****

**Quality Assurance & Prototyping**

**AIDI 2005 – CAPSTONE TERM 2**

**MARCOS BITTENCOURT**

AHNCH BALA 100424062

SONAKSHI KARKERA 100720763

SURBHI THAKUR 100732335

ARUN KALAESWARAN 100771700

**March 6, 2020**

# ***Executive Summary***

The following report will outline the results of various algorithms we used to test against our dataset. Our dataset contains building, weather and location information provided by ASHRAE in order to predict future energy consumptions of buildings. The following are some steps we took to examine the data:

1. Data Cleaning
2. Feature extraction: Analysing attributes having correlations among each other in order to avoid the unnecessary attributes. Extracting features like hour, weekday, is\_holiday from timestamp.
3. Feature engineering: We notice that the meter reading is extremely skewed. We will be applying a log transformation to the meter reading in order to see a normal distribution.
4. Datatype conversions for relevant features.
5. Dropped columns which consist of null values.
6. We replace columns with Boolean values with 0 and 1 and null values with most frequent values
7. Once we have the final dataset, we then split it into 80% test and 20% train.

# ***Model Building***

Traditional train test method suffers from high variance test problem. We usually the data to train and test set. We will not touch test set until the end of the computation and the final performance evaluation. Then, we can divide the train set to train and validation sets. We use the validation data set to tune the model.

We have used 3 different classification algorithms to compare from:

1. Linear Regression

2. Light GBM

3. Keras Neural Network

Limitations: We faced a few limitations with out data when examining. We realized that in order to achieve even more accurate results we needed more information, particularly the financial information for the restaurants, which were not provided by ASHRAE.

Future Improvements: We will scale the data and do more of feature extractions and engineering in order to get more accurate results. We can use some technologies like SMOTE to produce ample amount of synthetic data. This will help in getting efficient predictions.

# ***Algorithm Evaluation***

## Linear Regression

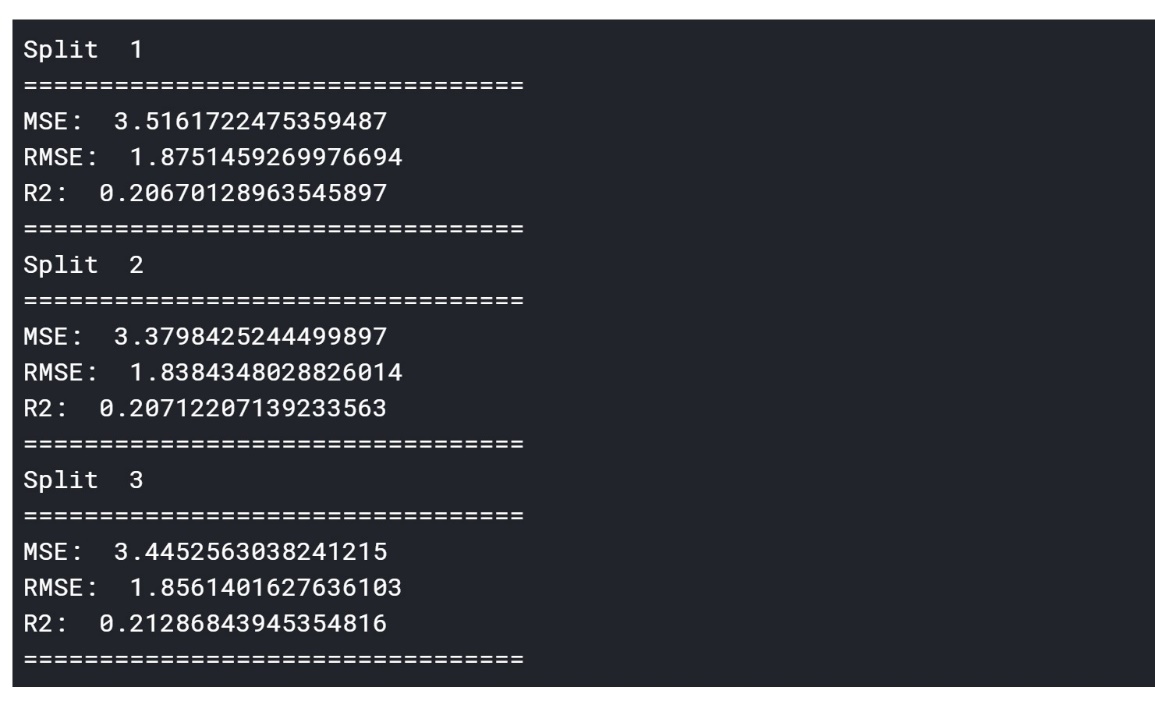
Linear Regression is one of the basic algorithms for prediction problems. It is the supervised learning algorithm. It works well on the linear models.

