ISYE6501 – HW8

Power Company Case Study

The goal of this report is to identify which power company customers need to have their power turned off, and also to suggest efficient ways to make that happen. The main topics to address are as follows:

1. Will customers pay?
2. How much power will be used in the next month?
3. Which customers should have their power shut off?
4. How much money can the company save?

We will start by looking at the data that we would like the power company to have access to. It was suggested in the lectures that some data that could identify ethnicity and gender were illegal to use, so we will choose to use the following reduced list of items:

- Credit score
- Income
- Past default history
- Past power bill payment history
- Rent or own accommodations
- How long has customer lived there

Part 1: Given the above data, we would use a support vector machine model to classify the customers into the 3 following groups: those who will pay, those who can't pay, and those who can pay but won't. Once we can successfully identify the characteristics of those in the "can pay but won't" group, we can also use this system to classify new customers.

Part 2: Cost / power estimation – cost of leaving power on or off

Given the above data, we use a linear regression model to predict the amount of power a "can pay but won't" customer will use next month. We will also include an error term to reflect the fluctuations in power usage as follows: given the same data above, we will use a GARCH model to estimate the variability (sigma^2) in a customer's power usage next month. We will add this variability to our regression model by adding an error term with distribution N(0,sigma) to our regression equation.

Part 3: Power-off determination

We would like to figure out which customers should have their power shut off. Given the classifications determined in Part 1 and the estimated cost of leaving power on determined in Part 2, we will use an optimization model to help us determine which customers should lose power. The customers who are selected for shutoffs will be coded with a binary variable where 1 indicates that power should be shut off. The objective function would be the cost difference between shutting power off or not and the savings the company would get by shutting off power times the binary variable. The constraints of this function would be the worker's hours/availability, ensuring that the chosen customers are in the "can pay but won't" category, and some additional routing constraints that are to be determined in the next section.

Part 4: Vehicle routing problem & Money-saving simulation

We would like some additional data for this part in the form of past data on travel time. This data would hopefully include customer address, time of day, amount of time it takes to shut off power, amount of time in transit to location and route taken, so that we could have a better idea of flow rates throughout the day in the specified city.

Given the information determined in Part 3 and the additional information above, use a clustering algorithm such as knn to determine communities of "can pay but won't" customers that are located close together. The distance function used would possibly be Manhattan distance, as this would model the street layout of cities and homes more realistically. Then we would weigh the respective clusters based on the average value each cluster would provide in regards to shutting off power (sum up the total customer power usage in each cluster and divide by the number of customers). We would assign workers to start shutting off power at the cluster with the highest average total usage value. Another approach would be to use the Louvain algorithm to determine the clusterings, and it would be interesting to see if these clusterings are different from what knn provides.

In the final stage of the solution, we would hope to use a simulation. Given all of the above data, we create a simulation of the city to figure out how many workers/trucks should be doing the shutoffs in order to maximize savings for the power company. We could first use distribution fitting to determine if the "can pay but don't" customers are a good fit to a Weibull or other distribution. Then we will model these customers with the relevant distribution and send varying numbers of workers/trucks out into the city to help determine the optimal number of vehicles to use. The nice thing about a simulation is that it's easy to keep track of many different aggregate variables, such as hourly cost of workers/trucks and potential savings of shutting off customer power. We could also choose to model the cost/labor of turning back on power to customers who have started repaying in this simulation, depending on the prevalence of this in historical data. Ultimately, we would have to make sure that the simulation is an excellent match to reality in order to use the results that it provides.

Final thoughts: I wonder what the accuracy rate would be for the initial SVM classifier - would it be able to completely discern between people who can pay but won't, and people who can't pay but would be eligible for an assistance program? Probably not. I also wonder if we would be legally allowed to make decisions based on geospatial data at all, since our clusterings would possibly identify low-income or ethnic minority areas as candidates for losing power. We would need to run this by the power company's legal team to make sure we are not violating any laws.