Team 1: Mallika Andukuri, Raisa Chowdhury, Lauren Earle, Sanjay Kasi, Yuchen Zhang

ECON 7720: Kaggle Contest - Banking Recovery

Context and Questions

Even after a debt has been legally declared "uncollectible", the bank associated with the debt still wants to collect as much of the debt they are owed as possible. The bank will implement different recovery strategies at different thresholds of expected recovery amount. The higher the expected figure, the higher the recovery strategy it employs. The trade off, however, is the costs associated with different recovery strategies. The higher the recovery strategy, the higher the costs incurred by the bank. Each subsequent recovery strategy level costs an additional $50 per customer.

We have been tasked with building a model to analyze the cost-benefit relationship of the Level 1 recovery strategy in order to answer the questions below. The industry relevant to this project is the banking industry and the relevant stakeholders are the bank, its shareholders, employees, and customers, in particular those who owe the bank a debt.

The questions we are trying to answer are:

- Does the extra amount that is recovered at the higher strategy level exceed the extra $50 in costs?
- What banking recovery strategy should the bank adopt, e.g., is it worthy to pursue Level 1 over. Level 0?

These are useful questions to answer because understanding the cost-benefit relationship will allow the bank to adopt a recovery strategy that maximizes the amount recovered from debts. This is crucial for financial stability and sustainability. They are also useful to answer because the bank has limited resources and determining the most effective recovery strategy helps them allocate resources optimally.

Data and Descriptive Statistics

The data set includes 6 variables, which are named and described in detail below. It also includes 1882 observations. The data is adequate as it contains the necessary variables, actual recovery amount and recovery strategy, to address the questions we are trying to answer. The target of interest is the actual recovery amount and the predictor that we plan to analyze is recovery strategy. The confounding variables are expected recovery amount, age, and sex.

The size of the data set is sufficient for calculating relevant descriptive statistics, such as averages and standard deviations. The representativeness of the data set is also sufficient as diverse cases are included, allowing for comprehensive analysis. There are no missing values that could impact the analysis.

Name, description, and type for each variable:

Variable Name	Description	Type
id	Unique identifier for the customer account	Numeric
expected_recovery_amount	Estimated amount the bank expects to recover	Numeric
actual_recovery_amount	Actual amount the bank recovered	Numeric
recovery_strategy	The level of recovery strategy employed to recover the debt	Categorical
age	Age of the customer	Numeric

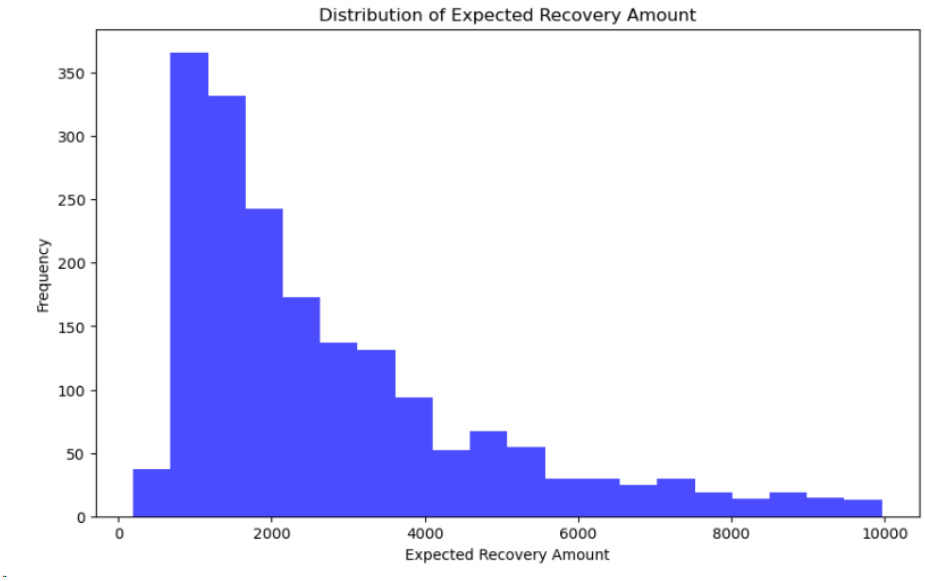
sex	Gender of the customer	Categorical

Relevant descriptive statistics for each variable:

