Time-Series Forecasting of Hot Water Demand in Residential Complexes

Prepared by: Praveen Sivashangaran
03/24/2022

Introduction

Plentify’s HotBot controls and monitors geysers and records readings of measured data every 5 minutes. The objective of this project is to forecast the volume of hot water drawn from a geyser using the sensor measurements 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 ensuring hot water is available when needed. The data used in this project is over a period of 4 months from October 2020 to January 2021 inclusive. The data was preprocessed and relevant features were generated, followed by exploratory data analysis, model selection and evaluation. Finally, the optimal model was used to forecast the water volume drawn using the test dataset.

Data Preprocessing and Feature Engineering

The initial dataset included 34842 readings across 8 features. There was no missing data in this dataset. This data had to be processed and aggregated in such a way that each row corresponds to a “water draw event”, which was defined as all consecutive measurements where there is a flow of water. Following this, relevant features were generated using aggregation functions and feature engineering. The maximum, minimum, average, range and standard deviation 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. A correlation matrix was plotted and highly correlated independent variables were dropped.

Exploratory Data Analysis and Inference

A time-series plot of the target variable was plotted to identify any trends in hot water demand over the 4 months. 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 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. Relationships between the numerical features and the target variable were plotted using scatter plots to identify any linear trends. A linear relationship between the target variable and the range of the internal temperature was identified. An extreme outlier was identified and dropped. The skewness of the temperature based features indicated that feature scaling may help. The hot water demand over the course of the day was visualized and, as expected, a greater demand for hot water was observed in the morning. This demand gradually decreases and picks up moderately in the late evenings. One thing to note here is that there is no HotBot data between the hours of 3 AM and 5 AM, 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 continuously 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 and Evaluation

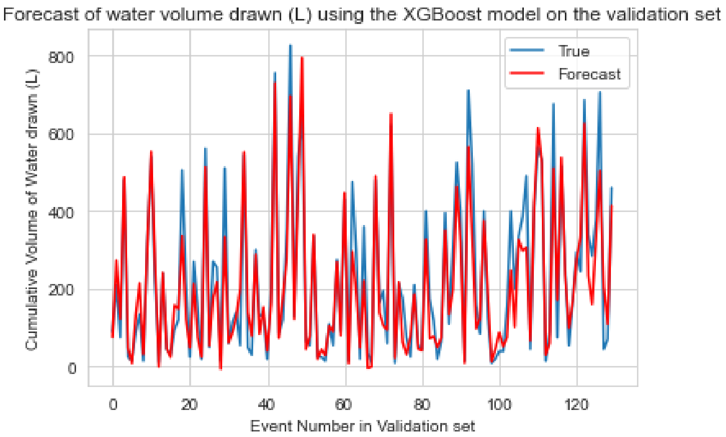
Since we are dealing with data indexed with a time order (time-series data), we cannot split the data at random, as this would lead to data leakage. For example, we may be using data from the future to forecast the past if the data from the future is in our training set, and the past data is in our test set. Therefore, it is wise to incorporate structured splitting after the entire dataset is sorted in increasing order by time to ensure that we are only using data from the past to forecast the future. There is a temporal dependency between observations, and we must preserve that relation during validation. A 4:1 ratio was used to split the data into a training set and a validation set according to the aforementioned stipulations. In addition, for the purposes of cross validation, time-series splits were used to prevent data leakage

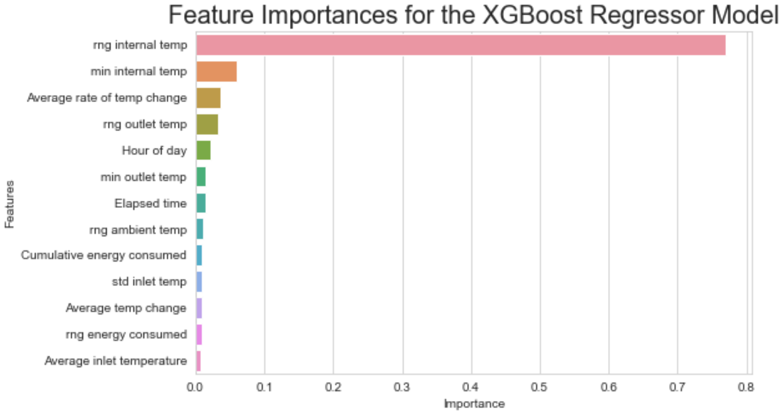
A baseline model, where the mean of the target variable in the training set was used as the prediction for all instances, was used as the baseline error. A total of 6 different learning algorithms were considered. Three different regression models were considered (Linear, Lasso and Ridge) and three different tree-based algorithms (Decision Tree, Random Forest and XGBoost). Following the training of the models on the training data, the hyperparameters were tuned using cross validation. The validation mean squared error (MSE) and the validation adjusted R^2 score were calculated as the metrics of evaluation.. The model with the best performance on the validation set was the XGBoost model, and the importance of the features is shown in the plot below. Finally, the XGBoost model with the optimal hyperparameters found during cross validation was trained on the entire dataset (training + validation) and used to forecast the water volume drawn using the provided test set.