Pros

* Small number of hyperparameters
* Easy to understand and explain
* Can be regularized to avoid overfitting and this is intuitive

Cons

* Input data need to be scaled.
* May not work well when the hypothesis function is non-linear.
* A complex hypothesis function is difficult to fit
* Prone to overfitting with many features are present
* May not handle irrelevant features well
* Very sensitive to outliers



# Light GBM

Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks. It produces much more complex trees by following leaf wise split approach rather than a level-wise approach which. is the main factor in achieving higher accuracy.

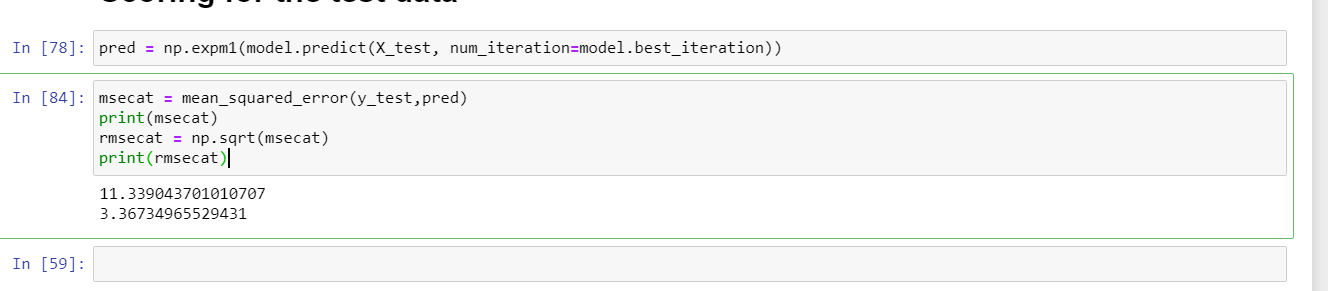
Pros

* Faster training speed and higher efficiency
* Lower memory usage
* Better accuracy than any other boosting algorithm
* Compatibility with Large Datasets

Cons

* Has many hyperparameters to tune, so model development may not be as fast
* Feature importance’s may not be robust to variation in the training dataset





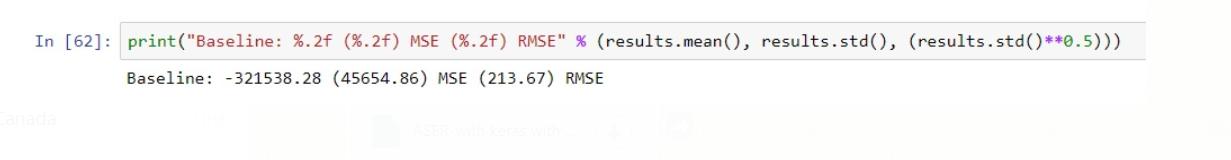
# Keras Neural Network

Keras is a high-level neural networks API, written in Python and capable of running on top of [TensorFlow](https://github.com/tensorflow/tensorflow), [CNTK](https://github.com/Microsoft/cntk), or [Theano](https://github.com/Theano/Theano). It was developed with a focus on enabling fast experimentation. Pros

* Capable of running on top of either TensorFlow or Theano.
* High running speed
* Easy to extend
* Runs seamlessly on CPU and GPU

Cons

* Lack of pre-trained models
* Errors thrown are difficult to debug.
* Not enough examples.



# ***Candidate Algorithm Selection and Rationale***

Using the above information regarding the algorithm evaluation, we can conclude that the top algorithm is Light GBM. After that it comes Keras Neural Network and Linear Regression respectively.

The RMSE score for the algorithms are below:-

**LightGBM – 3.36**

**Keras Neural Network- 213.67**

**Linear Regression- 1.8**

Going forward, we will prioritize data scaling, feature selection and engineering to ensure that the we get better results from the selected candidate algorithms.