Eric Zan
Divya Karthikeyan

Usage Behavior in Rural Microgrids

Introduction

Microgrids are a local self-contained electricity generator. UC Berkeley's Center for Effective Global Action (CEGNA), the Center for Information Technology Research in the Interest of Society (CITRIS), and the Electrical Engineering and Computer Science (EECS) department have partnered together to develop microgrid projects in rural India and Kenya. The microgrids record household power usage metrics and the current credit amount.

We looked at data from one village in Kasii, Kenya and attempted to discover types of users based on their power usage and credit behavior, to look at whether there is a relationship between power usage and credit behavior, and see how peak load times vary among users.

Using clustering methods, we discovered the households were categorized into the same groups when clustering around either power usage or credit amount. From this, the type of credit transaction behaviour can be predicted if power usage of a household is known. However, further work is needed to overcome the limitations of the polynomial regression model used for classifying a household into a certain user behavior type. It was also found that each user type had a unique peak usage profile which can help predict heavy load times.

These insights may be enough to form the basis for economic experiments for understanding demand as well as contribute to better maintenance through balanced load distribution.

Problem

There is a desire to study the behaviors of these newly electrically connected populations in an effort to help efficiently bring reliable electricity to the remaining 1.5 billion people on the planet without it. This large population suggests there is likely to be a significant increase in demand for energy in the near future and energy forecasts have implications for the appropriate scale in energy infrastructure and impact on the environment. So, the UC partners are trying to understand the development impacts of rural electrification, as well as its demand and usage behaviors to inform economic experiments, all with the goal of efficiently expanding electrical usage in a sustainable manner.

Economists at CEGNA were previously limited to survey data on power usage. To provide more accurate and timely data, a smart grid approach was taken and every microgrid records household power usage metrics and its current amount of credit. A prepaid metering system is used where customers use SMS text messaging to pay for and monitor their usage on demand.

The data is then sent daily over SMS where it is stored and then sent back to Berkeley. This granularity of data presents opportunities to predict types of usage behaviors and power usage for particular times of day or week.

The Data

The data are provided by a Phd student from the EECS department at UC Berkeley who is partnering with CEGNA. The data set contains 3 months worth of transactions of power consumption from eleven households captured once every 3 minutes and aggregated to the hour. It is composed of:

- ID: House Identifier
- Date: Time of Day
- Watts: Total cumulated watts during the present day
- Day Watts: Derived cumulated watts during daytime for current day
- Night Watts: Derived cumulated watts during nighttime for current day
- Credit: Current amount of credit for the household

Data Challenges

A more granular data set contained samples at the 3 minute level but only contained complete data for 2 weeks. Thus, an aggregated data set described above was chosen. Only 2 days are missing from the aggregated set.

Solution

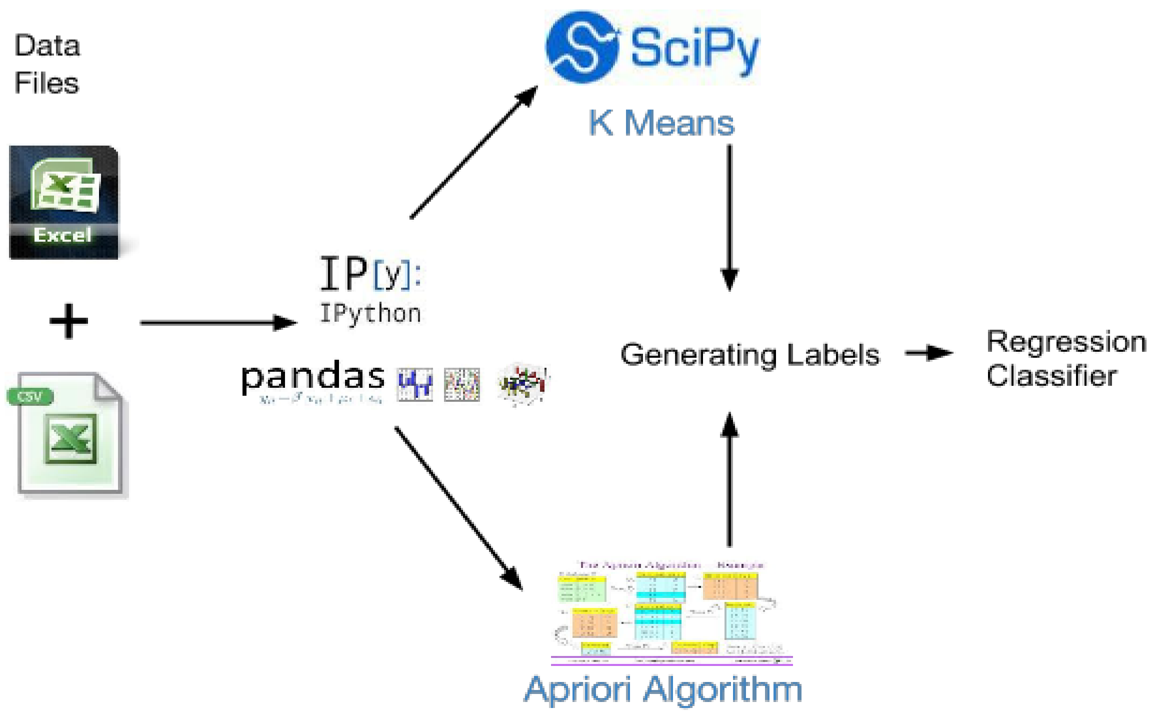
The time of day data was truncated to be on the hour. This simplified the model and enabled time series clustering with K means since the length of each time series needs to have an equivalent euclidean distance.

