Hw 13: Power Company Case Study

Introduction

My solution splits the problem into 3 parts:

1. Defining a non-paying household
2. Classifying them against future-paying households
3. Optimizing their power shut-offs

Part 1: Determining Thresholds

Given

- Payment history: amount and time stamps of transactions, billing, and contact with power company
- Duration of residancy and service

Use

- Cusum model
- Sample Probability Distributions
- K Means Clustering

To

- Determine a threshold for determining when a customer is not going to pay

Probability Distribution
Defining how we can classify non-paying vs future-paying customers means defining thresholds to determine if a household is non-paying. We can split the data into groups: one for those with late payments, and one for those without. We could also split the data between types of housing to reduce bias in the data (e.g. businesses may not miss payments as much). Then, we can plot the time between payments. If the distributions look very similar, it may be possible to combine them and have a larger sample size to analyze.

With the distribution of payment times, we can determine what is a reasonable threshold to decide whether or not the customer will pay again. This can be performed as an outlier detection problem and would be made even easier if the distribution resembles, for example, Poisson for the payment arrivals or Geometric for the intervals. It would be fairly straightforward to look at the probability of a payment later than T periods (days, weeks, months, etc.). Essentially, we would be defining a metric such as "customers who have no paid in 60 days are considered as customers who will not pay". If we data is available on past shut-offs that did not result in re-installations due to evidence of financial troubles, we can use it to help determine the threshold as well.

Later, we can use this as part of our cost calculations.

CUSUM
We can also try to build CUSUM models on the scaled data. For example, rather than looking at the time between payments in days, we can look at the relative distance from each payment to model. That way, houses with a tendency to be consistently very late but always pay will have leeway.

After building CUSUM models, we can decide which houses are to be considered non-paying. By adding a categorical variable for each of these houses, we can further out analysis by looking for patterns in the data in relation to this new variable. This can involve correlation analysis, information coefficients, or even simply performing the distribution method above to see how these non-paying houses compare to the rest of the data. This can enforce the threshold from before or help create a new one.

K-Means Clustering
K-Means clustering can be used both in Part 1 and in Part 2. In Part 1, it can be used to reinforce the decision threshold for setting up our classification data either by refining the houses that should be deemed as non-paying or by adding houses to the list. Similar to the CUSUM model, it can group non-paying houses and the results can be used to determine non-paying vs future-paying.

Or, perhaps there are other groups present that we had not accounted for - while this is a binary classification problem, by including houses that lie in a potential third group, our optimization model may change. For example, imagine a group of houses that are definitively on the fence about being non-paying or paying. While this might come up in the Logistic Regression portion of Part 2, the presence of a powerfully distinct third group may change our binomial classifiers into multinomial models. This is importance since a house that is "unpredictable" in their paying habits introduces more variability in our cost analysis.

Part 2: Classifying Non-Paying Customers

Given

- Threshold found in Part 1
- Same data as Part 1
- Location (zip codes, streets, latitude/longitude), number of residents

Use

- CUSUM
- Logistic Regression
- SVM

To

- Classify Non-Paying vs Future-Paying Households

CUSUM
If historical shut-off data is available, we can use it to aid in our CUSUM model training. By using

them as training or proxy data, we can build CUSUM models on them to match their shut-off decision point. We can use the parameters obtained from these trials to build new CUSUM models. We may be lucky and find that this is all we really need to determine non-paying customers after tuning our parameters to account for the cost of true/false positives/negatives.

We may even build CUSUM models for multiple variables and determine that x of X number of models must pass their threshold for the company to determine a household as being non-paying. This will help in accounting for independent factors in the decision.

Classification Modeling
Another approach is to treat this as a true classification problem. Since we have a threshold for whether or not households are non-paying vs future-paying, we can train, validate, and test a basic machine learning model such as SVM or Logistic Regression.

Using SVM or LR has the advantage in that we can involve more factors and see their relationship with another instead of relying on individual CUSUM models. With the SVM approach, we can strictly classify the houses and report them to the company. However, a preferred method may be to use Logistic Reression since it produces probabilities. This will be much easier for the power company to interpret, and the probability threshold can be adjusted accordingly with more ease in reality. This also follows the idea of the threshold being decided based on a probability. While the cost of shut-offs can be calculated even with SVM outputs, it is easier in industry to explain such concepts using probabilities rather than cost functions.

Part 3: Optimizing Shut-Offs

Given

- Location of shut-off houses
- Number of workers
- Time it takes for workers to shut off power in each home
- Amount of resources available

Use

- Optimization modeling
- Simulation if required

To

- Determine which houses to shut-off

Optimizing
Producing an optimization model might be fairly complicated for this. We have to account for the cost of shutting off a house, the cost of re-installing based on probability of it needing to be re-installed, the cost of travel including going between houses (non-memoryless) in the network, and the additional cost accrued as time passes for each household.

The objective function will be the sum of keeping the house binary (on or off) multiplied by the cost. The cost, however, will be a combination of the factors listed above. We are contrained by the number of workers available and the time it takes to shut-off power in the homes. If the process takes weeks, our model might change the decision of several houses and we must adjust our optimization model accordingly if the cost is worth it.

Simulation

With my current knowledge on optimization, simulations would be very beneficial to this process. Particularly, having a dynamic set of houses to shut-off will be challenging to include in the optimization model. Not only that, but we need to constantly be running our model (SVM, Logistic Regression, or CUSUM) since the data changes literally every second. While we may only be modeling in terms of days, the size of the power company will affect whether or not the process duration should be accounted for.