ECON 7720

Kaggle Contest Group Project

Zhongjian Lin

Fall 2023

Team Members:

Carson Marchetti

Charlie McCollough

Savar Jaitly

Armen Amirkhanyan

Funso Adewuyi

Banking Recovery Model Discussion

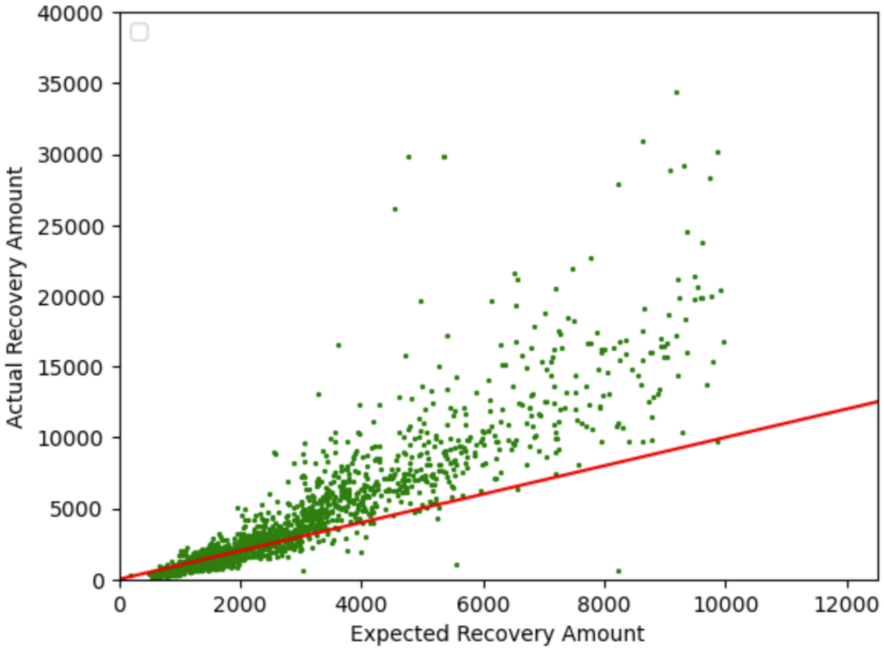
Project Summary

After a bank deems an account “uncollectable,” the account is “charged-off.” After which, it is up to the bank to determine how much money is expected to be recovered and the extent of resources that should be expended in the hopes of reclaiming the capital. This is a function of the likelihood of the customer paying, the total debt outstanding, and other debt-recovery-related factors. We are told that the bank has implemented recovery strategies at different thresholds ($1000, $2000, etc.) where the greater the expected recovery amount, the more effort the bank puts into contacting the customer. The recovery amounts are specifically ranked in levels, where each additional level requires an additional $50 per customer. In other words, it increasingly costs the bank to contact the customer. For debts with low expected recovery (Level 0), the bank uses automated systems like dialers and emails. In contrast, higher levels involve more direct human intervention.

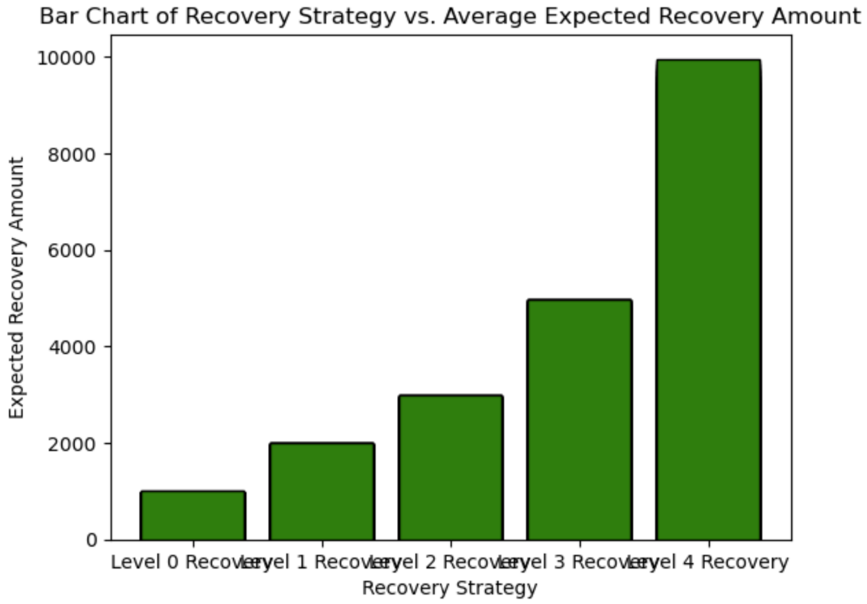
Therefore, the project is focused on analyzing and optimizing debt recovery for this bank. The primary goal is to determine whether the additional expenses incurred at higher recovery levels are worth the additional amounts recovered. Specifically, the project centers on the cost-benefit analysis of the Level 1 recovery amount, which involves a threshold of $1000.

