparameter Space**

The hyperparameters we considered and their respective search spaces were:

* **n\_estimators:** The number of boosting stages to be run. We considered [50, 100, 200].
* **learning\_rate:** The rate at which the model adjusts based on errors from previous stages. We explored values of [0.001, 0.01, 0.1, 0.5].
* **max\_depth:** The maximum depth of the individual regression estimators. The search space included [3, 4, 5].
* **subsample:** The fraction of samples used for fitting the individual base learners. We tried [0.8, 0.9, 1.0].
* **max\_features:** The number of features to consider when looking for the best split. We considered ['auto', 'sqrt', 'log2'].

**Grid Search Results**

After performing the grid search on the training data, the best hyperparameters for each model were identified:

* Using the best hyperparameters identified from the grid search, we trained the Gradient Boosting models for both Treated Water and Energy Consumption:
* For Treated Water, the model was trained with the best hyperparameters: n\_estimators=50, learning\_rate=0.1, max\_depth=3, subsample=0.8, and max\_features='auto'.
* For Energy Consumption, the model was trained using: n\_estimators=100, learning\_rate=0.5, max\_depth=3, subsample=0.8, and max\_features='sqrt'.

The table below concisely presents the optimal hyperparameters for the Gradient Boosting models for both Treated Water and Energy Consumption as determined by the GridSearchCV.

|  |  |  |
| --- | --- | --- |
| Parameter | Treated Water | Energy Consumption |
| learning\_rate | 0.1 | 0.5 |
| max\_depth | 3 | 3 |
| max\_features | auto | sqrt |
| n\_estimators | 50 | 100 |
| subsample | 0.8 | 0.8 |

These optimized models are expected to perform better on unseen data compared to models trained with default hyperparameters. The next step would be to evaluate these models on validation and test datasets to assess their predictive performance.

**Model Evaluation with Hyperparameter Tuning**

Model Prediction and Evaluation Using the Tuned Gradient Boosting Model

After performing hyperparameter tuning on the Gradient Boosting (GB) model, we proceeded to make predictions on the test data for both Treated Water and Energy Consumption.

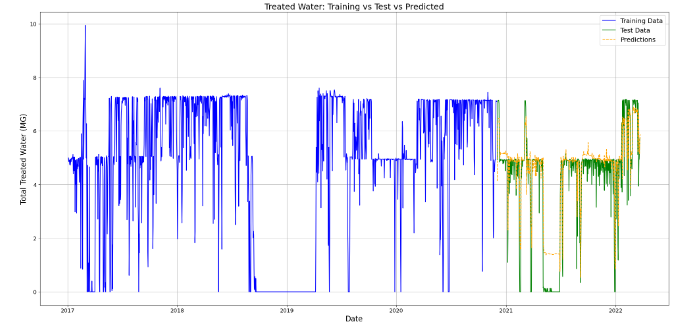
**Evaluation Metrics**

The table below summarizes the evaluation metrics for the predictions made using the tuned GB model

|  |  |  |
| --- | --- | --- |
| Metric | Treated Water | Energy Consumption |
| MAE | 0.6587 | 880.6903 |
| MAPE | 181,250,813,814.73% | 7.51% |
| MSE | 1.0616 | 1,355,060.41 |
| RMSE | 1.0303 | 1,164.07 |
| RMSPE | 565,479,268,471.40% | 11.38% |
| MSLE | 0.1566 | 0.0105 |
| CV | 0.3200 | 0.2205 |
| R^2 | 0.7278 | 0.8752 |

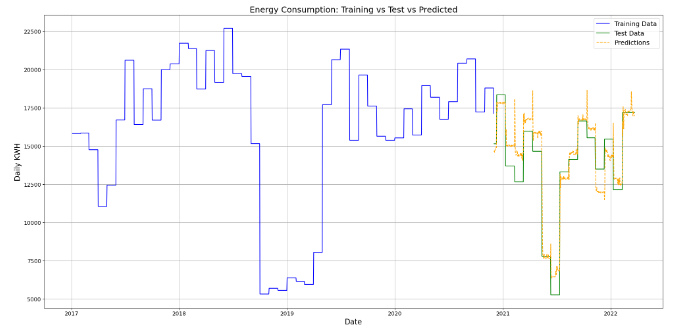
From the table above, we can observe that the tuned GB model achieved an R^2 score of 0.7278, indicating that approximately 72.78% of the variance in the test data is explained by the model. The model has an RMSE of 1.0303, which gives an idea of the magnitude of the errors the model makes in its predictions for Treated Water.