- **ID:** The id variable has a count of 1882, meaning there are 1882 entries. The IDs range from 6 to 2056, which suggests that the dataset may have been sampled from a larger database or that the IDs are non-sequential.
- **Expected Recovery Amount:** On average, the bank expects to recover approximately $2,759.97 per account, but this amount varies, with a standard deviation of approximately $2,019.83. This indicates that there's a wide range of expected recovery amounts. The smallest expected recovery amount is $194, and the largest is $9,964.
- **Actual Recovery Amount:** The actual amount recovered is, on average, higher than expected at $4,000.97, and the variation is considerable (standard deviation of $4,576.51). The highest amount recovered is significantly larger than the average at $34,398.48, which could indicate outliers or particularly successful recovery cases. The lowest amount recovered is $200.43.
- **Recovery Strategy:** In the data set, there are 247 customers that fell under Level 0, 670 under Level 1, 333 under Level 2, 368 under Level 3, and 264 under Level 4.
- **Age:** The average customer age in the data set is approximately 39.65 years, with a standard deviation of 15.45 years. The youngest age is 18, and the largest is 84.
- **Sex:** In the data set, there are 973, or 51.70%, male customers and 909, or 48.30%, female customers.

Descriptive Visualizations

Distribution of Expected Recovery Amount:

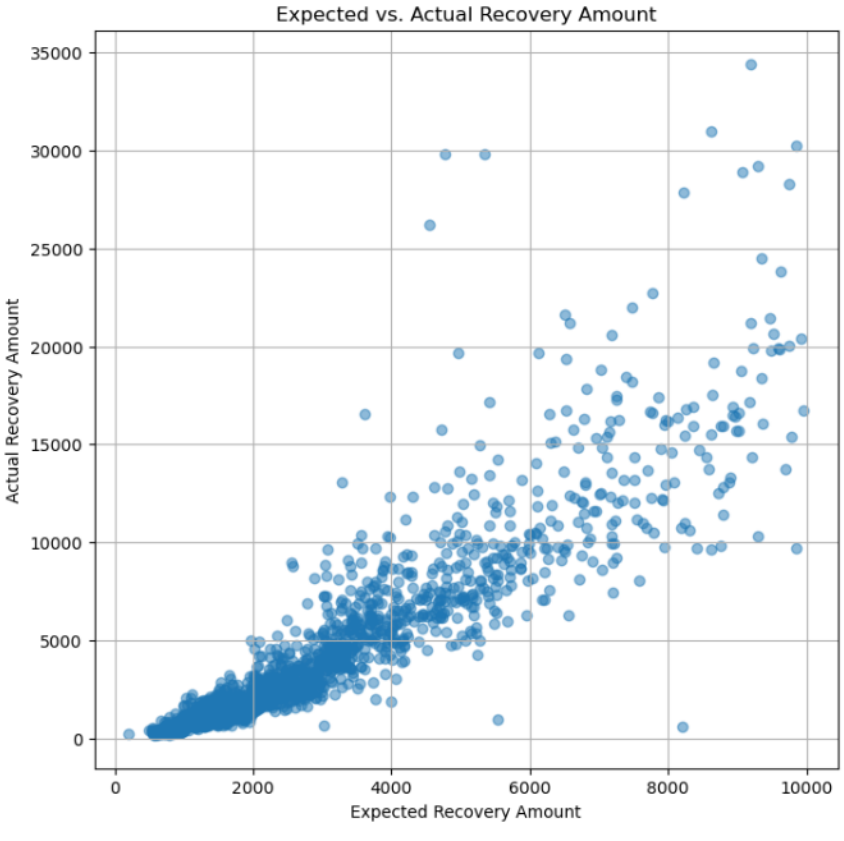


The histogram depicts the distribution of the expected recovery amounts from the dataset. It is evident that the distribution is right-skewed, meaning that the majority of the expected recovery amounts fall at

the lower end of the scale. There's a high frequency of accounts with smaller expected recovery amounts, as indicated by the tall bars on the left side of the histogram. The frequency decreases as the expected recovery amount increases, showing that there are fewer accounts with higher expected recovery amounts. This trend continues towards the right tail, which drops off, suggesting that very high recovery amounts are rare.

The right-skewness implies that the bank deals with a larger volume of accounts with smaller expected recovery amounts, which could be less profitable to pursue aggressively due to the costs associated with the recovery process. Accounts that fall into the higher expected recovery amount brackets are less frequent, but may warrant a more aggressive recovery strategy. This initial analysis sets the stage for a more granular investigation into whether the additional efforts and costs invested in higher expected recovery amounts are justified by the actual amounts recovered.

