# **Data Science Work Product: Predicting Volume of Water Draws**

[boris\_plentify\_internship.ipynb](https://drive.google.com/file/d/189HVpAQfBx13adEAUmXg4nZt5gFGHQoP/view?usp=sharing) and [geyser\_prediction.csv](https://drive.google.com/file/d/1AGlb3eHoubN5T53RV8xqR_ebeHYrCali/view?usp=sharing).

After importing the [geyser\_dataset.csv](https://drive.google.com/file/d/1rEPBnTY_Fcmasv8S8GAZWFVXHDuvRZqq/view) and the [geyser\_test\_set.csv](https://drive.google.com/file/d/1N8gz0O6j70iD7C7NivCcopuHMs5UOR7_/view). The first step was to identify the number of observations in both data sets which was respectively (34842, 8) and (2895, 8). Appreciating the presence of 8 features in both datasets, the target variable in the geyser\_test\_set.csv (test data set) was dropped which is the **Volume of water drawn.** The material provided at this [link](https://docs.google.com/document/d/1NX7w77hecOzaWFFuDG68ZZlxKewSUXRpV-w6PEue8Ww/edit) described the various features.

Commenced by checking for missing values in both test and train dataset and was happy to observe that there were no missing values. Proceeded to check if the dataset was balanced and it was observed using the **Water flow** feature that over 30,000 of the records were obtained when there was no water flow from the geyser in the train and over 2500 in the test data set. For better manipulation of DateTime in the **time\_index** feature was set to type datetime64 (this feature was previously an object) both for the test and train data sets.

In other to appreciate how frequent data was being collected, periodicity was evaluated in both train and test and the average was seen to be 299.24849 seconds that is approximately 5 minutes time interval and it was also observed that the volume of water drawn is always a multiple of 5 litres. For better manipulation, the **Water flow** True/False were mapped to 1/0 respectively. Then proceeded to the creation of new features. First element, mean, sum were used amongst other calculations to evaluate new features and how they were used in generating water draw events after adjusting the timezone to GMT + 2 as presented in the dataset can be seen in [boris\_plentify\_internship.ipynb](https://drive.google.com/file/d/189HVpAQfBx13adEAUmXg4nZt5gFGHQoP/view?usp=sharing).

The features considered where, time\_index, Internal temperature, Inlet temperature, Outlet temperature, Ambient temperature, Energy consumed, Water flow, **inlet\_outlet\_temp\_diff, internal\_external\_temp\_diff, duration\_of\_flow, inlet\_outlet\_pressure, internal\_external\_pressure, time\_index\_year, time\_index\_month, time\_index\_day\_number, time\_index\_day, time\_index\_hour, time\_index\_minute, time\_index\_quarter**, Volume of water drawn. The pressure calculations were obtained based on the relationship PV = nRT where it was assumed that nR/V is proportional to count of Water flow multiplied by 5 minutes (duration between consecutive records in initial dataset.

Inlet\_outlet\_temp\_diff = Inlet temperature - Outlet temperature

Internal\_external\_temp\_diff = Internal temperature - Ambient temperature

duration\_of\_flow = sum of time\_index(t + 1) – time\_index(t) for water event

Monday and Thursday were observed to be the days of the week with the highest water flow events with 12 – 17 being the time interval of the day with the highest water draw events. In the train data set, approximately 1200 (majority) of water draw events lasted 5 minutes that is 5 litres of water with the highest water draw event lasting 897.248 seconds (14.65 minutes) with an average volume of 380 litres (1140 litres total) of water drawn on the 7th of October 2020, started at 09:38:25. A description of the test and train events was observed to check for any anomalies like negative values (min) of which none were observed.

The dataset was split in the ratio 1:3 with 25% for validation and 75% for training. The following untuned models were used: LinearRegression, Ridge, Lasso, ElasticNet, BaggingRegressor, RandomForestRegressor, ExtraTreesRegressor, KNeighborsRegressor, DecisionTreeRegressor, SVR, LGBMRegressor, CatBoostRegressor. CatBoostRegressor performed best on the validation dataset with an MAE (mean absolute error) of 18.03 litres. The MAE metric was used given that we have a 1D regression problem and we wish to know by how many litres can we adjust our predictions to be able to satisfy our clients. Using GridSearch with cross-validation of 5 folds (given the small size of the training dataset that is 1844 water draw events), parameter tunning was done and the most suited hyperparameters were obtained and the whole dataset was used for training and prediction done to obtain the average volume of water drawn in litres which was adjusted by multiplying with the **Water flow** and rounded up to the nearest multiple of 5 given we previously established that water was drawn from the geyser in multiples of 5 litres. Any forecast that was less than or equal to 0 was adjusted to the minimum allowable volume of water that can be drawn which is 5 litres.