Thomas Huelsnitz
ISYE 6501
Week 8 Homework

Question 18.1

Describe analytics models and data that could be used to make good recommendations to the power company. Here are some questions to consider:

• The bottom-line question is which shutoffs should be done each month, given the capacity constraints. One consideration is that some of the capacity – the workers' time – is taken up by travel, so maybe the shutoffs can be scheduled in a way that increases the number of them that can be done.

• Not every shutoff is equal. Some shutoffs shouldn't be done at all, because if the power is left on, those people are likely to pay the bill eventually. How can you identify which shutoffs should or shouldn't be done? And among the ones to shut off, how should they be prioritized?

Think about the problem and your approach. Then talk about it with other learners, and share and combine your ideas. And then, put your approaches up on the discussion forum, and give feedback and suggestions to each other.

You can use the {given, use, to} format to guide the discussions: Given {data}, use {model} to {result}.

Have fun! Taking a real problem, and thinking through the modeling and data process to build a good solution framework, is my favorite part of analytics.

The first step in forming a solution to this problem is determining which customers we expect to pay their bills and which customers will not. I would start by gathering data that would seem to be indicative of how likely a customer is to pay.

Data Gathering

Below are some factors I would consider along with some hypotheses that I would test in an exploratory data analysis.

- **Customer payment history**
  This is something that the power company would certainly have access to. New customers would have little to no payment history. Customers with extensive payment history are people who pay their bills – otherwise they would have no service (it seems safe to assume that, even in the absence of a sophisticated predictive model, the power company would have some existing procedure for cutting power off for customers who have not paid their bill). We will assume that new customers and customers with a history of late payment are more likely to not pay their bill than long-standing customers who pay their bill on time.

- **Customer credit score**
  According to the Federal Trade Commission website (https://www.consumer.ftc.gov/articles/0220-utility-services#What%20Should), this is information that the power company would have access to. A credit score

is intended to be a measure of someone's credit-worthiness. Someone with a poor credit score is someone who pays bills late or not at all and is likely someone that would pay a utility bill late/not at all. I do wonder if utility bills may be unique as compared to other debts – I can imagine that the utility bill would be one of, if not the highest priority bill for most people to pay. I can imagine that there are many people with bad credit scores that reliably pay their utility bills.

- **Customer rents or owns their residence**
  Owning a residence can be indicative of higher financial stability.

- **How long customer has lived at current residence**
  Amount of time that the customer has rented/owned their residence probably correlates with their financial stability

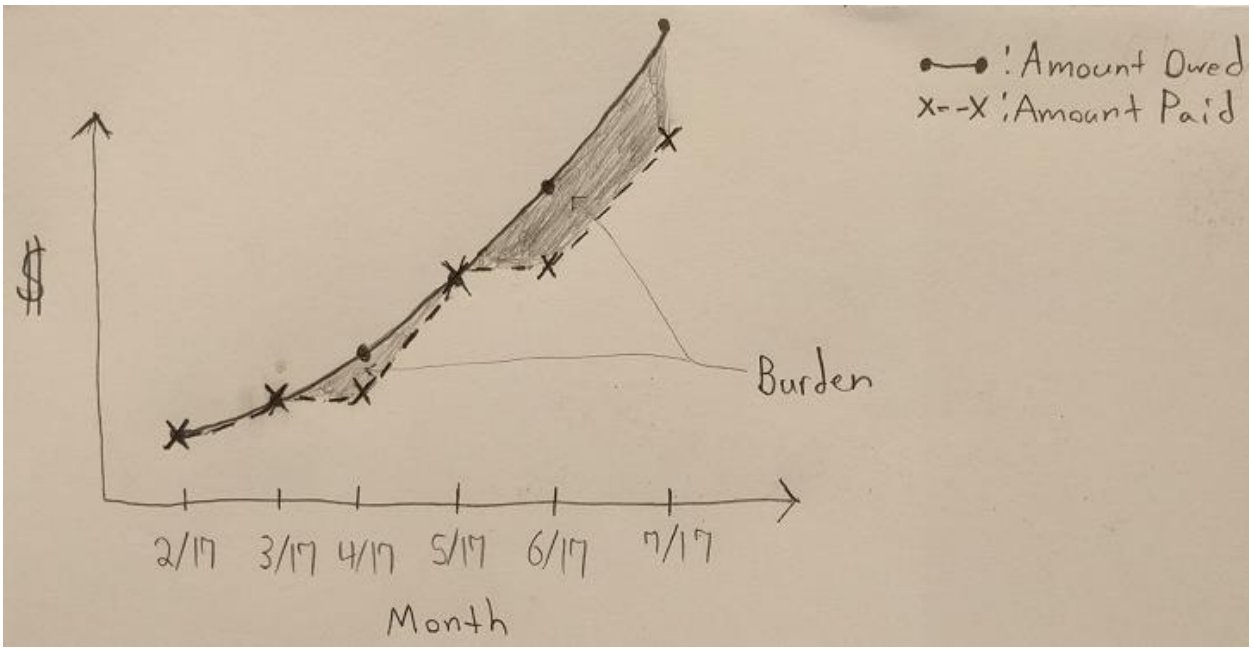
- **Customer is married or single**
  Married households will most likely be more stable than households that are not married.