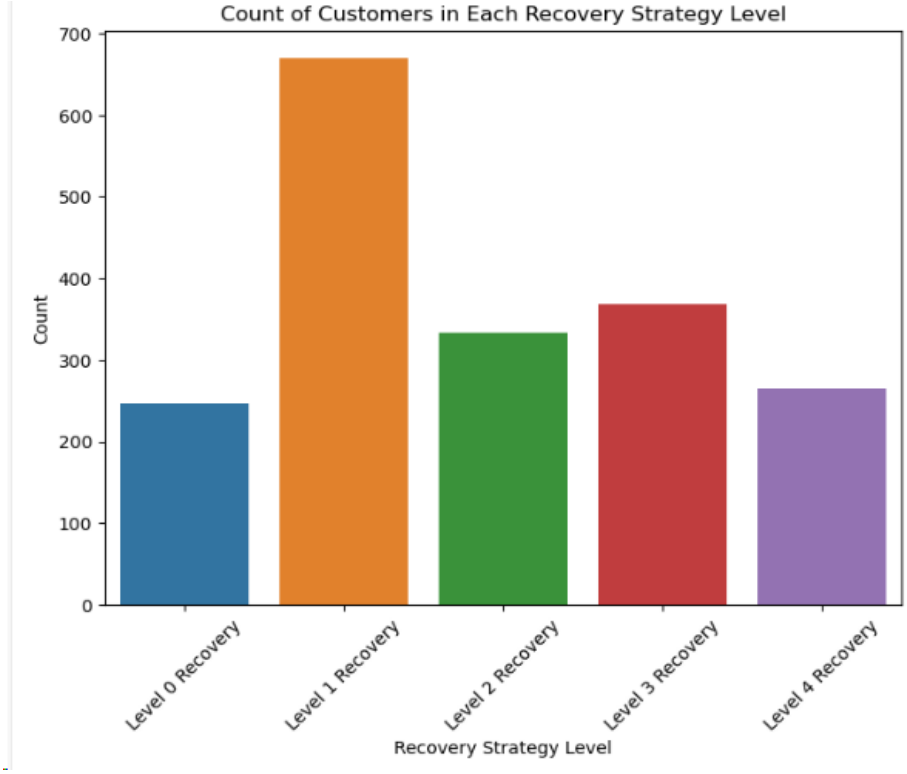
Expected vs Actual Recovery Amount:



In this scatterplot, we see the bank's estimated recovery amounts on the horizontal axis and the actual recovered amounts on the vertical axis. The dots scattered across the plot represent individual debts. Ideally, if the bank's predictions were perfect, all dots would align along a diagonal from the bottom left to the top right. However, the spread of the dots indicates room for improvement in the bank's predictive model.

Points above this diagonal line show where the bank has been conservative in their estimates - they actually collected more than they expected and might consider a more optimistic model. Conversely, points below this line are instances where the bank has been too optimistic, suggesting a need to understand the factors that led to lower recoveries than expected. The overall variability, especially noticeable in the middle section of the plot, suggests that incorporating more nuanced data or reassessing the factors that influence recovery could lead to more accurate predictions. For the most significant deviations, the bank should delve into individual accounts to fine-tune their predictive approach.

Count of customers in Each Recovery Strategy Level:



This bar chart provides a clear visual breakdown of the number of customers engaged at each recovery strategy level. Level 0, which is the least intensive effort, has the least number of customers, suggesting that the bank employs more effort for a significant number of cases. The most noticeable aspect is that Level 1 Recovery is the most populated category, with the highest count of customers. Levels 2 and 3 show a reduction in customer counts, which indicates that as the intensity of the recovery strategy increases, fewer customers are targeted, likely due to the higher costs involved.

Level 4, which we can assume to be the most aggressive and expensive strategy, understandably has a low number of customers comparatively. The pattern here suggests a strategic approach by the bank, escalating their recovery efforts to a point as the expected recovery amounts increase, but not beyond the point where costs also increase to an unacceptable level. This visualization is key to understanding how the bank allocates resources across different levels of debt recovery efforts.

Model

The two methods used to investigate the causal effects of recovery strategy were Regression Discontinuity Design (RDD) and Propensity Score Matching (PSM). RDD is a quasi-experimental pretest-posttest design that aims to determine the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. PSM is a statistical method that matches individuals based on their propensity to receive the intervention, thus controlling for confounding factors and allowing for a causal estimate of the intervention's effect.