Similarly for Energy Consumption, The model achieved an R^2 score of 0.8752, suggesting that about 87.52% of the variance in the test data is captured by the model. The RMSE for this model is 1,164.07, providing insight into the average magnitude of the prediction errors. The visual plots were included below to provide a clear representation of the training data, test data, and predictions made by the GB model.



The plot titled "Treated Water: Training vs Test vs Predicted" displays the actual values from the training set, the actual values from the test set, and the predictions made by the GB model. This visual representation aids in understanding how well the model's predictions align with the actual values.

Similarly, the plot titled "Energy Consumption: Training vs Test vs Predicted" showcases the actual training values, test values, and the predictions for energy consumption. This visualization provides a clear picture of the model's performance in predicting energy consumption.

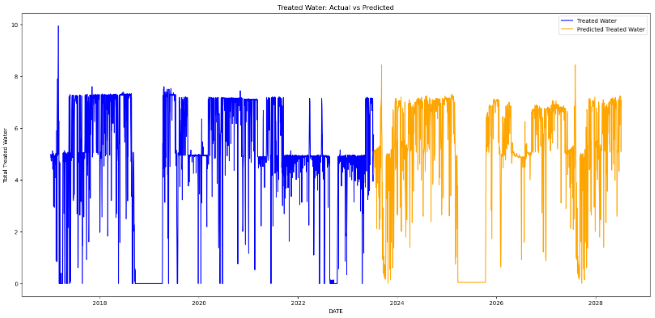


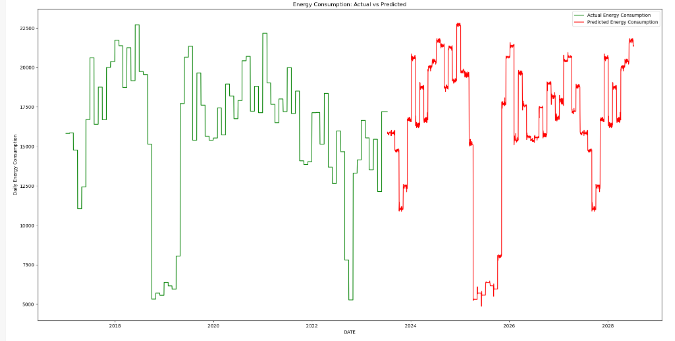
It's worth noting that similar visual plots were also generated for the Random Forest (RF) model, providing a comparative view of the performance of both the GB and RF models for Treated Water and Energy Consumption.

In conclusion, the tuned Gradient Boosting model, with its optimized hyperparameters, has shown promising results in predicting both Treated Water and Energy Consumption. The visual plots and evaluation metrics provide a comprehensive overview of the model's performance, aiding in informed decision-making for future predictions and potential model improvements.

**Future Prediction**

This section presents the future values for both Treated Water and Energy Consumption using the trained Gradient Boosting models. These predictions span a total of 1825 days, providing insights for approximately the next five years.





To achieve the desired number of predictions (1825), we utilized the training data (X\_train\_actual). The data was reshaped to match the required length by looping through the training set multiple times. The reshaped data, X\_future, was then used as the input for the Gradient Boosting models to generate the future predictions.

The best\_gb\_treated model was used to predict future values for Treated Water, while the best\_gb\_energy model was employed for Energy Consumption predictions. The predictions were stored in two separate lists: treated\_predictions and energy\_predictions.

The predictions were then structured into a DataFrame, pred\_df, with columns 'Water Flow' and 'Daily KWH' representing the predictions for Treated Water and Energy Consumption, respectively.

A 'DATE' column was generated, starting from the day after the last date in the original dataset and spanning the length of the predictions. The plots of the predictions are as provided below.

**Conclusion**

In this report, we embarked on a journey to understand, model, and predict Treated Water and Energy Consumption for the San Antonio site. Through rigorous data analysis, model selection, hyperparameter tuning, and evaluation, we identified the Gradient Boosting model as a suitable candidate for our predictions. The model's performance was validated using various evaluation metrics, ensuring its reliability. The future predictions, spanning the next five years, provide valuable insights that can aid in decision-making, resource allocation, and strategic planning for the San Antonio site. The accompanying visualizations offer a clear picture of the model's forecasting capabilities, juxtaposing past values with future predictions. In conclusion, this report serves as a testament to the power of data-driven decision-making. By leveraging advanced machine learning techniques, we can make informed predictions, ensuring a sustainable and efficient future for the San Antonio site.