- **Count of people in household**
  Larger households where the residence is rented or the customer is single will likely be less financially stable than smaller households. Household size likely has less of an effect for customers that are married.

Building Model for Identifying Non-paying Costumers

As mentioned above, a customer's payment history is an important piece of predicting whether they are likely to pay an outstanding bill or not. Here is one way of building a metric for customer payment history. Below is an example plot where the x-axis is the time period from February of 2017 to July of 2017 and the y-axis is dollars. Here we are comparing a plot of the cumulative amount of money owed over time to a plot of the cumulative amount paid.

Thomas Huelsnitz
ISYE 6501
Week 8 Homework



Let's call the area between the curve of money paid and money owed the burden which the customer is imposing on the power company.

The equation for a customer's burden at the current time, t_b, would be

$$\int_{t_a}^{t_b} O(t) - P(t)\, dt$$

where O(t) is the cumulative amount owed as a function of time, P(t) is the cumulative amount paid and t_a is the date that the customer's service started.

By working with the power company to learn more about their financials and looking at historical data we would determine a threshold value for burden, beyond which we expect that the customer is not going to pay their bill.

Once this threshold is determined, I would look at all the available historical data and divide it up into a data set with one entry for each customer for each billing cycle. Each of these entries would have the corresponding burden for that customer at that time and a binary variable indicating if the customer is above or below the burden threshold. Below is a snapshot of what this data might look like:

customer_id	billing_cycle	service_start_date	credit_score	own	months_at_residence	married	num_in_household	burden	over_burden
312415	5/1/2018	1/1/2012	700	0	36	1	3	0	0
312415	6/1/2018	1/1/2012	700	0	37	1	3	0	0
182749	5/1/2018	4/1/1992	780	1	120	1	5	0	0
182749	6/1/2018	4/1/1992	780	1	121	1	5	0	0
192837	5/1/2018	2/6/2018	440	0	3	0	5	130	1
192837	6/1/2018	2/6/2018	440	0	4	0	5	180	1
172638	5/1/2018	2/2/2000	720	1	100	0	2	70	0
172638	6/1/2018	2/2/2000	720	1	101	0	2	0	0

This historical data would give us testing and training data. I find it tempting to try to fit a regression model to this data where burden would be the response and credit_score, own, months_at_residence, married, num_in_household would be the input factors. In the lectures for this week, Dr. Sokol mentions that this would not be a good approach in this case because the burden factor will be dominated by instances of 0 (customers who always pay their bill on time.) He recommends that the classification and regression be done separately – specifically, we should try to classify who is likely to be over_burden and who is not. I would try a Support Vector Machine or Gradient Boosted Tree with various combinations of hyperparameters and choose whichever performs the best. I would do an 80/20 split for training/validation and testing. I would use k-fold cross validation on the training/validation portion of data to compare the effectiveness of the various SVM and Gradient Boosted Tree models with various hyperparameters. The best of these models would then be tested on the test set to get a more accurate estimate of the model's performance on real data. Many implementations of SVM and Gradient Boosted Trees allow you to also return the estimated probability of the assigned class. I would want to return these probabilities because they are useful in the subsequent step of dispatching power company resources for cutting off power.