Conclusion, Implications and Future Work

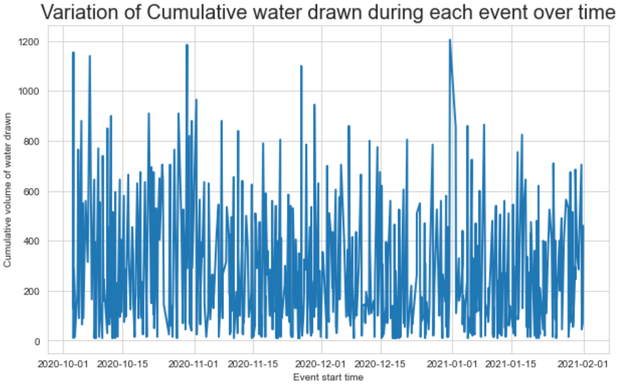
Although the XGBoost model showcased the highest prediction performance, the interpretability of ensemble methods is not as comprehensive as that of regression. This is to do with how feature importance is determined for the two methods. Meanwhile, the regression based models had around a 38% decrease in performance when compared to the XGBoost model but are more interpretable. This tradeoff between performance and interpretability should be further assessed in the context of the problem at hand to determine which of the two is valued more. From the forecast plot below, it can be observed that the model tends to underestimate the water volume. Also, the range of the internal temperature is by far the most significant feature in terms of feature importance in the XGBoost model.

As for future work, it was noted that the XGBoost model was overfitting by observing the training and validation errors. More comprehensive hyperparameter tuning over a larger parameter space with specific focus towards variation reduction hyperparameters including the maximum depth of each individual tree and the minimum number of features to consider for each split will be explored. Furthermore, more features can be created in order to potentially identify features with higher predictive power. However, it is expected that the most significant increase in performance should come from the collection of more data since the final training dataset only had 523 samples. Assuming that the cost of collecting HotBot sensor data is low, collecting data over a longer period of time should lead to a more robustly trained model. In addition, learning and implementing time-series forecasting algorithms such as Facebook Prophet and LSTM may also prove to be worthwhile.

