Homework Twelve

Chelsea Hughes

Question 18.1

Describe analytics models and data that could be used to make good recommendations to the power company.

Here are some questions to consider:

(1) The bottom-line question is which shutoffs should be done each month, given the capacity constraints. One consideration is that some of the capacity – the workers' time – is taken up by travel, so maybe the shutoffs can be scheduled in a way that increases the number of them that can be done.
(2) Not every shutoff is equal. Some shutoffs shouldn't be done at all, because if the power is left on, those people are likely to pay the bill eventually. How can you identify which shutoffs should or shouldn't be done? And among the ones to shut off, how should they be prioritized?

Think about the problem and your approach. Then talk about it with other learners and share and combine your ideas. And then, put your approaches up on the discussion forum, and give feedback and suggestions to each other.

18.1.1: Identify shutoffs

I think the first thing to do is to cluster the population so we can see if and what patterns exist. This should help us determine predictors. To avoid protected classes, we want to measure the wealth of the family but not so directly. So, we'll use data that suggests a higher probability of consistent payments being able to be made. We also don't want to use zip codes, but we do want to know what the climate is because daily average temperatures could influence the amount of power being used and how high the bill could be. Not that that will give us exactly what we need so we can also get around that class by measuring regional population density and distance.

Given Data

- Household size
- Age
- Employment status
- Home ownership (vs. renting)
- Car ownership (vs. leasing)
- Income
- Payment history of electricity bill
- History of electricity bill (size)
- Number of days temperature is above 80
- Number of days temperature is below 55

Use Model
K-Means Algorithm

To Result
Creating categories to later use to suggest probability of payment and non-payment.

Given Data
Once we have the clusters, we can add other predictors to help estimate cost so we can more efficiently and effectively prioritize shutoffs. *Such as*: number of credit lines open per household, regional population density, filed exemptions in tax records or current IRS delinquencies and of course, number of weeks delinquent, the distance to the nearest K (loads of looping required).

Use Model
Logistic regression (probability classification)

To Result
Calculate probability from a fitted model on historical data of non-payment households. If the data is updated every week, or every 5 business days (let's say every Thursday) then hopefully you would want to predict 10-12 weeks in advance. So, from here we could use **Holt Winters** We may also want to estimate the size of the household power bills and we could use **ARIMA** for that. *On second thought, I'd probably just use ARIMA since it would allow for more parameters than Holt Winters.*

18.1.2: Prioritize shutoffs

Given Data
- Size of power bill per household
- History of non-payment in household
- Regional traffic (for technician routes)
- Payment schedule (for technicians)

Use Model
(1) Clustering
(2) Maximization Optimization

To Result
The idea is that we calculate our return investment by maximizing our net benefit to determine who we should shut off. We would calculate that by taking the non-payment probability and multiplying it by a function of shutoff cost (payment schedule and technician routes) while considering the cost of turning power back on after a shutoff. So, this could determine which households would need to have the power shutoff to where the least amount of money would be spent (or lost) with the most amount gained in payments made. I think maybe the best way to visualize this would be a **confusion matrix** or a **simulation**.

So, originally my plan was to use **SVM** but opted out of a supervised classification approach so that I could have more flexibility with less binary information but ideally, I would probably use both and compare the two. Also, maybe an easier route to have taken would have been to utilize measuring Manhattan distance following my clustering to find the highest expected value of power usage.