Building Model for Estimating Cost of Non-paying Customers

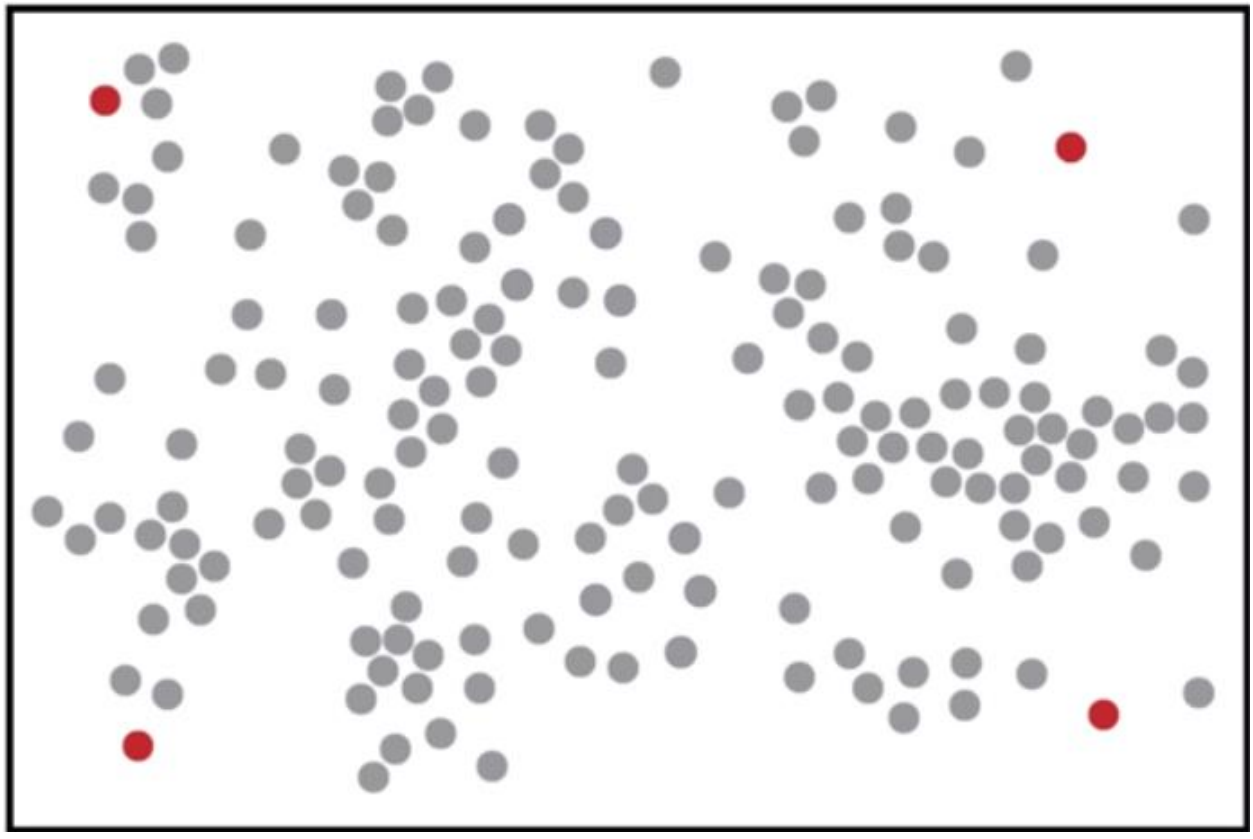
I would build a regression model using the same historical data as before, but this time I would only use the subset of customers which have at least one billing cycle where over_burdened is 1 (true). I would choose to subset the data, because at this stage of our analytic, we are focused only on customers who we predict to be non-payers, so this is the same subset of customer classification that we should focus on for our training data. This also addresses the issue with the hybrid classification/regression approach described in the prior section.