The crux of the project is to develop a model that determines how efficient the bank’s strategy is. This model analyzes data from the provided CSV file to predict the recovery amounts at different levels and compare them with the costs incurred. The goal is to evaluate whether upgrading a customer from Level 0 to Level 1 results in a sufficient increase in recovered funds to justify the additional expenditure. If yes, this would mean that the level upgrade was accompanied by an additional recovery amount greater than $50.

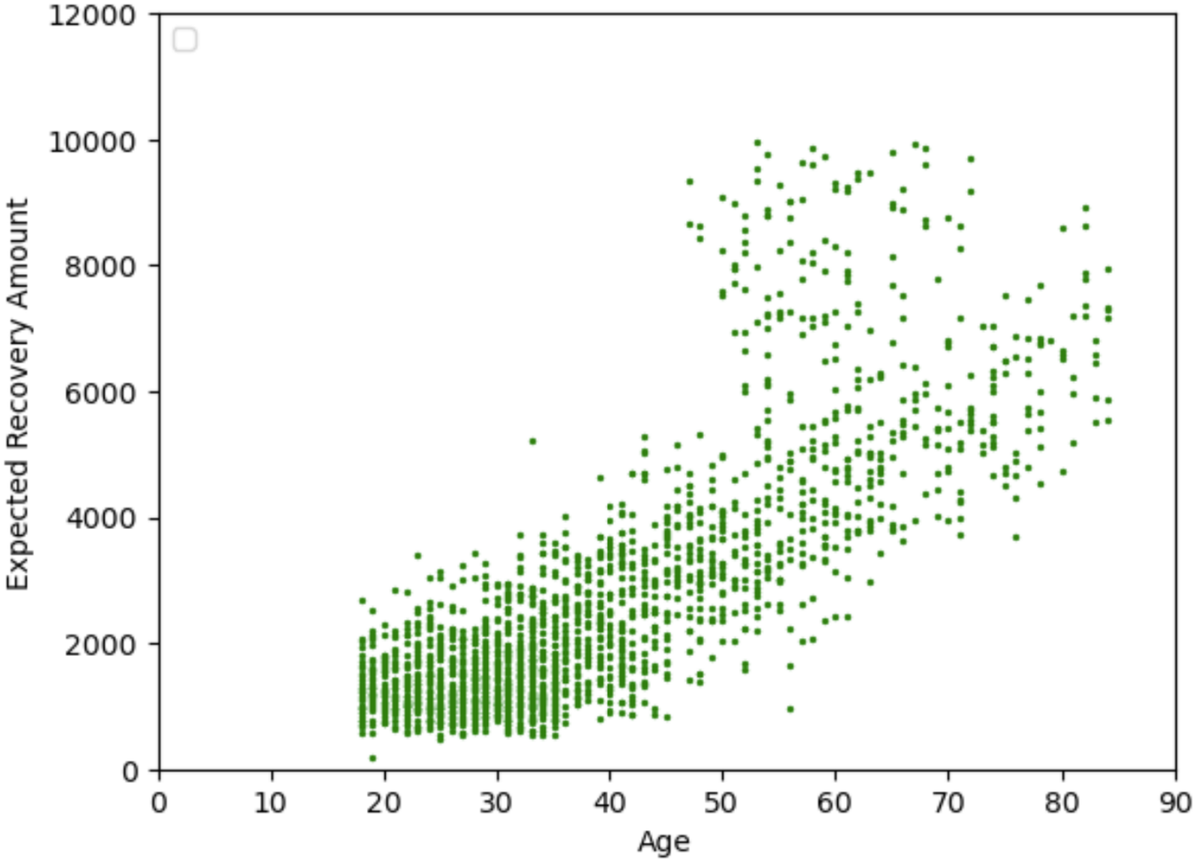
Exploratory Analysis



To the right is a scatterplot of the data with the actual recovery amount on the y-axis and the expected recovery amount on the x-axis, both in US dollars. Each individual green dot represents a customer, and the red line is a line we included to represent a perfect expected recovery amount (1:1 slope with respect to the x-axis). As one can see, the bank actually recovered well over the expected amount. This demonstrates that their strategy is working well.



