ISYE 6501 Homework 8

Due: July 15, 2021

1 Power Company Case Study

1.1 Question 18.1

The question of which customers' power should be shut off at first glance sounds simple but is in fact quite complex when this analytics problem is broken down further component by component and detail by detail. To express the intricacies of this problem I have listed considerations and constraints to be explored during the construction of an analytics model. Recall the goal is to turn off the power for those who can pay but are never going to pay their bill and not those who can't afford to, forgot to, or or got behind.

1. Customer identification
   (a) Some of those who don't pay should not have their power shut off
   (b) There are several factors for identifying the people whose power should be shut off: credit score, income, history of defaults on payments to any company, power company payment history, renter vs. home owner, home value, length of residency, married vs. single, number of people in household
   (c) Which model should be used: classification, clustering, or logistic regression, etc.
   (d) Should a single model be used, tree-based approaches or a hybrid?
2. Model for cost estimation
   (a) For some customers, it is more costly than others to wait to shut off their power (and turning back on is also costly)
   (b) There are legal, societal and reputation costs as well
   (c) The cost estimation can be based on several factors like those listed in 1.(b)
   (d) It may also be useful to estimate variability
   (e) Should time-series analysis, factor-based regression, or a hybrid approach be used?
3. Model for shut off selection
   (a) One could find the expected cost of leaving power on or not
   (b) Consider using a vehicle rating model or similar
   (c) The data to consider is: travel time to site data, time to shut off power, estimation of future usage, probability of non payment, gas price, worker's wage, etc.
   (d) Which model should be used: optimization models, clustering, simulation, or a combination
   (e) Traveling to n closer sites whose cost sum to more than m further sites may be worth it even when the m further sites are ranked higher because they individually have higher expected costs
   (f) Is it worth hiring new employees?

The solution to this problem that makes the most intuitive and quantitative sense to me, capturing the subtle complexities without having a large runtime, is first completing logistic regression. I can use logistic regression to first identify the probability of each customer paying their next bill using the data explained in 1.(b). Of course using logistic regression is a supervised learning approach so this assumes that I have access to ample past data on those who have and have not paid their bill or using the current status of those who have or have yet to pay their bill. This data serves as the predictors and labels. If no such data existed then I may have to go with another approach like clustering but for now I will assume I have access to this data. Using this approach would also give me the ability to make easy modifications or tune the separator/threshold.

Now I effectively have the probabilities of each customer paying their next bill and can complete some manual tuning of the probability threshold used for further consideration i.e. how many customers lie beneath the threshold? If I threshold at 50%, are there thousands of potential non-paying customers? Is this number feasible given the fact that there are only 10 workers to shut off their power etc.? This threshold will have to be tuned later in the analysis.

Previously, I wanted to use all customers to have a rough equal representation of each class during training but now there is no need to further analyze those who have no remaining balance. So at this time I would filter out all data points where the remaining balance is \$0 or even slightly higher, maybe a threshold of \$1 because the remaining balance is likely due to a calculation mistake or a mistyping on the user's side. Next, because I have already used the several factors to assign probabilities to each customer, I want to avoid overfitting and so I choose to use a time-series model like ARIMA to estimate the cost of each customer for the next time period. This model works well because I am solely focused on the next time period and can simply rerun the model next month so there is no need to forecast any further into the future. Now I will essentially have the expected cost of each customer (probability times the expected cost).

Lastly, I will preprocess the model for shut off selection step by spatially mapping or plotting the locations of the n thresholded potential non-payers. Note that the distance of each customer is with respect to the location of the power company and where the drivers would leave from. I choose to use an optimization model to optimize for number of workers (expected savings) using data: expected cost, travel time to next location, travel distance to next customer and time to shutoff power, constrained by worker's wage, gas price, and hours each worker works in a day. The number of workers to hire will be when the company's expected cost is minimized. As it has been seen, it may be more worth it to the company to shut off the power of 10 spatially close customers of lower expected cost than 5 spatially far customers of higher costs and this is what the optimization model can solve for me.