ISYE 6501 : Homework 12
4/7/2021

Table of Contents

18.1 Power Company Disconnects

Given a power company case with customers who are delinquent on their bills, analytics are needed to determine the following:

1) Customers who should be disconnected. The goal is to disconnect customers who have no intention of paying their bills.
2) Evaluate capacity to turn off customers who should be disconnected.
3) Evaluate the order by which to turn off customers who should be disconnected.

Analytic Methods Considered

- KNN
- Logistic Regression
- Categorical Variables
- Random Factor Trees
- Linear Regression
- K-Means Clustering
- Vehicle Optimization
- Queuing Simulation

Step 1

Data gathering and preparation

It is assumed that the power company will have the following data available for modeling:

- location,
- payment history,
- payment plan,
- account owner income and
- prior energy usage data.

I would limit the data sampled to the last 5 years since market conditions impacting an individual's ability to pay their energy bill can change drastically within this timeframe.

If we wanted to narrow the data even more, the beginning of 2020 with the pandemic would be a relevant timeframe for determining how to differentiate between customers

who want to pay their bills versus those who do not. In this case I would choose data for the past 3 years since there are a significant number of individuals whose income was negatively impacted by job losses during the pandemic.

Determining the amount of data to use in modeling should make use of some practices. Since we didn't dig into that in-depth in this course my statements are based on research that I did outside of the class. There will be seasonality in the data. Seasonality requires that a slightly higher % of data be validated (Box and Tiao, 1975). Summer months may understandably be warmer with a higher use of power. 3 years of monthly data provides 36 months to evaluate. If attempting to evaluate the data for one of the local power companies in TX (TXU Energy), 10% or 210,000 customers over 36 months (~7.5M observations) will be sufficient to create training, validation and test data sets. This may however, be too large a sample for initial observations but, may provide a relevant amount of information should the power company be asked to justify cutoffs during unusually hot periods before the state public utility commission.

Removing outliers, determining if data needs to be scaled and imputation are the next steps in data preparation.

Using the Box-and-Whiskers plot method would be too cumbersome for such a large number of observations. Random sampling can be used to extract a much smaller subset for the purposes of observation to determine if there could be outliers that could impact the quality of the analysis.

Scaling may be beneficial for income and energy use data points, especially when considering the impact of job loss on an individual's ability to successfully pay their power bill when it is in arrears. Both Python and R have scaling functions. If using an extremely small sample of data, scaling can be accomplished in Excel with mean, standard deviation and normalization calculations

Missing data should also be evaluated. There would be an assumption that data hygiene would be high at a public utility company. However, without strict data validation rules during the input or update of customer records it is possible that customer care representatives could miss data points when SLAs for time to resolve issues are taken into consideration. This would be data that is MNAR – missing not at random (outside research) so the observations shouldn't be deleted without further evaluation. Imputation is best in this case.

Two possible methods for evaluating the missing data points are KNN and logistic regression. If there is missing data that is significant (I would submit >5% of the sample population), then imputation based on the specific data point should be evaluated. For example, if there is missing payment data then this observation may need to be deleted since there may not be accurate methods to impute this data point. However, if there is missing data from the payment plans, it may simply be that the customer doesn't have a payment plan. In this case categorical variables can be assigned to represent 'no payment plan'.

Step 2
Classification of customers

In order to determine if a customer is a non-payer(delinquent), after data preparation the data will then be evaluated for classification using either KNN, logistic regression or random factor trees. The classification values will be defined as:

- Customers with accounts past due > 6 months (I am submitting that this is the threshold for determining if a customer intends to pay their bill)
- Customers with payment plans
- Customers based on recent income who are past due > 2 months who qualify for a payment plan or payment assistance programs
- Customers who do not have accounts past due > 6 months.

Customers who do not have accounts past due > 6 months, customers who qualify for payment assistance programs and customers on payment plans will be excluded from the observations for the remaining steps.

Step 3
Evaluating the cost of energy for customers included in the observations

Customers with past due accounts > 6 months are the target group for determining the cost to the power company of keeping their accounts active in the event that there are not enough technicians to shut off the power for all delinquent customers on a daily basis. In order to determine the prioritization (ranking) a linear regression model can be used.

Additional data such as average energy rates (considering seasonal fluctuations) will be needed for these calculations. The customer's past energy usage will also be used to predict the amount of energy that may be used in an upcoming month. This prioritization would be based on a range of 1 to n. A customer at 1 will be of the highest cost to the power company if not shut off.

Step 4
Evaluating the location of delinquent customers within the locations of technician territories

Assuming that technicians work in specific territories to optimize their travel time, a clustering model (K-means) can now be used to determine those clients where the delinquent clients are located. This is where the data that has now been ranked based on highest cost to the power company, location and the number of technicians available in a respective territory is utilized.

The clusters will show is there are more customers to shut off than technicians available in a territory. If there are clusters where there are more shutoffs in a day than there are technicians available a decision can be made the evening before to temporarily

reallocate technicians to high priority areas (keeping in mind that you don't want to have an area without any technicians in the event of an emergency).

Vehicle optimization tools for route planning of truck rolls to minimize waste in travel time, wear and tear on vehicles and fuel can also be used to maximize the number of daily shut offs.

Step 5
Exploring options for adding/reducing technician and reallocating assignments

Using Queuing simulations based on the delinquent clients, prioritization of those clients and the geographical clusters that they are assigned to, a queuing simulation can provide predictive modeling to determine how many technicians are needed for each cluster daily. The results of Queuing simulations based on historical capturing of the data can also provide the basis for leadership to make decisions on the need to permanently reallocate or hire additional technicians.