Forecasting the Demand of Hot Water

Introduction

Supplying hot water at lower cost, with less energy, and lighter load on the network infrastructure is always a primary goal of utility stakeholders. Short-term hot water demand forecasts can be taken to tackle the imbalance and gaps between supply and demand. It can be used to better schedule operations, maintenance and to provide better services.

Electric water heaters or geysers account for almost half of household energy, especially during peak hours. The objective of this project is to forecast the volume of hot water drawn from a geyser using the sensor measurements from pre-installed geysers as the underlying data. Creating a robust model capable of accurately forecasting the demand for hot water can be used to save electricity by drawing power during off-peak hours while still ensuring hot water is available when needed

This report provides utility companies with valuable insights and supportive analysis for hot water demand forecasting. Utility companies will be able to increase reliability and affordability of their services, while better managing energy loads and reducing electricity costs to further improve their margins.

Dataset Description

The data is collected over a period of 4 months from October 2020 to January 2021 inclusive. The initial dataset includes 34,842 readings across 8 features. Each row in the dataset corresponds to a reading of measured data over a time interval of 5 minutes, and each measurement is explained in detail in the table shown in [Appendix 1a]. This data was obtained from an energy tech startup company based in South Africa and New York.

Data Preprocessing & Feature Engineering

There is no missing data or null values in this dataset. The goal is to process and aggregate in such a way that each row corresponds to a “water draw event”, which is defined as all consecutive measurements where there is a flow of water. Only the consecutive measurements where the Water Flow feature is true were extracted and consecutive blocks were successfully grouped together so that each individual water draw event is a row. This resulted in 1,844 rows, with each row corresponding to a “water draw event.”

Following this, relevant features were generated using aggregation functions and feature engineering. Seeing that consecutive individual 5 minute measurements where water is being drawn are lumped together. The maximum, minimum, average and range of the temperature and energy data was used to generate features. In addition, the hour of day, average temperature change, average rate of change in temperature, average power and elapsed time were some of the other features that were created. Elapsed time is the total time elapsed for each water draw event. Also, for each event, the target variable is the Cumulative Volume of Water Drawn. The names of the original features were changed to make them more interpretable.

The aforementioned feature generation step resulted in a total of 30 features. In order to find the most relevant features, a correlation matrix was plotted and highly correlated independent variables were dropped. Based on the correlation matrix shown in [Appendix 2a], a function was written to return a set of correlated features such that these features have a correlation greater than the specified threshold with at least one other independent feature. This threshold was set to 0.85 and the highly correlated independent variables were dropped, resulting in a total of 20 features and the dataset is now ready for modeling.

Exploratory Data Analysis and Inference

First of all, in order to have a general acknowledgement of the volume of water drawn in different time periods, we got a plot as shown in [Appendix 3a]. From the plot, it can be noted that for the most part, the variation over the 4 months seems to be quite consistent, with abrupt surges that should be further explored. Seeing that South Africa gets warmer between the months of October and February, one might expect the demand for hot water to drop but this is not explicitly evident from this plot.

The target variable is heavily positively skewed (skew = 1.45) as indicated by the distribution plot. When a log transform was taken, it appears to follow a bimodal distribution and is not convincingly transformed into a normal distribution. The comparisons are shown in [Appendix 3b]. Box-Cox transformation and square root transformation were also explored, but the best fit seemed to be the log transform.

A series of plots to explore the relationships between the numerical features and the target variable. The plots can be found in [Appendix 3c]. Through these series of plots, some importance inferences are included below:

- There seems to be an extreme outlier, where the elapsed time of the particular event is more than 8 hours. To provide some context, the rest of the events are less than 2 hours. This data point will be dropped although it is not clear if this data point is in fact erroneous. However, dropping this outlier will improve performance of regression based models, which are not robust to outliers.
- Cumulative energy consumed and Power are sparse features. Typically, learning algorithms do not learn well from sparse data. However, ensemble tree-based methods such as XGBoost and Random Forest do so it is worth exploring.
- The temperature-based features are rather skewed so it makes sense to scale the feature matrix.
- Hour of day, Day of event and Month of event are all discrete features. Perhaps we can one-hot encode these discrete features at the expense of increasing model complexity. Feature selection techniques can be used to handle model complexity. However, it is clear from the correlation matrix that Day of event and Month of event have almost zero correlation with the target variable. It makes sense to simply drop these two features.
- As expected, there tends to be a greater demand for hot water in the morning, which drops later in the day, and picks up in the late evening.

