## 

## **AIDI 2005 – CAPSTONE TERM 2**

## **MARCOS BITTENCOURT**

## 

## **AHNCH BALA 100424062**

## **SONAKSHI KARKERA 100720763**

## **SURBHI THAKUR 100732335**

## **ARUN KALAESWARAN 100771700**

## **FEBRUARY 7, 2020**

## 

## 

## 

# Problem Statement

# Older buildings have not been built to the energy efficient standards of today. Because of this, buildings are more expensive to maintain and are making detrimental impacts to the environment. Determining the specific problems of these buildings can ensure that costs are decreased, and the environment is protected.

# Rationale Statement

The main goal of this project is to apply machine learning to predict the energy usage of a building based on electricity, hot water, chilled water and steam meters. We’ll be using the dataset provided by ASHRAE as part of their initiative to advance the arts and sciences of HVAC and related fields. The dataset includes information about energy usage of 1000 buildings over the last 3 years.

Along with the main goal, we have an additional goal. A lot of buildings are getting older and have not been built according to modern energy efficient standards. This means that owners of those buildings are looking to invest money to make the buildings more efficient. The question is how does the owner know that the money they are investing is being put into good use? Are the improvements working in reducing energy usage? With our model, we can provide better estimates which makes business owners, financial institutions and big investors more inclined to invest in these areas in regarding building efficiency.

# End User

We will be the end users of the solution. We will propose the solution to various construction companies, building owners. People can sign up for the subscription packages to get the details of their energy consumption.

In Basic Subscription, customer’s will be provided by their quarterly and yearly energy consumption with detailed Kwh usage connected with the dollar amount. This will give the general report to the customers made using Tableau. The Premium Subscription package will include the detailed Kwh usage connected with the dollars for each month, quarter and year. Premium subscription will also give benefit of email alerts if their logged consumption and price limit is exceeded.

# Metrics

For this regression prediction problem, we will be measuring precision, recall and accuracy as evaluation metrics for testing the performance of different machine learning models on our dataset. We will also use the confusion matrix with accuracy in order to determine true positives, false positives, true negatives and false negatives.

The metrics used in our dataset are meter readings which is the target variable, measured in kWh (kilowatts per hour). Another metric is square feet for each building. We need to measure energy use on a per unit level like per Square feet of a building (kW/Sqft). We will be providing energy consumption on a monthly basis for clients. Another metric that will be considered is weather. Metrics like air temperature, precipitation and wind speed will be measured and provided as inputs. The metrics we will be providing for our business model is the amount of kW saved per month and used per month which will also be shown in monetary value as well.

# User Interface

As mentioned before, clients or end users will have a subscription with our business. Once the subscription has been established, we will provide users with login information. The users will log in and be presented with a dashboard which provides them with all insights and analytics related to energy consumption by their building.

# Data Analysis

## Data Requirements

We will be using a data set that has been provided by ASHRAE that they created for the Kaggle competition. Kaggle wants individuals or groups to use ASHRAE’s data in order to predict future energy consumption of buildings. ASHRAE stands for the American Society of Heating, Refrigeration, and Air Conditioning Engineers. ASHRAE was created in 1894 to advance the science of heating, ventilation, air conditioning refrigeration and their associated fields.

We require our dataset to have meter readings , when they were taken (time stamps), types of energy (electric, heating, etc.), the weather outside (precipitation, wind, temperature, etc.) and description of the building that we are taking the readings from (square feet, floors, etc.).

**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 making the assumption 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 definitely undergone wear and tear.

**Limitations -** 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.

**Constraints -**  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.

## Data Exploration

Our data has already been split into train and test datasets. The train data is from January 2016 and the test data was from approximately January 2017- December 2017. The dataset contains 5 csv files: building metadata, test, train, weather test, and weather train. Since these are authentic and detailed datasets provided by Kaggle, the file sizes are quite high (2.5 GB). Due to this, we may need to switch to cloud options for training our model.

|  |  |
| --- | --- |
| *Independent Variables* | *Dependent Variables* |
| Building ID | Meter Reading |
| Meter |  |
| Time stamp |  |
| Site ID |  |
| Primary Use |  |
| Square Feet |  |
| Year Built |  |
| Floor Count |  |
| Air Temperature |  |
| Cloud Coverage |  |
| Dew Temperature |  |
| Precipitation Depth 1HR |  |
| Sea Level Pressure |  |
| Wind Direction |  |
| Wind Speed |  |
| Row ID |  |

# Data Preparation and Cleaning

In order to prepare the data, the first step we would take is to remove all null values from the provided dataset as these will hinder our results. We will also be looking at correlations and seeing which variables might not be needed as a part of our analysis. Lastly we noticed that the meter reading is extremely skewed. We will be applying a log transform to the meter reading variable in order to see a normal distribution.

## Data Engineering

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.

**Methodology and Testing Procedures**

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 Gradient Boost, 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.

In our context, We will be predicting the value for “Meter Reading” based on various features. To come up with the accurate Meter Reading we will analyze the multicollinearity between the dependent and independent variables to know the impact of the features on the final output. To improve the accuracy of the model, we can further include  PCA(Principal Component Analysis) to deal with the high dimensionality and to optimize the model. We can further look into neural networks by adding more dense layers, increasing the size of epochs and batch size.

We will use a test/train split on our training dataset to divide the dataset for training and validation purposes. After the data is fed to the model and model is trained. It will be tested on the available test dataset.

The metrics we will be using for our problem will be as follows:

1. Mean Squared Error (MSE) : An average of the squared difference between the target value and predicted value by the regression model.
2. 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
3. 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.