ISYE6501 Homework 8

Dylan Peters

July 11, 2017

Power company shutoff question

The assignment is to define data and analytical models to determine 1) what delinquent accounts to shut off and 2) how to optmize the shutoffs in a given month. I consider these as two separate questions, though the answers question one can inform question 2.

Who to shut off

The outcome of this decision is either a classification (leave on/shut off) or a probability that each delinquent account is not recoverable and should be shut off. A third class is customers who qualify for an assitance program. Here is some of the data we can use:

- Account balance
- Age of account (newer accounts are less likely to pay)
- Credit history of accountholder, if available
- Monthly usage
- Type of account: house/apartment/residential/commercial/etc.
- Whether they have already been sent a warning/disconnect notice
- Whether they have been disconnected in the past
- Months since last payment

I would be careful about using specific location data such as zip code or neighbodhood. There is the possibility of grouping people just by demographic factors such as race or nationality that may be illegal.

One initial question is whether this is in fact a supervised classification problem, or an unsupervised problem. Technically, there is no predetermined outcome to predict as a classification problem. Whether a customer is delinquent enough to have their service disconnected (basically expected to never pay) is more of a human decision. The decision is impacted by overall business profit and loss considerations. One way to use supervised learning is to take a dataset of past customer history including cutoffs and try to predict what a human would choose. This has the drawback of reinforcing previous biases.

In order to make a true classification decision, we can set as our target decision whether the customer made a payment in the current month. When used against the next month's data, the model can be used to predict whether each customer will make their next payment.

The model

We can use logistic regression to find a model that will give a probability that the next payment will be made. We should use Lasso regression to help find the best combination of variables. The final model will give a range of probabilities, with most being high probability of payment, several with low probability, and a few in the indeterminate range, e.g. 30-70%. In order to determine the best threshold to dictate a disconnect, we will need to use the probability with some more information.

The cutoff decision

Using the probabilities from our test data set, we can create a confusion matrix for a variety of general thresholds. To calculate the optimal threshold, we would need to assign a general cost to each quadrant:

- The cost of a true positive (payment predicted and made) is 0.
- The cost of a true negative (no payment predicted nor made) is the cost of service disconnect, which generally can be considered to be constant.
- The cost of a false positive (payment predicted but not made) is the cost of service for that customer.
- The cost of a false negative (payment not predicted but is made) is the cost of disconnect and reconnect, plus any lost service billing (and customer goodwill).

For the last two costs, we can either decide on an average customer cost to use, or use the actual cost for each customer. When we use the model in production, we likely want to use the actual customer cost. All other variables being equal, it is much more costly to keep a customer spending $350 a month who isn't paying than one spending $70 a month.

After determining that the customer is unlikely to make a payment, the next step is to contact them and give a warning notice about the disconnect. This may itself induce the customer to pay. As part of this process it may be determined that the customer is suitable for a program to help with their payments. For the rest, the customer is added to the disconnect list and passed to the next step.

For the next step we can convert the probability of payment to probability of delinquency by taking one minus the value.

Optimizing the disconnects

The output of the previous model is a set of customers who are unlikely to make their next payment and the probability that they will continue to be delinquent. We can combine this with their average monthly usage to determine a priority: monthly usage times expected delinquency. A user with $160 a month in usage and a 60% default probability is a higher priority than a customer with $90 a month in usage and a 90% default probability, because the expected cost is higher ($96 vs $81).

With the expected cost of each customer and their address, we can look at optimizing the routes for the service operators. Assuming we have good mapping software and traffic data, this becomes a network problem. Each vertex has a value–the expected reduction in cost for disconnecting the customer. Each edge has a cost–the time it takes for the worker to drive to the address and complete the disconnect. Each worker must also start and stop the work day at the service center. For each worker, the network algorithm can find a route that optimizes the value while keeping the cost (time) within the bounds of an 8-hour workday. After worker one is assigned the optimal route, those customers are removed from the network and a new optimal route is found with the remaining vertices. This continues until all workers are assigned a route. In order to avoid lopsided routes, an alternative algorithm would look at all the drivers at once and determine a balanced route for each one that is similar in value/cost.

Another factor is that we would assume that the workers are also performing connects and other service related tasks, so routing would need to take that into account.

Given a set of N disconnects each month and an average number M of disconnects performed by one worker each day, we can estimate how many workers are needed each month. We can use simulation software to find an optimal number of workers to ensure that the number of customers awaiting disconnects (or other service) does not grow continuously. We would need to know the distribution for both the number of service calls each month, and the number of service calls completed daily by one worker. If the workload is distributed evenly, then the worker daily calls would probably be a normal distribution, with for example a mean of 12 disconnects and a variance of 4. The monthly total disconnect count might similarly be normal, or might have a different distribution such as Poisson.

Conclusion and feedback

The models need to be updated on a regular basis, perhaps monthly. After each month, we can revisit the decisions and compare them to actual results. Did we miss any delinquent accounts that are still not paying? Did we make any cutoffs that were quickly reversed? In each case we can rebuild the model and adjust our threshold. We can calculate the costs saved for each successful cutoff and compare it to the costs incurred for missed cutoffs and reversed cutoffs.

We also need to track the routing algorithm to make sure that all cutoffs decided on are actually executed. There exists the possibility that a low-value customer will never rise to the level of an actual cutoff due to the cost of reaching that customer being higher than the cost related to their usage. However over time that usage will accumulate so the lack of cutoff becomes more and more costly in hindsight.