Finally, the demand for hot water over the course of the day was further explored[Appendix 3d]. As expected, there is a great demand for hot water in the morning, which gradually decreases and picks up moderately in the late evenings. Lowest demand seems to be from 12am - 2am. One thing to note here is

that there is no sensor data between the hours of 3am and 5am, which is worth exploring. Is there absolutely no hot water being drawn during these hours (unlikely) or are there no sensor measurements during this time? Ideally, sensor measurements should be recorded always but the cost of collecting this data may be rather expensive, which may be why the sensor is turned off during the hours where the expected demand for hot water is relatively lower.

Model Selection

In the following parts, the supervised learning framework was incorporated to predict the cumulative volume of water drawn in each water draw event, which is defined as an event where there is continuous flow of water / all consecutive measurements where Water flow is True.

1. Splitting the Data and Feature Scaling

The first step is to split the whole dataset into training and testing dataset. Since the data is indexed in time order (time-series data), splitting the data at random would lead to data leakage. Therefore, it is wise to incorporate structured splitting after the entire dataset is sorted in increasing order by time to ensure that only data from the past is being used to forecast the future. There is a temporal dependency between observations, and this relation must be preserved during testing. Also, seeing that a separate test dataset is provided, this entire current dataset can be considered as the development dataset. Therefore, this dataset should only be split into a training dataset and a validation dataset for the purposes of hyperparameter tuning and model selection. 80% of the data will be used for training purposes and the remaining 20% of the data will be used as the validation dataset. In addition, for the purposes of cross validation, time-series split were used to prevent data leakage. A standard scaler was used to scale the features so that they have zero mean and unit variance.

A baseline model, where the mean of the target variable in the training set was used as the prediction for all instances, was used as the baseline error. The baseline model's mean squared error on the validation dataset was 27153 and the adjusted R-squared score is 0. This is a common baseline used in regression problems. The other trained models' MSEs should be much lower than these baseline mean squared error. A total of 6 different learning algorithms were considered. Three different regression models were considered (Linear, Lasso and Ridge) and three different tree-based algorithms (Decision Tree, Random Forest and XGBoost).

2. Regression Models (Linear, Ridge and Lasso Regression)

All selected independent variables were used as the features for the linear regression model. It should be noted that the mean squared error on the training dataset was 3821.77, and the mean squared error on the validation dataset was 3544.83. The adjusted R-squared score on the validation dataset was 0.868, indicating that model performance is relatively good. The feature importances of the linear regression model using the coefficients of the features were visualized in [Appendix 4a].The relatively low

difference between the training and validation error indicates that the model is not overfitting. Further reducing the bias at the cost of increasing variance so that the model learns better from the training dataset is a viable tactic. However, regularized variants of linear regression, such as lasso and ridge regression were first explored, to empirically observe their corresponding evaluation metrics on the validation dataset.

Firstly, the default lasso regression model was trained and evaluated. The resulting mean squared error on the training dataset was 3866.07 and 3514.98 on the validation dataset. The new adjusted R-squared score is 0.869, which did not improve much compared to the basic linear regression model. Next, the L1 regularization hyperparameter, alpha was hyperparameter tuned. Typically, the higher the alpha, the fewer the number of non-zero coefficients. Instead of regular k-fold cross validation, time series splits are used to ensure that there is no data leakage (predicting the past using future data). Two criteria are met here: every validation set contains unique observations and observations from the training set occur before their corresponding validation set. After a series of testing, it was observed that the optimal alpha value was 0.75.

To decide the most important features, the features with the highest absolute coefficients were chosen. This is especially useful as a variance reduction technique, as choosing fewer features reduces model complexity. The top 5 features from the final lasso regression model in terms of highest absolute coefficients are: Range of Internal Temperature, Range of Outlet Temperature, Range of Ambient Temperature, Range of Inlet Temperature and Minimum Internal Temperature.

Next, a ridge regression model was trained. For the ridge regression model, it was noted that the optimal alpha value was 20 after hyperparameter tuning. Under this situation, the optimal mean squared error on the validation set was 3531.47, and the adjusted R-squared score was 0.869.

