Question 18.1

To address the question, we need to consider several aspects in order. First, we need to classify homes based on their payment behavior, distinguishing between those that pay on time, those that are likely to pay but missed the deadline, and those that are unlikely to pay.

Once classification is complete, we can develop a predictive model to assess the benefits and costs of shutoffs, taking into account classification and classification errors.

Finally, we can optimize the shutoff process while considering the company's constraints after completing the prior analyses.

First Analysis: The Home Classification Analysis

For this case, a significant amount of data is required to make accurate predictions and classifications. The topic involves several aspects, such as bill payments, household identification, power usage, and other relevant information. Therefore, the first step should be creating a database containing all the necessary data if it doesn't exist. Once we have the database, we can determine the key factors required for the classification model:

- Home IDs to identify the customers
- Home addresses to analyze geographic patterns
- Number of people living in each home
- Credit scores of the bill payers to predict their ability to pay bills
- Previous classifications indicating likelihood to pay or not
- Cumulative values tracking the number of months paid and not paid
- Time since last bill payment as a potential indicator of a bill being unpaid
- Energy usage history to prioritize shutting off power for households with high energy consumption

These are some potential features we could use for our classification model. It is important to consider the data in distinct parts. We can assume that some homes are consistently on-time payers, some may pay but often be late, and some may never pay. For our analysis, we can focus on the last two groups because they are the most relevant to the company’s issues. Using the features mentioned earlier, such as home IDs, addresses, number of occupants, credit scores, previous classifications, and energy usage history, we can train a logistic regression model to predict the likelihood of a household being late or not paying their bills.

In addition, change detection can be useful in detecting changes in the cumulative values of the number of months bills have been paid and not been paid. We can use CUSUM to identify any changes, indicating that a previously reliable household may now be likely to miss bill payments or vice versa. This can help us assign households to the appropriate group or remove them from one.

To summarize, our end goal is to classify households into two categories–likely to pay but might be late and likely to never pay. Logistic regression is a suitable model for this classification as it can give us the probabilities of a data point being in a group alongside its classification. To handle missing data, we can use various imputation methods. Additionally, cross-validation can be used to select the best model, which can then be trained on all the available data.

In the desired format:

Given

- A comprehensive database containing various customer data

Use

- Data imputation methods
- CUSUM
- Logistic regression
- Cross-validation

To

- Properly fill in missing data
- To detect changes that could help with customer classification into a group
- To predict customer classes
- To help choose the best model

Second Analysis: The Shutoff Cost-Benefit Analysis

Performing a cost-benefit analysis requires consideration of several critical factors, and it necessitates the use of various models. To determine the appropriate classification threshold for our logistic regression, we must take into account the cost of not shutting off power earlier and the cost of misclassifying a customer. We can begin by examining a customer's monthly energy usage and creating a forecasting model, such as ARIMA or exponential smoothing, to predict their energy usage and costs. This will allow us to prioritize shutting off the power for customers who consume a lot of energy and have not paid their bills for an extended period. It also lets us see if the person is not likely to pay their bills in the future. Similarly, conditional probability is a vital consideration as well. If a customer has not paid for a long time, their probability of not paying for the next billing cycle is also likely to be high. Formalizing this into a probability distribution can provide a more precise estimate of the likelihood that they will not pay their current bill given their history of unpaid bills. Furthermore, it is essential to evaluate the present value of future unpaid bills (as is taught in MGT 8803) to gain a better understanding of the potential financial loss incurred by not shutting off power for non-paying households. It is also critical to analyze the costs involved in shutting off power for customers, such as worker travel time, equipment expenses, and potential productivity loss. To determine the appropriate threshold for the logistic regression model, all of these factors must be considered in tandem, balancing the costs and benefits of accurate classification and misclassification while adhering to the company's constraints.

In the desired format:

Given

- Our logistic regression model
- Customer energy usage data
- Shut off cost data

Use

- ARIMA or exponential smoothing
- Present value discounting
- Probability distributions

To

- Determine an appropriate threshold for logistic regression classification
- Determine future customer payment and energy usage forecasts
- Determine the present value of money to be paid in the future
    - Identify customers that the company would lose a lot of value on if they did not pay (this can be done in conjunction with the forecasting step)

Third Analysis: The Optimization Analysis

To optimize the process, we need to consider several constraints. Some are: travel time, the costs involved in shutting off power, and worker productivity. Geographic factors must also be taken into account, such as distance between households and the concentration of non-paying customers in a specific area. For instance, is it worthwhile to send a worker a very long distance for one non-paying household or would it be better to send workers to a more concentrated area of non-paying households? An optimal route should also be produced by taking into consideration traffic data and other relevant factors. The company could potentially use decision trees for this process. Beyond that, the number of workers available to the company for this task should also be considered. Another key factor to consider is the immediate impact of shutting off power for a household. If doing so is likely to put a household in harm's way and the harm cannot be prevented, then perhaps we should add another constraint ensuring that we do not shut off power for them. This constraint complies with the ethics behind shutting off power and the safety issues associated with it. Finally, when it comes to actually optimizing, we should try and optimize for cost. In addition, it would be worthwhile to use stochastic optimization since it would better replicate some of the randomness associated with traffic and the other factors we are considering. Overall, it should provide us with the best utilization of our resources.

In the desired format:

Given

- Traffic data
- Customer geographic data
- Company constraints
- Potential risk of harm to households due to shutting off their power

Use

- Decision trees
- Stochastic optimization

To

- Determine the best use of the company’s resources in a way that minimizes the costs associated with shutting off the energy for non-paying customers (while also complying with company constraints)
- Determine the best travel routes the employees can take to minimize time on the road