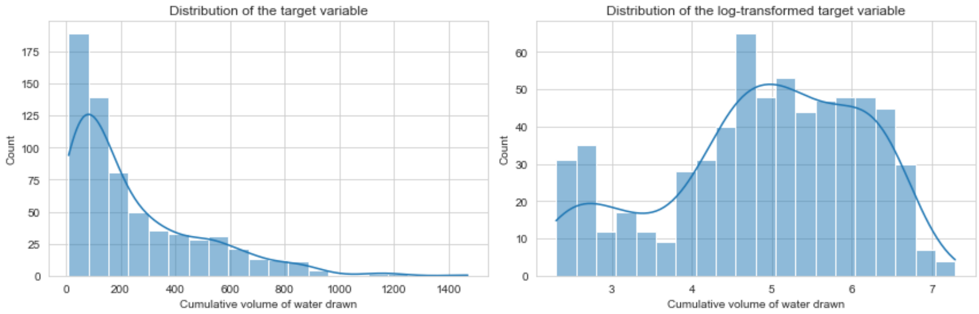




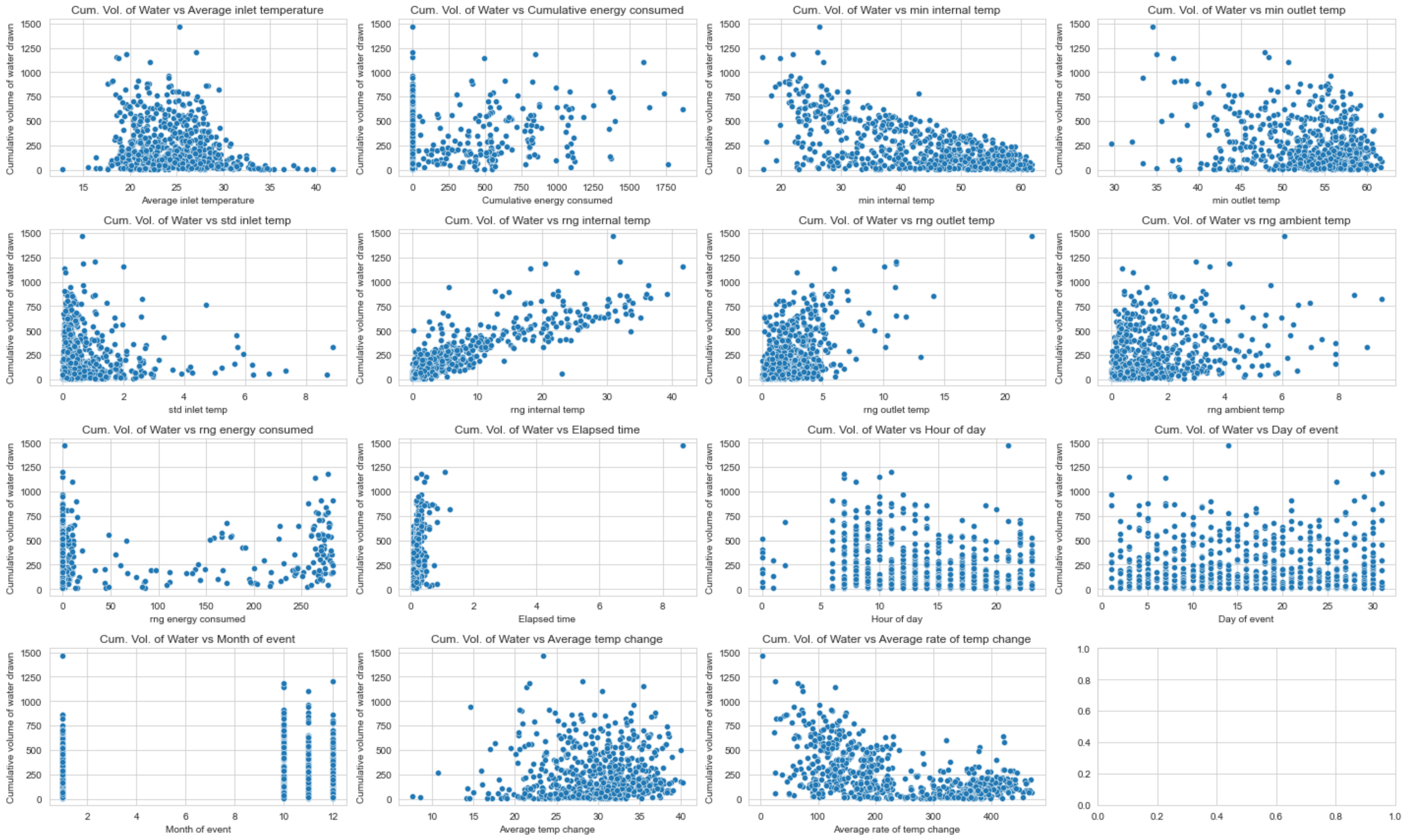
Appendix



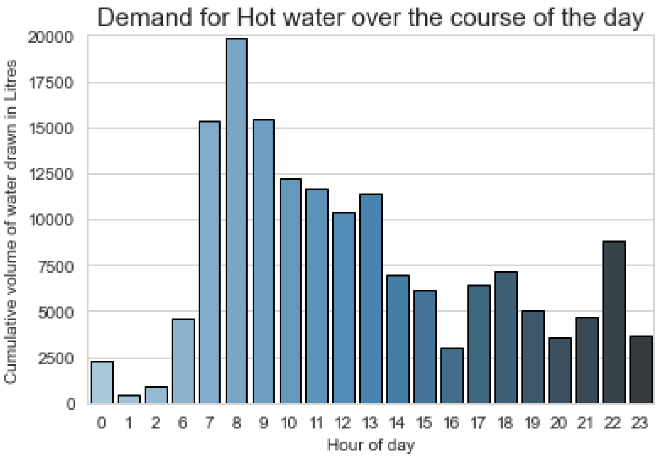
Variation of Cumulative water drawn during each event over time
Cumulative volume of water drawn
0
200
400
600
800
1000
1200
2020-10-01
2020-10-15
2020-11-01
2020-11-15
2020-12-01
2020-12-15
2021-01-01
2021-01-15
2021-02-01
Event start time