3. Tree-based Models (Decision Tree, Random Forest and XGBoost)

Three different tree-based models were trained and evaluated. The basic decision tree model led to a mean squared error of 6545.596 on the validation dataset, and an adjusted R-squared score of 0.756. The feature importances are shown on [Appendix 5a]. However, a decision tree is more interpretable than ensemble methods such as a random forest model, so there is value in training a decision tree to plot the feature importances.

Next a vanilla Random Forest Regressor model was trained and evaluated on the validation dataset. The adjusted R-squared score was 0.896, which was the best yet. The mean squared error on the training set was 472.72, and 2804.23 on the validation set. The significant difference between the training error and validation error indicates that this model was overfitting to the training data. Hyperparameter tuning can be performed to improve validation performance using variance reduction techniques specific to tree-based algorithms such as controlling the max depth of each individual tree as well as controlling the maximum number of features to consider at each split. The vanilla random forest model's feature importances are shown in [Appendix 5b].

After hyperparameter tuning the random forest model over three different hyperparameters, the optimal random forest model uses “squared_error” as the splitting criterion, a max depth of 15 and a max features to consider at each split of 0.5 or 50%. This hyperparameter tuned random forest model’s adjusted R-square was 0.909 on the validation set. The percentage decrease in validation mean squared error after hyperparameter tuning was 12.84% . The feature importances are visualized in [Appendix 5c], and the top 3 features are Range of Internal Temperature, Elapsed time and Range of outlet temperature.

Reducing the number of features using ensemble feature selection as a variation reduction technique was explored. On the one hand, we included only the top 5 features selected from the lasso regression model, and on the other hand, we included only the top 5 features from the random forest model. For the model including features selected from the lasso regression model, the best validation mean squared error was 2719.72, and the adjusted R-squared was 0.899. For the other one, the best validation mean squared error is 2893.89 and the adjusted R-squared was 0.893. The hyperparameter tuned random forest model including all features remains the best in terms of performance.

The final model that was trained and evaluated was the XGBoost regressor model. A vanilla XGBoost model was trained, after which hyperparameter tuning was done to find the best hyperparameter configuration. After hyperparameter tuning the boosted tree model, the mean squared error of the best XGBoost regressor was 2677.83 with 200 estimators, learning rate of 1.5, alpha value of 0.5 and lambda value of 1.5. The adjusted R-squared score on the validation dataset was 0.9, and the percentage decrease in validation mean squared error after hyperparameter tuning was 17.25%. The feature importances are visualized in [Appendix 5d].

Model Evaluation, Optimal Model Training and Model Deployment

All 6 models were evaluated on the validation dataset using the mean squared error and the adjusted R-squared score as the two evaluation metrics. The final results table is shown in [Appendix 6a]. The optimal model is the random forest tree model with all features included. The forecast on the validation dataset using our optimal model was visualized by plotting the forecast of water volume drawn and the actual water volume drawn on the same plot. The plot is shown in [Appendix 6b], and it can be observed that the model does a good job in forecasting hot water demand.

Finally, the Random Forest model with the optimal hyperparameters found during cross validation was trained on the entire dataset (training + validation) and was used to forecast the water volume drawn using the provided test set. A preprocessing pipeline was built to preprocess the test dataset. Note, that the energy tech startup did not provide the true forecasts for the test set so evaluation on the test set was not possible, but the hot water demand on the test set was successfully forecasted.

Conclusion, Implications and Future Work

The main objective of this study was to investigate the influence of using supervised machine learning techniques to accurately forecast the hot water demand so that utility companies can use the proposed model’s forecast to improve reliability and affordability of their services while managing energy loads better and reducing electricity costs.

The results of this study have highlighted the following concluding remarks:

1. Tree-based models have shown better forecasting performance, especially after hyperparameter tuning, than the regression models. The final optimal model achieved an adjusted R-squared score of 0.909 on the validation set.
2. The demand for hot water is highest during the morning time, and peaks at 8am. Then there is a decreasing trend until 4pm in the afternoon, following which the demand picks up moderately in the late evening.
3. The range of the internal temperature within a water draw event is the most important feature in predicting hot water demand, with the second most important feature being elapsed time

Below are some suggestions for our potential stakeholders, the directors of utility companies:

