DURHAM
COLLEGE
SUCCESS MATTERS

AIDI 2005 – CAPSTONE TERM 2

MARCOS BITTENCOURT

Modeling Part II

AHNCH BALA 100424062

SONAKSHI KARKERA 100720763

SURBHI THAKUR 100732335

ARUN KALAESWARAN 100771700

FEBRUARY 21, 2020

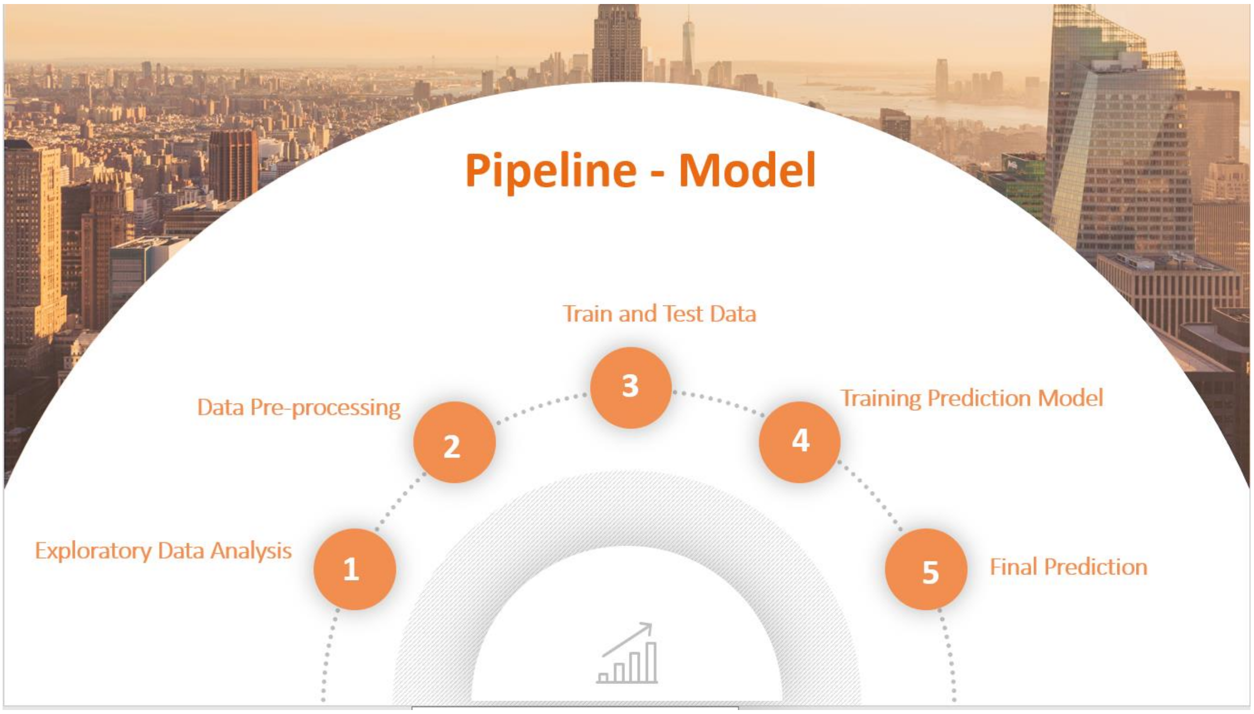
Our project is about the prediction problem. The aim of the project is to predict energy based on different meter readings like electricity, hot water, chilled water and steam meters. The dataset provides 3-year readings for 1000 buildings depending upon geographic domains.

Model Architecture

We have been provided by five csv files – Building_metadata, weather_train, weather_test, train and test. We have merged all the files into two files train and test. We are dealing with the prediction problem so we will be training prediction models based on train.csv file and make the predictions based on test.csv.

We are going to engineer the data so that we combine datasets. Our train and test dataset need to be combined with the building metadata as well as the weather train and test datasets. This is because it will be easier to process and understand the data in a bigger picture instead of analyzing individually. They will need to be inline using the site ID and the building ID as the data.

Pipeline



Pipeline - Model

1 Exploratory Data Analysis

2 Data Pre-processing

3 Train and Test Data

4 Training Prediction Model

5 Final Prediction

Data Assumptions

The assumptions we are going to be making is that the buildings have not been modified or fitted with any energy conserving material after the readings that we have been given. That will affect the future readings of meters and would show a greater disparity from our predictions. Also, that waste of energy doesn't occur, so that meter readings are spiked due to someone mistakenly leaving the windows open, etc. With global warming impacting climate, we are assuming that weather conditions stay constant year-round despite the notion that global warming is affecting climate in areas. Another assumption that is made, is that the buildings have stayed in the same condition as it was when the readings were taken, even though the buildings have undergone wear and tear.

Data Limitations and Constraints

We are limited to only a month of data in our training dataset. More data in our training dataset would usually yield a more accurate model for us to predict with.

We are constrained to the data that we have in 2016 and 2017. To predict now based on data on those years, would not take into consideration changes that may have occurred in 2020.

Model Selection Scorecard

The problem at hand is a regression problem, so using Linear regression would be an ideal classifier in this situation. We will also be using complex models like Light BGM, Deep Learning and Keras Neural Networks to compare it with a simple model like linear regression and choose the most appropriate one based on its performance. The evaluation metric we will be using to measure the performance of the model is RMSE, R-square and Training and Testing Accuracy Score.
The metrics we will be using for our problem will be as follows:
Mean Squared Error (MSE): An average of the squared difference between the target value and predicted value by the regression model.
Root-Mean-Squared-Error (RMSE): The square root of the averaged squared difference between the target value and the value predicted by the model. It is important when avoiding large errors
R^2: Otherwise known as the Coefficient of Determination. It helps us compare our model to the baseline and tells us how much better our model is.

Model Name	Matrix Score		
	MSE	RMSE	R^2
Linear Regression			
LightBGM			
Keras Neural Network			