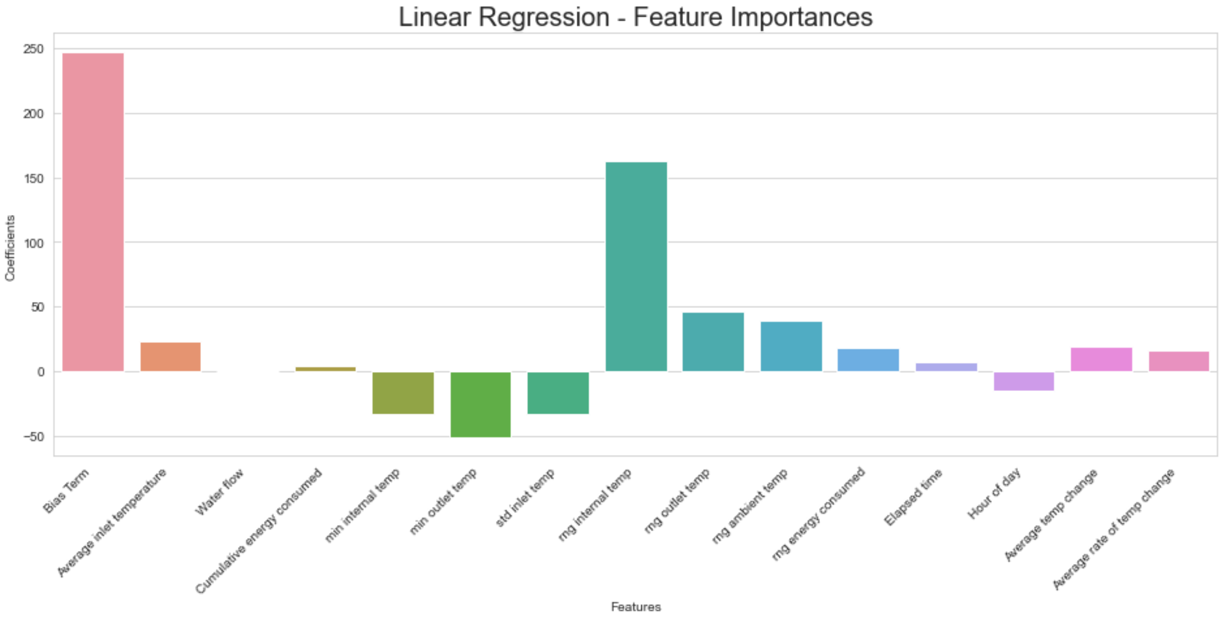
Distribution of the target variable
Count
0
25
50
75
100
125
150
175
0
200
400
600
800
1000
1200
1400
Cumulative volume of water drawn
Distribution of the log-transformed target variable
Count
0
10
20
30
40
50
60
3
4
5
6
7
Cumulative volume of water drawn