1. Rate Structure: Utility companies can take utility time into their rate structure, namely, establishing rates that escalate during peak hours. For the peak hours, a higher rate can be set while in the non-peak hours, a relatively low rate can be set. On the one hand, this can help to release the pressure in peak hours; on the other hand, this is a good choice to improve profits.
2. Maintenance & Improvement of Equipment: During the peak hours, the intensive usage of electricity and hot water heaters puts great pressure on related equipment, such as pipes. The regular maintenance of current equipment is of great importance. When it comes to maintenance, utility companies would want to avoid doing maintenance work during peak hours. Considering both the normal working hours of maintenance crews and peak usage time periods, the timeframe from 3-5 PM would be the best option for periodic check and maintenance.
3. Off-peak drawing of electricity: Since drawing power off-peak is efficient and less straining on the grid, it leads to more affordable and reliable energy. Using the forecasts from the proposed model, utility companies can identify peak hours (7 AM-11 AM), and draw energy during off-peak hours for heating purposes while still ensuring hot water is available for consumers when needed. It is anticipated that the retention of consumers will increase due to the increased reliability and affordability of their services, while also easing the strain on the distribution grid.

The model is not perfect, as there are still some limitations. In order to improve model performance and improve robustness, the following should be considered:

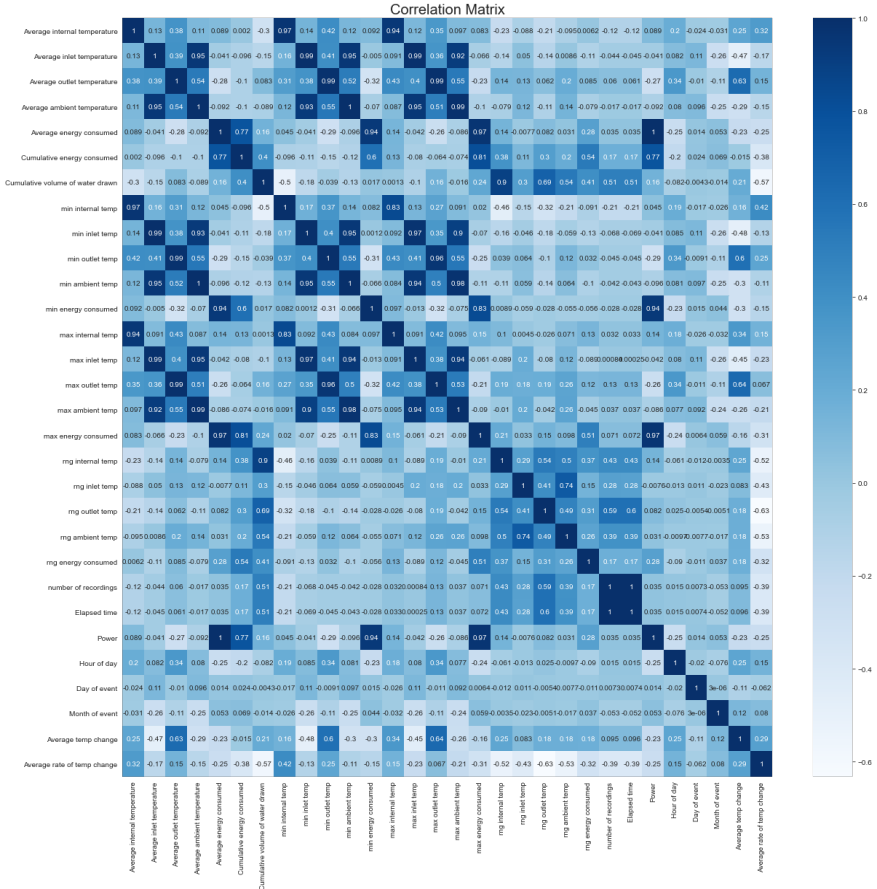
1. Time Span: This dataset covers only a short period of term as it only contains 4 months of data. Medium and long-term forecasts should expand well into the future, and can provide periodic trends to assist in the sizing and macro-level system operations.
2. External Data Sources: Currently, only temperature and energy data associated with geysers are considered. Factors such as weather changes and population changes should also be considered, as it can affect demand. Perhaps satellite imagery data of cloud cover could be bought to predict weather changes, and its relationship with hot water demand.
3. Algorithms & Techniques: Classical econometric techniques, such as ARIMA and SARIMAX, and advanced time-series forecasting algorithms such as Facebook Prophet and LSTMs may be worth exploring

