Assignment 12

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.

Answer 18.1

Step 1: Exploratory Data Analysis

In this section we will try to get more familiar with the data, data types, distributions, correlations, outliers etc. Often through visualization we can see some trends and get an overall sense of the data. Now, let us investigate what kind of data we might have for this work and how we can use it do our analysis.

Step 2: Creating a Model which can confidently segregate the defaulters and non-defaulters.

• Data Given:

a. Household Income-Numerical Data
b. Credit Score-Numerical Data
c. Property Owned/Rented-Categorical Data (Yes/No)
d. Bill Amount-Numerical Data
e. Years with the company: Numerical Data
f. Past Defaulters- Categorical Data (Yes/No) (People who never paid the past dues at all.
g. Late Payments-Number of times a client has missed a payment (Numerical).

• Use: Logistic Regression Model(s), a probability distribution model will be good as through various iteration we can adjust the threshold and see how well we can classify the data between defaulters and non-defaulters. Adjusting the threshold can give us the cushion of being more cautious and move any person into the defaulter category for further investigation.

• To: If we are satisfied with our model, we can get a good idea of the subset of number of potential defaulters. This will help us to answer the next question how we can prioritize the job of shutting off the power of the potential defaulters based on the available resources.

Step 3: Creating buckets of priority based on the output of the model created in step 2.

• Given: Amount owed by customer with probability outputs from Logistic model • Use: Decision Tree to create the priority buckets • To: Create power shutoff priority ranking for each customer

Now we have a subset of data containing the potential defaulters which we can further categorize into three group of based on the amount owed as high, medium, and low priority.

More scientific would be to use software to create a decision tree on the logistic model output along with the other data collected and create a regression model at each node or leaf. For the leaves that correspond with customers of a lower confidence level we would add a penalty or subtract away expected money owed to emphasize this point. A single decision tree is still easily explainable to a general listener, however depending on the sample size we may end up overfitting the data.

Step 4: This step helps in prioritizing our resources to execute the task of shutting of the power of the defaulting customers.

• Given: The amount owed by customer and the mean/estimated drive time to each other customer. • Use: Simulation • To: Determine which groups of customers can be reached in the month that will maximize the estimated savings to the power company.

Although it is illegal to use address when computing this priority list, the power company does have the address of all their customers. Overall data science is also an art and along with a science. It might take many iterations to get the best results. This should help us create an optimal model to minimize the overall loss to the company.