The next step was to use unsupervised learning and association rule learning methods to observe natural groups among the users in our dataset:

- Apriori: Discover association rules among users, power consumption and credit transaction
- K-Means: Group users based on their power consumption and credit history
- Classifier: Create a regression model for each cluster group for categorizing households

Process

Eric Zan
Divya Karthikeyan



K-Means

The K means clustering was performed separately on the power consumption and credit transaction data. This optimum value of K was identified by clustering with K equal to the number of observations in the data set and every value less until it was equal to one.

The resulting graphs at each step were analyzed to determine which value of K returned the best results. After inspection a K of three was chosen.

Apriori

The apriori association rule learning algorithm was used to identity rules between the household and behaviour. Since the result only generated few rules of significance, we just used them as a means of supporting the validity of clusters we generated through K means. This result was used to aid in labelling the clusters. For example, a set of households has a strong correlation with null power consumption at night and were labeled *"Day users"*

This process uncovered data integrity issue when probabilities for night users did not match insights gained from K-means. After inspecting the raw data, it was discovered the night wattage is not consistent aggregated correctly..

Labeling

Users were clustered in two different sets corresponding to their power usage and credit transaction behaviour. Labels were created for the groups by observing the graphs representing the average among the users in each cluster.

Classifying

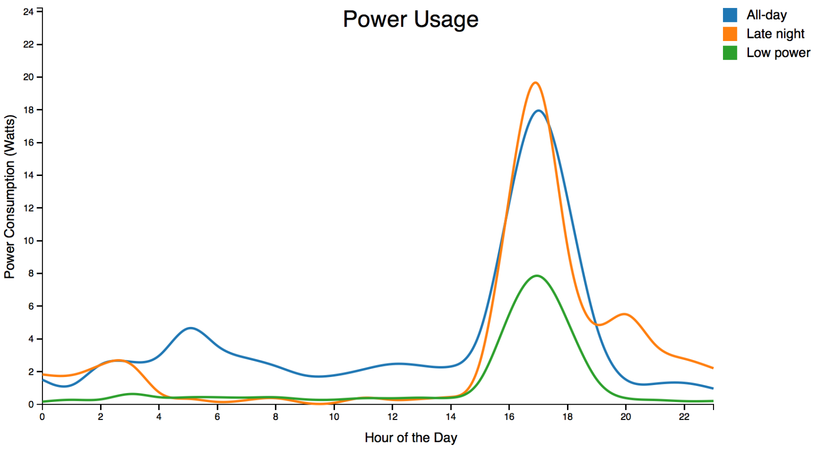
A polynomial regression model was generated for the waveform of watts throughout the day for each resulting cluster group. The large variance across the time of day affected the accuracy of the regression fit using the Python numpy model. The cluster group data points are similar enough to force any individual household data into the group which has the best regression fit. Therefore a classifier was not successfully generated and more work is needed to find an accurate model for these specific waveforms.

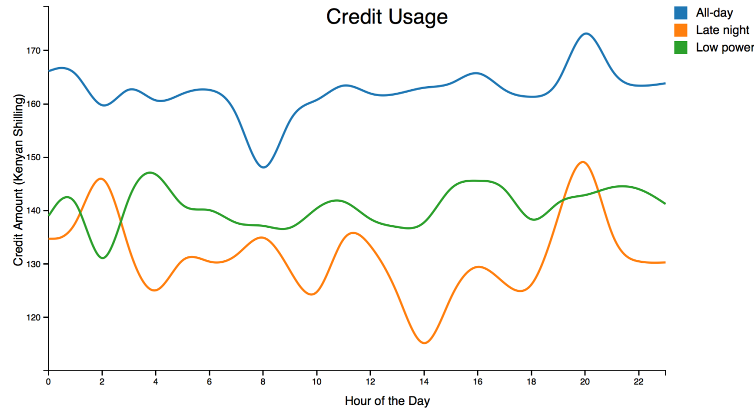
Results

The clustered groups for power usage and credit each contained the same households. This suggests a strong correlation between power usage behavior and credit transaction behavior.

User groups based on Power usage and credit behaviour	**Cluster Label**
Household 1	All-day
Household 2	Low power
Household 3	Low power
Household 4	All-day
Household 5	Low power
Household 6	Low power
Household 7	Low power
Household 8	Low power
Household 9	All-day
Household 10	Late Night
Household 13	Low power

Eric Zan
Divya Karthikeyan





Based on the power usage and credit transaction the users are grouped into 3 types:

1. *All day:* Power is used throughout the day, peaking in the evening. They maintain a higher credit value in comparison to the other types. The credit peak occurs after their power usage has a peak value.
2. *Late night:* There is near zero power consumption during the day. Consumption peaks in the evening and is non-zero through the night. There is a large variance in the credit behaviour. Peak credit increases happen before the maximum power consumption in the evening.
3. *Low power:* There is near zero power consumption except during the evenings. They maintain a relatively stable credit value in comparison with the other types. Credit peaks occur many times throughout the day.

Related Work

- Wang, Xiaozhe, et al. "A scalable method for time series clustering." *Unrefereed research papers* 1 (2004).
  - Helped justify rounding the time to the closest hour in order to perform time series clustering using K means since the length of each time series needs to have an equivalent euclidean distance.

- Thangsupachai, Noppol, et al. "Clustering Large Datasets with Apriori-based Algorithm and Concurrent Processing." Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 1. 2011.
  - Apriori is used for feature identification to cluster upon. Since we wanted to see some natural clusters forming in the dataset, Apriori seemed to be a good means of identifying relationships within the dataset to help identify features for clustering.

Future Work

- Analyze weather's effect on power consumption.
- Analyze how demographic data is associated with power and credit behavior.
- Produce an accurate classifier by tuning the regression model.
- Analyze day of the week's association to power usage.