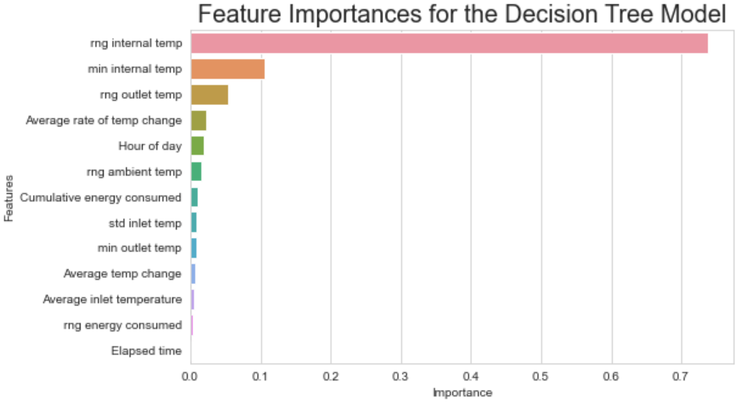


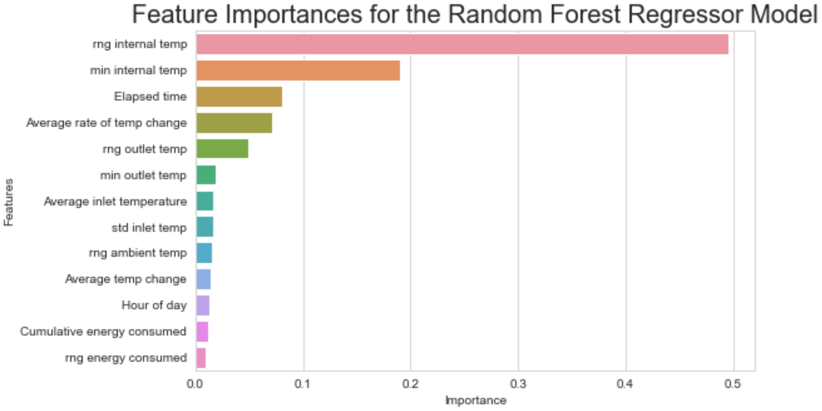
Cum. Vol. of Water vs Average inlet temperature
Average inlet temperature
Cum. Vol. of Water vs Cumulative energy consumed
Cumulative energy consumed
Cum. Vol. of Water vs min internal temp
min internal temp
Cum. Vol. of Water vs min outlet temp
min outlet temp
Cum. Vol. of Water vs std inlet temp
std inlet temp
Cum. Vol. of Water vs rng internal temp
rng internal temp
Cum. Vol. of Water vs rng outlet temp
rng outlet temp
Cum. Vol. of Water vs rng ambient temp
rng ambient temp
Cum. Vol. of Water vs rng energy consumed
rng energy consumed
Cum. Vol. of Water vs Elapsed time
Elapsed time
Cum. Vol. of Water vs Hour of day
Hour of day
Cum. Vol. of Water vs Day of event
Day of event
Cum. Vol. of Water vs Month of event
Month of event
Cum. Vol. of Water vs Average temp change
Average temp change
Cum. Vol. of Water vs Average rate of temp change
Average rate of temp change
Cumulative volume of water drawn
0
250
500
750
1000
1250
1500
0.0
0.2
0.4
0.6
0.8
1.0

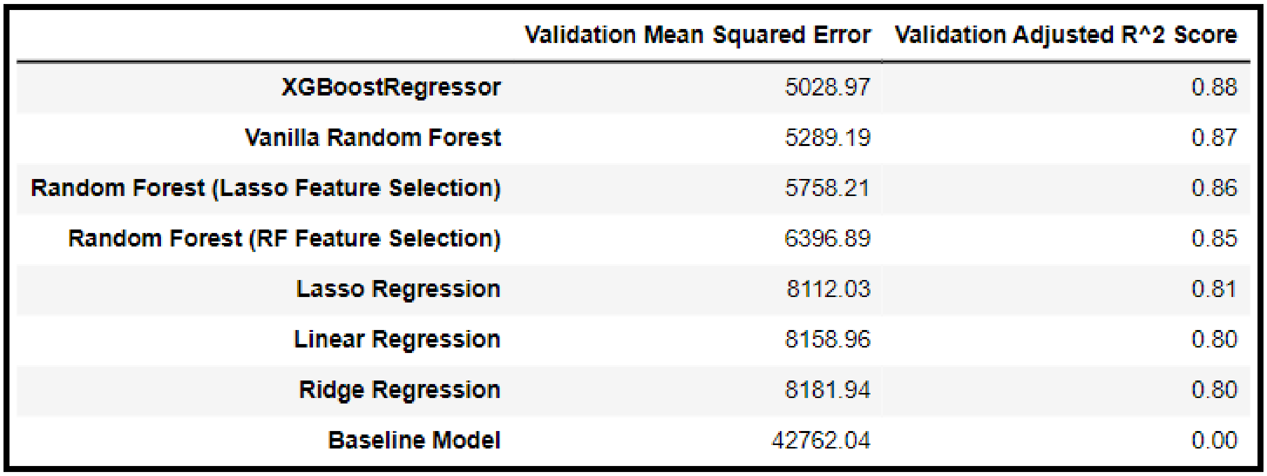


Demand for Hot water over the course of the day
Cumulative volume of water drawn in Litres
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of day


Linear Regression - Feature Importances


Feature Importances for the Decision Tree Model


Feature Importances for the Random Forest Regressor Model



	Validation Mean Squared Error	Validation Adjusted R^2 Score
XGBoostRegressor	5028.97	0.88
Vanilla Random Forest	5289.19	0.87
Random Forest (Lasso Feature Selection)	5758.21	0.86
Random Forest (RF Feature Selection)	6396.89	0.85
Lasso Regression	8112.03	0.81
Linear Regression	8158.96	0.80
Ridge Regression	8181.94	0.80
Baseline Model	42762.04	0.00