Appendix

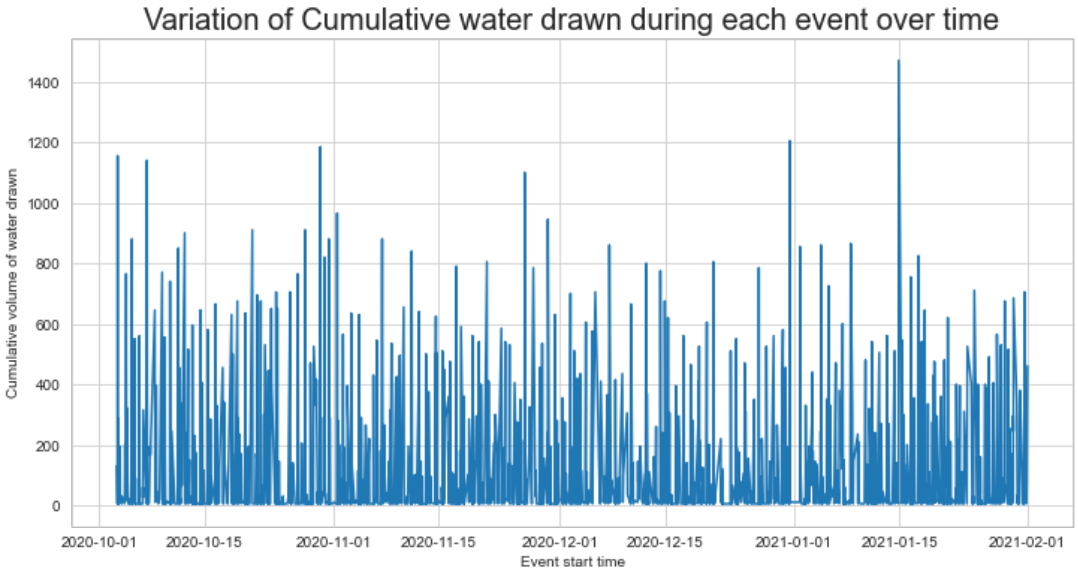
1a) Dataset Description Table

Measurement	Description	Measurement period	Unit of Measurement
Internal temperature	Temperature of the water inside the geyser	Instantaneous at the end of the measurement period	Degrees Celsius (C)
Inlet temperature	Temperature of water coming into geyser	Instantaneous at the end of the measurement period	Degrees Celsius (C)
Outlet temperature	Temperature of water exiting the geyser	Instantaneous at the end of the measurement period	Degrees Celsius (C)
Ambient temperature	Air temperature of the environment	Instantaneous at the end of the measurement period	Degrees Celsius (C)
Volume of water drawn	The volume of water drawn from the geyser	Cumulative over the measurement period	Liters (L)
Energy consumed	The amount of energy drawn by the geyser's heating element	Cumulative over the measurement period	Watt hours (Wh)
Water Flow	True if water has been drawn from geyser in this measurement period	Cumulative over the measurement period	Boolean Value