As before, I would divide this data into train/validate and test data, cross validate and choose the best model and then get a more accurate estimate of the model performance on test data. I would try simple ordinary least squares (OLS) regression for this portion and do some feature engineering to extract useful information from the existing data. For instance, I would make a new column called month_num which is just the number of the month of billing_cycle. This is interesting because different months have different power consumption habits and this can affect the expected burden of the customer.

One issue I see here is a lack of continuous factors for accurately predicting the burden. If the fits from the regression model are not adequate, then one possible solution would be to bucketize the burden values in our historical data by looking at the distribution of the values and choosing thresholds to define 3 different buckets indicating low burden, medium burden and high burden. Now we have converted the regression problem into a classification problem and we have a bit more flexibility in our model selection. I would then try to use SVM or Gradient Boosted Trees as a predictive model for these responses. Moving forward, I will assume that the original regression model gave satisfactory results.

Dispatching Power Company Resources

The models preceding here would give the result which is the starting point of the last lecture for this lesson. Specifically, the following map:



Which shows the locations of paying customers (grey) and customers who are over-burdened and likely will not pay their bill (red). Additionally, for each red point, we have a corresponding expected increase in burden from the prior billing period to the next one. This explains the benefit in shutting off power for that customer – the higher the expected delta in burden, then the more the company saves by shutting off power at that location . We must create a model which chooses the optimum set of locations to cut off power subject to costs due to driving

distance (labor costs per time driven and fuel costs due to distance) and the number of power company employees we have to dispatch.

To choose the optimum dispatch pattern, I would use a modified k-means algorithm where k corresponds to the number of power company employees we are able to dispatch. Below is the modified k-means equation which accounts for the benefit in cutting off power for points assigned to clusters and approximately accounts for money spent traveling within each cluster of points. This second part is approximate because the sum of the distances from a cluster center to each of the points belonging to the cluster is not the same as the distance that would be covered when moving along a path that hits all the points in a cluster.

Thomas Huelsnitz
ISYE 6501
Week 8 Homework

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} y_{ij} \left(\| (x_{ji} - c_{ji}) \alpha \|^2 - B_i^2 \right)$$

$$\text{s.t.} \sum_{j=1}^{K} y_{ij} = 1$$

Where $y_{ij} = \begin{cases} 1 \text{ if point } i \text{ in cluster } j \\ 0 \text{ otherwise} \end{cases}$

c_{ji} : Coordinates, d_i, of cluster j

x_{ji} : coordinates, d_i, of point i

α : Conversion factor for miles $\rightarrow$ \$ $\left[\frac{\$}{\text{mile}}\right]$

B_i : Benefit to cutting off power at point i [\$]

Minimize J in using the K-means algorithm.

1. Seed with K randomly chosen cluster centers. (c_{ji})
2. Construct y_{ij} by assigning points to their nearest cluster center.
3. Calculate J
4. Calculate new c_{ji} as the centroids of the new clusters.
5. Construct y_{ij}
6. Calculate J. Compare to prior. If significant change, repeat 4-6. If not, stop.

- Once converged, y_{ij} tells which points belong to which clusters.
- I would run this many times and then choose the y_{ij} which belongs to the lowest J-value.

(power company employees)

B_i, the benefit to cutting power off at point i, would be the probability of no payment for that point (as determined in the first model) multiplied by the expected burden delta as determined by the regression model.

At this point, the matrix y_{ij} would describe which points belong to which of the k clusters. In the context of the problem, this describes which customers are having their power turned off by which power company employees and we have completed the model requested by the power company.