Methodology :

Light GBM – In order to train the model and fit into the test data, which has around 20M and 41M rows respectively, the “groupNum\_train” feature has been implemented, it divides the data into 60 groups, using 15 site\_id’s and 4 meter types, during the training and testing phase, a FOR loop has been implemented to iterate within 60 groups and get X\_train, Y\_train, X\_valid, Y\_valid, X\_test and Y\_test. This files has been deleted after training into each iteration save the memory and only the output log, regression errors and model has been saved, suffixing with the groupNum\_train.

In order to spilt the data into train and validation, k-fold cross validation without shuffling has been used, so that at least near-term data is not included in validation. In the dataset, multiple combination of folds has been used to spilt the data, where it has been observed that 3 folds works the best.

Model Parameter Tunning:

LightGBM uses the leaf-wise tree growth algorithm, while many other popular tools use depth-wise tree growth. Compared with depth-wise growth, the leaf-wise algorithm can converge much faster. However, the leaf-wise growth may be over-fitting if not used with the appropriate parameters.

For better accuracy lower learning rate of magnitude 0.1 is used with large number of iterations, i.e 200, while the number of leaves has been kept as low as 31 to deal with the overfitting, going lower than that has resulted in lower accuracy.

By deafult bagging is disabled, it has been set enabled in the parameters to randomly sample from the train data without replacement. Bagging frequency has been set to 5, to control how often a new sample is drawn. Bagging fraction has been set to 0.1, i.e. 10%, as the training data is large and lower fraction ensures reduced trainging time.

To pick best iterations root mean square error has been implemented in the metric, and early stop has been set to 20 iterations. That means after 20 consecutive rounds if the model’s accuracy fails to improve then the LGBM stops the training process. Early stopping rounds reduce the overall training time.