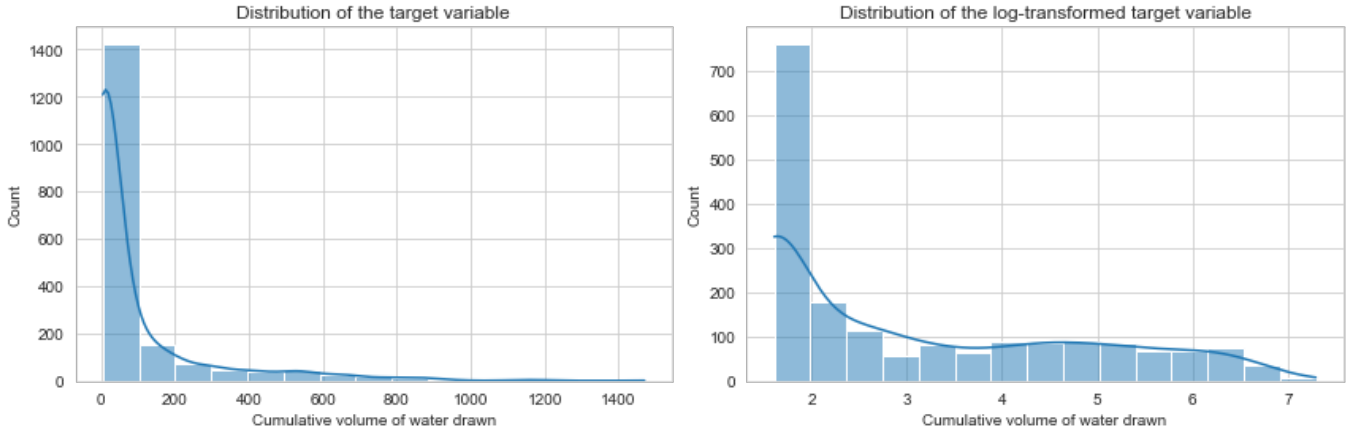
2a) Correlation Matrix Plot



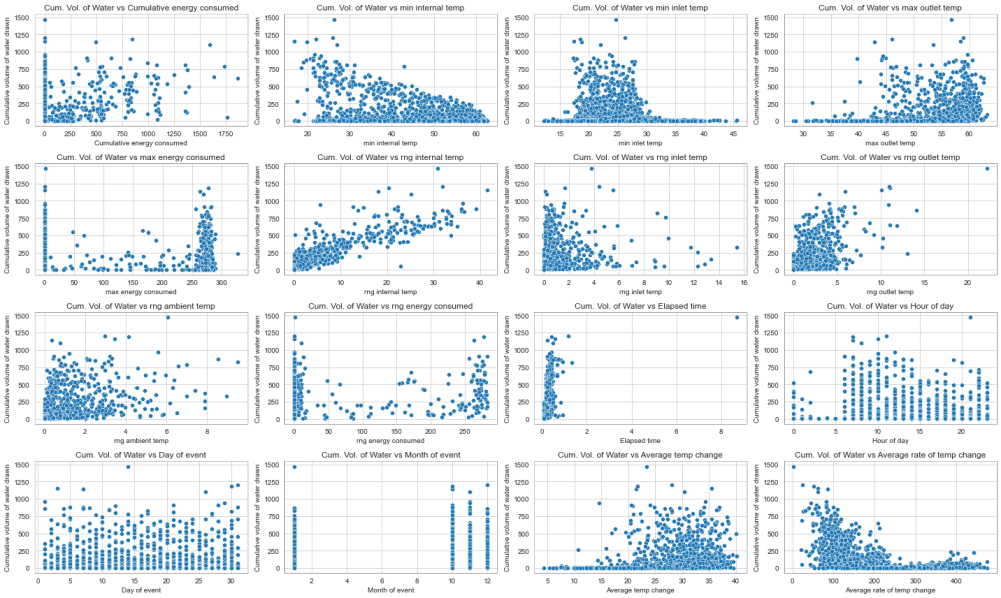
3a) Variation of Cumulative water drawn during each event over time