In the RDD approach, we established the independence of sex and age from the actual recovery amount and subsequently examined the impact of the elevated recovery strategy. We employed statistical tests such as the Kruskal-Wallis H-test to ensure similarity in age distribution above and below the $1,000 expected recovery amount threshold. The Chi-square test confirms the independence of sex from recovery strategy. Additionally, a linear regression, focusing on the comparison between Level 1 and Level 0 recovery strategies concerning the $1,000 threshold of expected recovery amount, is used. The model was applied to filtered entries around the $1,000 threshold, distinguishing between expected recovery amounts above and below. Utilizing Ordinary Least Squares, the model provided insights into the additional recovered amount linked to the higher recovery strategy. We also used a different threshold to further evaluate the efficacy of the higher recovery strategy.

PSM, on the other hand, utilized logistic regression to calculate propensity scores, ensuring balance in covariates between recovery strategy levels. Model evaluation relied on metrics such as the estimated Average Treatment Effect (ATE) and relevant statistical tests. We used a logistic regression model, incorporating covariates such as age, sex, and expected recovery amount, to estimate propensity scores, ensuring a balanced distribution of participant characteristics between Level-0 and Level-1 recovery strategy groups. To prevent data leakage, propensity scores were calculated only for participants with scores between 0 and 1. The positivity assumption, crucial for causal inference, was strictly maintained by excluding participants with propensity scores of 0 or 1, ensuring overlapping distributions between treated and untreated groups. Weights, derived from propensity scores, were then used to create populations representing everyone being treated or untreated. Thorough sensitivity analyses were conducted to assess the robustness of the results. This meticulous PSM approach significantly contributed to a nuanced understanding of the recovery strategy's impact while effectively controlling for potential confounding variables.

Results

The findings indicated a significant difference in the actual recovery amount between the two recovery levels, with Level 1 showing a substantial increase. Both RDD and PSM showed that the elevated Recovery Strategy (Level 1) had a positive and statistically significant effect on the actual recovery amount. RDD showed that the average additional recovery amount was $277.63, while PSM showed that the average additional recovery amount was $524.96. Considering the extra cost of $50 for Level 1, the strategy was deemed worthwhile, demonstrating a positive cost-benefit relationship.

Limitations

A limitation in the analysis is the potential existence of unmeasured confounders or external factors influencing recovery outcomes that are not captured in the available dataset. It is possible that other

factors, such as changes in the economy or the bank's collection practices, could have also contributed to the increase. Collecting additional data on customer engagement, communication channels, and economic conditions could have enhanced the model's explanatory power. Despite controlling for known variables, the observational nature of the data makes it challenging to establish true causation, highlighting the need for caution in interpreting the recovery strategy's effectiveness.

Another limitation is that the study is based on a relatively small sample size. This means that the results may not be generalizable to the wider population of customers with charged-off accounts. Despite the limitations, the results of this study suggest that the Level 1 recovery strategy is an effective way to increase the amount of money that the bank recovers from charged-off accounts.

Recommendation: Banking Recovery Strategy

Key Takeaways: Our comprehensive analysis, employing RDD and PSM, underscores the efficacy of implementing Level 1 in the banking recovery strategy. Both PSM and RDD consistently indicate a statistically significant increase in the actual recovery amount at Level 1, with an estimated ATE of approximately $524.96. Importantly, this exceeds the extra cost of $50 associated with Level 1, indicating a favorable cost-benefit relationship.

Question 1: Does the extra amount that is recovered at the higher strategy level exceed the extra $50 in costs? Affirmatively, our analysis confirms that the extra amount recovered at Level 1 significantly surpasses the additional $50 in costs, establishing Level 1 as a cost-effective choice for the bank.

Question 2: What banking recovery strategy should the bank adopt, e.g., whether it is worthy to pursue Level 1 v.s. Level 0? The recommended banking recovery strategy is unequivocally to pursue Level 1. This recommendation is substantiated by both its statistical significance, as indicated by the ATE, and its practical cost-benefit considerations. The strategic decision to implement Level 1 is poised to not only maximize recovery outcomes but also enhance the overall efficiency of the recovery process.

Call to Action: Embrace Level 1 in the recovery strategy. The cost-benefit analysis affirms that the extra recovery at Level 1 surpasses the $50 additional cost, promising enhanced financial returns and streamlined recovery efforts, aligning with the bank's goal of optimizing debt recovery.

Conclusion

In conclusion, our analysis supports adopting Level 1 in the recovery strategy, leveraging PSM and RDD. Visualizations and statistical outcomes demonstrate Level 1's effectiveness and cost-efficiency, surpassing the $50 additional cost. This data-driven recommendation positions the bank strategically for optimal financial returns and effective debt recovery.