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.

Solution:

Given customer payment history data, use logistic regression models to predict likelihood of payment

Given the customer payment history data, financial data, and geographic data, we can use machine learning models to predict the likelihood that each non-paying customer will eventually pay their overdue balance. Specifically, we can train classification models like logistic regression, random forests, or neural networks to estimate the probability that a customer will pay based on features like:

- Number of missed payments
- Days past due
- Credit score
- Income level based on zip code demographics
- Account balance
- Past disconnections and reconnections

The models can be trained on historical examples of customers who did and did not end up paying after being shut off. The probabilities output by the models for current non-paying customers can help prioritize cutoffs - those with very low probability of payment should be prioritized for disconnection.

For the classification models, logistic regression is a good choice due to its interpretability and ease of implementation. The logistic regression model estimates the probability of a customer paying their overdue balance based on customer attributes:
P(Y=1 | X1, X2, ..., Xn) = 1 / (1 + e^-(B0 + B1X1 + B2X2 + ... + BnXn))
Where Y is the binary outcome of customer paying (1) or not (0), X1 to Xn are the input features like credit score, income level, past payments, etc. and B0 to Bn are the coefficient weights estimated by the model. Features with positive coefficients increase the estimated probability of payment.

For example, a positive coefficient for the credit score feature would increase the estimated probability of payment for customers with higher credit scores. Interactions between features can also be included, so income level and past missed payments could interact to provide better probability estimates.

Given customer and calendar data, use feature engineering to capture seasonal trends in payment

In addition to the core customer data, we could engineer calendar variables to capture seasonal trends. For example, a binary feature for winter months could indicate higher heating bills leading to greater difficulty paying. Interactions between calendar variables and features like income level may also prove insightful.

For instance, high income customers may be less impacted by seasonality compared to low-income customers. The winter indicator variable could be interacted with income level to account for this relationship in the model.

Given predicted payment probabilities, use optimization algorithms to schedule disconnections

In addition, we can use optimization algorithms like linear programming or constraint satisfaction to solve the operational problem of scheduling the limited number of disconnections each month. The optimization would take into account the predicted probability of payment for each customer as well as operational constraints like geography, capacity, and worker schedules. The result would be an optimal plan for which customers to disconnect each day and week to maximize the number of likely non-payers shut off.

For the optimization model, we can use an integer programming formulation to determine the optimal set of disconnections for each period. The objective function aims to maximize expected revenue based on the predicted probabilities of payment. Constraints include worker capacity, geographic proximity of customers, and operational costs.

For example, optimization may determine it is most efficient to group disconnection orders by geography to minimize worker travel time between orders. The output would be an optimal plan specifying which customers to disconnect each day or week to maximize shutting off likely non-payers given the constraints.

Given disjointed customer data systems, use data integration pipelines

To implement these models, key steps involve data collection, cleaning, feature engineering, and model validation. The power company likely has customer data in multiple systems that need to be integrated. Domain expertise in the collections process is essential for constructing useful features and validating model performance. Monitoring changing customer behavior patterns over time is also critical to keep the models accurate.

There are also important ethical considerations around bias, transparency, and fairness when using machine learning models for societally impactful decisions like service disconnections. But overall, combining rigorous predictive modeling with optimization provides a powerful framework for efficiently managing power collections and balancing the needs of the business with those of customers struggling to pay their bills.

For example, past payment data may come from the billing system, while credit data comes from the customer management system. Cleaning and mapping are needed to integrate the disparate data sources into a single analytical dataset.

Given disconnection optimization models, use ethical checks to ensure fairness

While data can inform and optimize disconnection decisions, important ethical considerations around bias, transparency, and fairness should be incorporated. For example, human review of the model scores could identify recommendations that seem biased against certain customer groups. Also, the models may unfairly penalize customers going through temporary hardships.

Auditing the data and providing recourse for negatively impacted customers can help provide accountability and transparency around the system. No machine learning model is perfect, so techniques to monitor for and mitigate unintended consequences are critical.

Conclusion

This combined approach of analytical tools for prediction and optimization for planning provides a data-driven way to determine disconnections. We balance the need to increase revenue through disconnections with the goal of avoiding cutting off customers who are likely to pay if given more time. The models can be continuously improved by adding new data and re-training such as considering probabilities where threats of shutoff may encourage customers to pay. And the operations can be optimized on an ongoing basis.