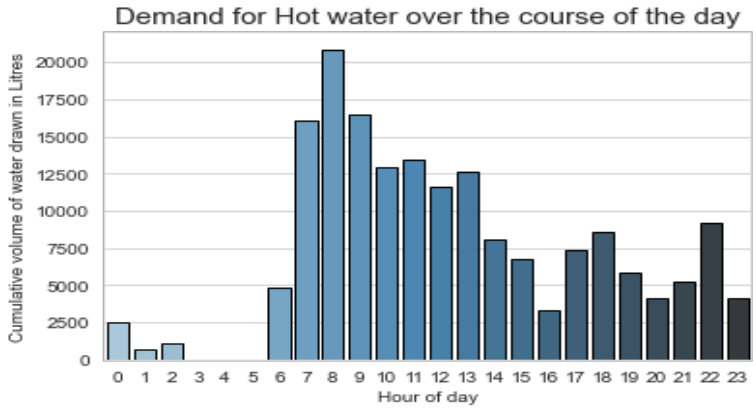
3b) Distribution of Target Variable



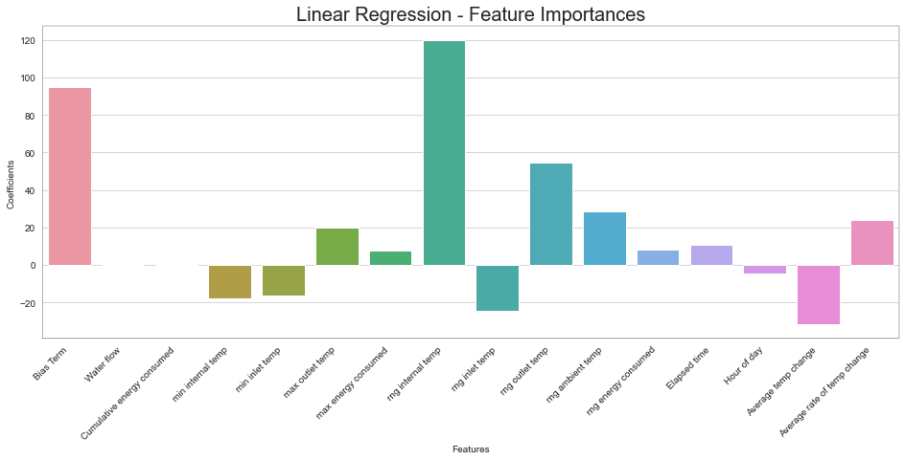
3c) Relationships between the Numerical Features & the Target Variable



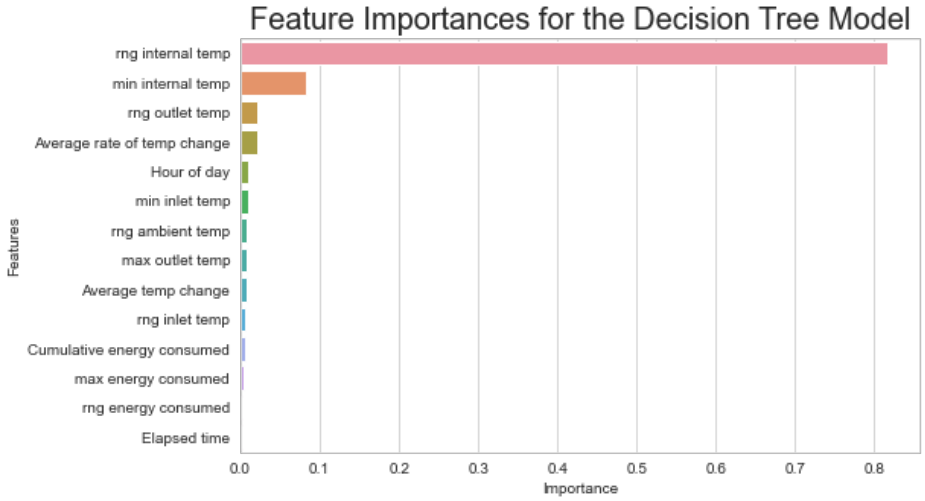
3d) Demand for Hot Water Over the Course of the Day



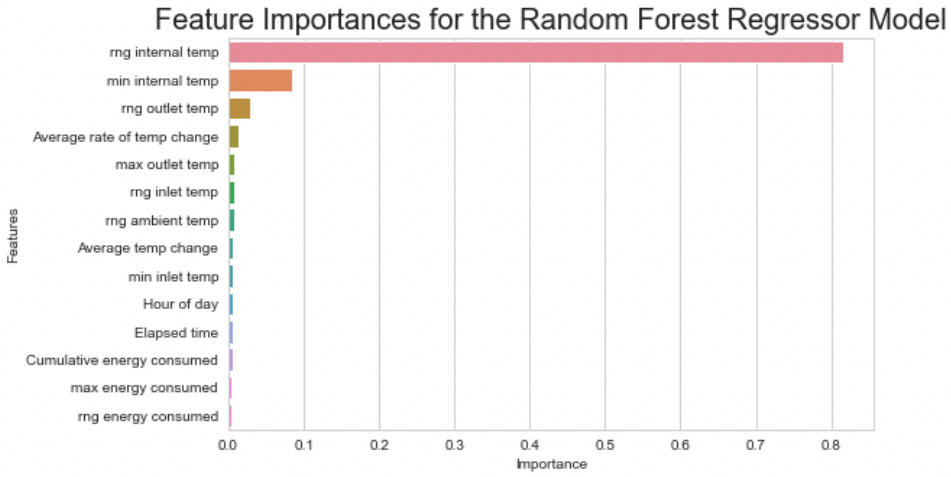
4a) Linear Regression - Feature Importances



