**Introduction**

This is group 10’s proposal which includes Jesse Franks and Joshua Hall working on this project together. Our project is called “**Classification, Clustering, and Time Series Analysis of London Smart Meter Data**.”

I work at an electric and gas utility company in Baltimore, MD. This means I have domain knowledge of the electrical grid which I have used to form my research question. My decision to use the London smart meter dataset was based on a survey paper of machine learning applications in the utility sector that I read last year. That paper said that not a lot of research was done on that dataset and that is one of the reasons that I am exploring this dataset. Another reason I am interested in this dataset is I am interested in doing customer segmentation so that populations of customer can be identified by their electric consumption profile. This information can aid in further analysis like detecting energy theft by utility customers by comparing an individual’s electric consumption profile with respect to the customer population they belong to by checking if there are any outliers. This saves energy companies money and creates values for customers. But for now, for now I intend to focus on 3 main lines of analysis which I will describe later.

**Dataset**

My dataset has a sample size of 5,567 London households. This mean that there are 5,567 smart meters. Smart meter readings were sampled every half-hour between November 2011 and February 2014. Each sample is represented by a unique customer by ID and is in fact a whole time series. So, this dataset has 167 million rows or about ~30,000 readings per household. The columns in this dataset are:

* **LCLid** – Is the household ID for the whole series.
* **StdorToU** – Is a Boolean on whether the household is in the ‘Dynamic Time of Use’ program or not, which is a service that text the consumer on their smart phone what the price per kilowatt hour will be the next day (low, normal, or high) via day-ahead forecasting model. Which is supposed to allow the customer to make more informed decision about how to control their electrical bill. Only 1,100 customers were in this treatment group. The other household received no information.
* **Datetime** – This column is a datetime stamp of when the energy reading of the customers smart meter was taken.
* **KWH/hh** – Is the amount of power the household was using at time t.
* **Acorn** – Is a categorical column of classifications of economic class based the household’s income or place of living.
* **Acorn grouped** – is discretized column of ‘Acorn’ classification from ~20 groups to 3 labeled as affluent, comfortable, and adversity.

The data can be found here: <https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households>

Text

Description automatically generated

Fig 1.1 – This is a five number summary of the amount of kilowatt hours per half an hour per group used. Broken down by acorn group and whether they participate in texting day ahead service.

A picture containing table

Description automatically generated

Fig 1.2 – Number of households that belong to each group.

Chart, histogram

Description automatically generated

Fig 1.3 – example of time series of customer MAC004221 (std and affluent).

**Research questions**

1. Are the different populations (std or ToU) energy consumption profiles different? If so, can they be broken down further by economic group?
2. Can you cluster customer profiles in such a way that allows you to do outlier detection to prevent energy theft? Is this the enough data to detect energy theft?
   1. Note: This dataset doesn’t have labeled example of energy theft so I can’t prove that the models I build will detect energy theft, but I think this is a good approach-based papers I’ve read in the past.
3. Can I build an accurate time series model that predicts the next day’s, next weeks, and next year’s energy consumption?
   1. The value in this is that you only want to produce as much power as you’ll need so the more accurately you forecast future energy needs the more money you can save as a company.
4. Can I classify the customers economic acorn group by their energy consumption alone?
   1. This is valuable because I know from my job, we typically don’t have a lot of economic information about the customers. So, this would be useful for placing in the right category for energy theft detection.

Research questions 1, 2, and 4 are about finding useful insight into my customer population which will allow me to serve them better if I know more about their power use habits. Question 3 is about predicting how much power needs to be generated at peak times each day because generating extra power when you don’t need it cost money. So, the idea is to not make more than you need.

**Methods**

My analysis will be in effort to answer the above questions. I have four plans right now.

1. I intend to answer this research question by averaging the consumption patterns for each customer class present in the data(there are 6 labeled classes as of now as seen in Fig 1.1) to make a line graphs to see if they are different enough.
2. My next plan is to use the methods described here to do time series clustering:
   1. <https://www.kaggle.com/code/izzettunc/introduction-to-time-series-clustering/notebook> and <https://github.com/JustGlowing/minisom>
   2. Minisom is a machine learning technique using a neural network to do unsupervised clustering on time series data. I hope the results from the minisom will be of use allow me to do outlier detection for each customer class because I would compare the customer suspected of energy theft to the cluster profile that they belong to.
3. For the time series prediction models, it is my intention to capture all the seasonality from each type of customer using harmonic regression because I know meter data has multiple seasonality.
4. For the customer classification question, we are going to try use algorithms such as SVM, decision tree, neural network, Naive bayes, random forest, and KNN.

We anticipate having to learn a lot about how to work with time series data when it is applied to the classical machine learning algorithms, we used in the past but both Josh and I have worked with time series data before so its just the new applications we’d have to learn. **NOTE: Josh just joined my project on 4/11/2022 so we will be meeting to discuss the project this week as he hasn’t had a chance for input yet.**