Homework #8 ISYE6501

Power Company Case Study

Zach Olivier

7/10/2018

Summary:

My analytics solution for the Power Company follows this general format:

Given:

- Customer Payment History, Family Composition, Age, Estimated Household Income,
- Estimated Home Value, average Energy Consumption, Energy Consumption Trends,
- Macro-Economic Factors, Weather, Energy Costs, Occupation, Seasonality

Use: flexible classification model - random forest, SVM, Gradient Boosting

To: identify customers who are most likely to never pay their energy bill

Given: Customer Location, probability to never pay (model output)

Use: supervised clustering algorithm

To: target shut-downs of high probability default customers in close areas (high value low cost clusters)

Recommendation:

The end result a recommendation each month to send the minimum amount of workers to the least costly, highest value geographic clusters to perform shut-offs. I would also recommend trying to preemptively solve the problem by pushing a message to high probability customers asking them to state their intentions of payment. Using this outreach and the model's classification probabilities - we could set up an experiment to optimize which customers would be most responsive to our outreach. Not only would sending a message to customers be less costly - the results from this experiment could also add additional information to help us decide which clusters, and how many technicians we should send on the shut down run.

Below are the components of my analytics solution - with notes aiming to explain the thought process of my choices.

Data:

The foundation of this solution is a model that can accurately classify customers who will never pay their energy bill. Without an accurate model the entire solution - however complex - will break down. The foundation of a highly accurate model is the data. I came up with a select few “stock” data points the energy company would most likely have on hand, plus a few exogenous factors that could explain a missed energy bill. I think a host of customer payment history coupled with variables that explain their ‘make-up’ as a customer will help accurately identify customers at risk. Younger customers with spotty payment history may be the most at risk to never pay, while families might be prone to weather issues but eventually pay off their debt. Large economic shocks could also effect certain groups to default completely on their payments.

My one concern is a lack of actual differentiation from customers who will never pay and customers who are late but eventually pay. To accurately classify customers we will need many examples of customers who default completely and customers who default for some period of time then pay their remaining balance. This can be tricky as we would have to examine the repayment distribution to determine when a “never pay” customer can be labeled as such. In theory a customer we label as a true default could pay their entire bill next month, data that we do not have.

## Model:

I see this problem as a “prediction” problem rather than an “inference” problem. Carefully choosing our variables up front and ensuring they do not violate any laws will allow us to fit highly complex and flexible models to our problem. This does not come without drawbacks - gaining trust in the solution will be more difficult. Explaining why the predictions are the way they are will not be easy.

Overall, I think fitting a Classification Tree model and drawing quick insights (view the decision tree splits and classification scores per leaf) will open up the ability to model using a Random Forest Model (cannot view the entire tree - but can bridge the gap with the business that this model is a combination of many trees).

Both should work well to fitting our data if there is true differentiation between classes of non-payment customers.  Both tree based models can output probabilities of non-payment in the same way a logistic regression would. Using probabilities as opposed to hard classification would allow us to custom fit the company’s goals based on their available budget. By toggling our threshold of classification we can choose to target shut-offs for customers who are “sure things to not pay” ($p > .9$) or choose a lower threshold if there is a lot of resources developed for this problem.

## Experiment Design:

Having the probabilities from the model will also allow us to recommend getting ahead of this problem through outreach. We can set up an experiment by emailing customers across all ranges of non-payment probability and see which ones are the most responsive. Ideally, a section of our customers will be reminded of their payment, and based on their situation, pay the minimum fee to keep the power on, or let us know they will eventually pay. Doing this will not only help us hone our actual shut-offs, but we can also take what we learn from the experiment - like the type of people who will pay with a reminder - and add it back as a factor in our original model. This will help reduce shut-offs “up-front” but also help us iteratively improve on our model by “investing” in data that will better guide us towards customers have a physical shut-down.

### Clustering:

Assuming we have a good model that is highly accurate - I would develop a clustering solution to group customers with high probability to default forever, distance between customer addresses, geography (census tract), and cost to physically shut off that customer’s energy supply (drive time and repair time cost estimates). Each centroid of the cluster can give us an highly interpretable “value index” of where to target our physical shut-downs. Results from our experiment can then remove some address from these clusters cutting down repair time. We can also “size” our prospective clusters based on our classification probability and cost. It may be that turning power off for customer’s who will pay is more costly than the alternative. Our clusters can then be determined by only customers with extremely high probability to default completely.

Powerful interpretable explanations can be developed to motivate the business i.e. target this cluster because it has with high probability to default, short drive time, and densely grouped customer addresses (low inter-drive time from customer to customer). We can expect targeting this cluster to cost x amount and shut-down a high percent of true total defaults.

My concern with this solution is if the data is sparse or does not cluster together well. Unsupervised clustering may not give us well-formed clusters. Supervised clustering may be a better solution to group customers based on their 3-dimensional distance of default probability, location, and cost.

### Optimization: (why not?)

In my experience optimization is very difficult to develop and also difficult eventually sell as a solution. Especially so when the optimization model tries to model the entire system of possible variables and constraints. Explaining the reasons for an optimization model’s solutions is on par with explaining a complex “black box” model to business units - sometimes there are simply too many variables to account for, and the cost to optimizing the solution for all cases could quickly exceed available resources.

I do believe in the power of optimization to handle components of our problem. If we can reduce the entire solution down into how to optimize the shut-downs for a specific cluster - we would have a much better chance at connecting an optimization model to this problem. Given a cluster we could also use mature optimization algorithm's (shortest path) to find the best way to perform the shut-down for a specific cluster. This would be my approach as a follow up phase to my proposed solution. We could then recommend the number of workers, the route they take (based on cost), the shut-downs they complete in order to most efficiently target one or more of our "high probability to default - low overall cost" customer clusters.