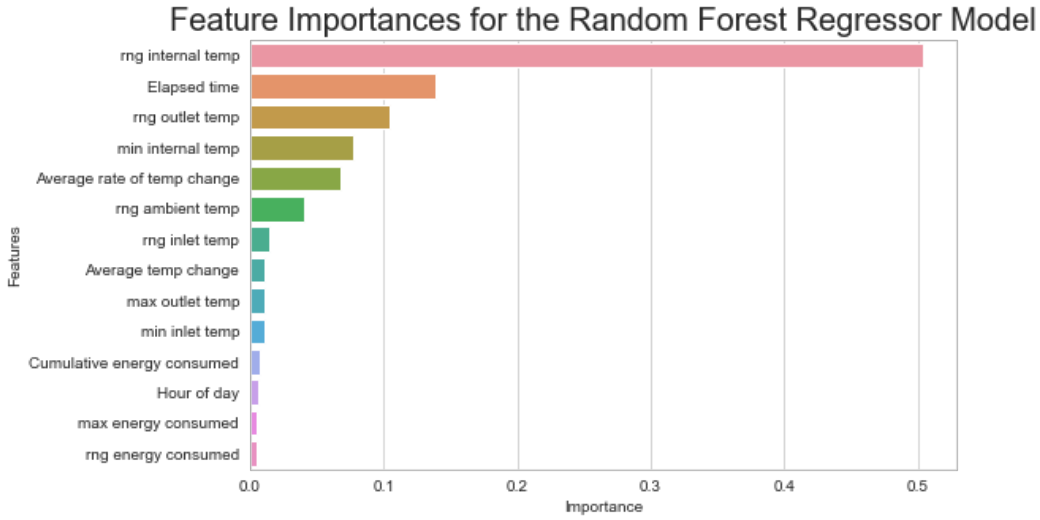
5a) Decision Tree - Feature Importance



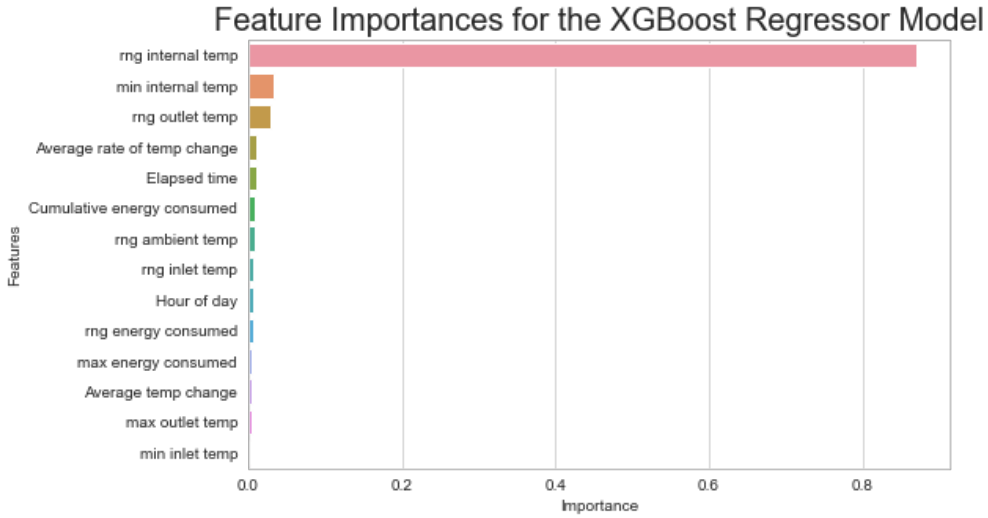
5b) Random Forest - Feature Importance



5c) Random Forest with Hyperparameter Tuning - Feature Importance



5d) XGBoost Regressor Model - Feature Importance



6a) Model Evaluations Result Tables

	Validation Mean Squared Error	Validation Adjusted R^2 Score
Vanilla Random Forest	2444.26	0.91
XGBoostRegressor	2677.83	0.90
Random Forest (Lasso Feature Selection)	2719.72	0.90
Random Forest (RF Feature Selection)	2893.89	0.89
Lasso Regression	3519.56	0.87
Ridge Regression	3531.47	0.87
Linear Regression	3544.83	0.87
Baseline Model	27153.04	0.00

6b) Comparisons between Predictions and Real Water Volume Drawn in Validation Set