The bar chart for recovery strategy versus expected recovery amount shows the average expected recovery amount for each strategy. This follows suit with the context of our problem and the data set, showing that each recovery strategy has a range of expected recovery. In the problem, it is explained that each recovery strategy has an additional cost of $50 compared to the strategy before it, which supports the question of the bank, is the recovery amount worth the additional investment per recovery?

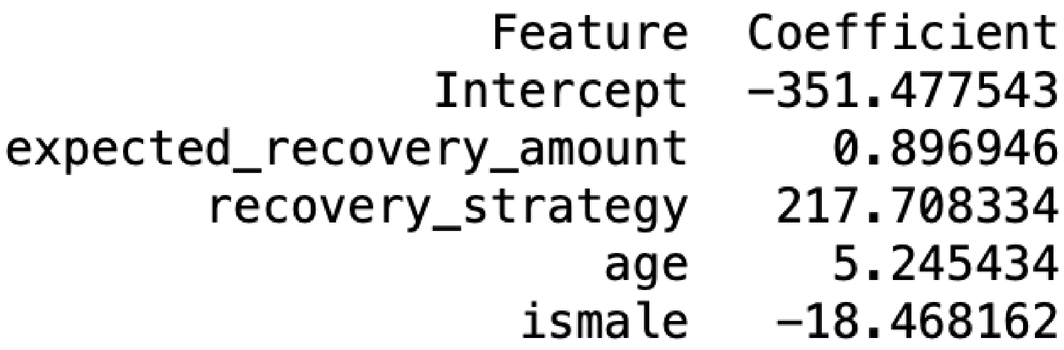


The above scatter plot shows the correlation of age with the expected recovery amount of their uncollectable account. The correlation shown in the scatter plot depicts that with the increased age of the account holder, there is possibly an increase in the amount of expected recovery that the bank could gain back. Some possible explanations for this correlation is that older individuals could have increased income and debts which would increase the size of their account collections and therefore their expected recovery amount if the bank chooses to pursue the collection.

Model Discussion

Linear Regression

We are running a linear regression to estimate the effect of the level 1 recovery strategy, however, this is likely not the true causal effect.

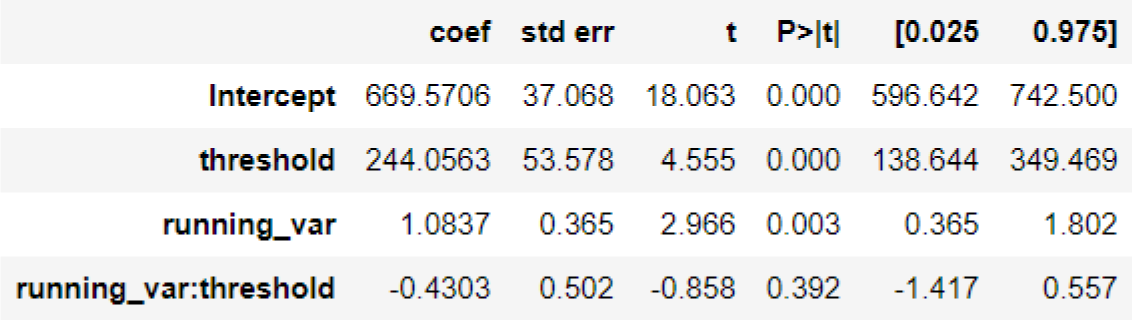


```
                 Feature  Coefficient
               Intercept  -351.477543
expected_recovery_amount     0.896946
       recovery_strategy   217.708334
                     age     5.245434
                  ismale   -18.468162
```

We filtered out the recovery strategies to only level 1 and 0. We then converted the values to 0's and 1's. We did this with gender as well. We then split the data into training (80%) and testing (20%) sets to fit the linear regression model to the training data. The primary variable of concern was recovery_strategy, which yielded ~217.71. This indicates that when the recovery_strategy level is 1, the actual recovery amount is on average holding all else constant 217.71 dollars higher than when the recovery_strategy level is 0. This value is well over 50.

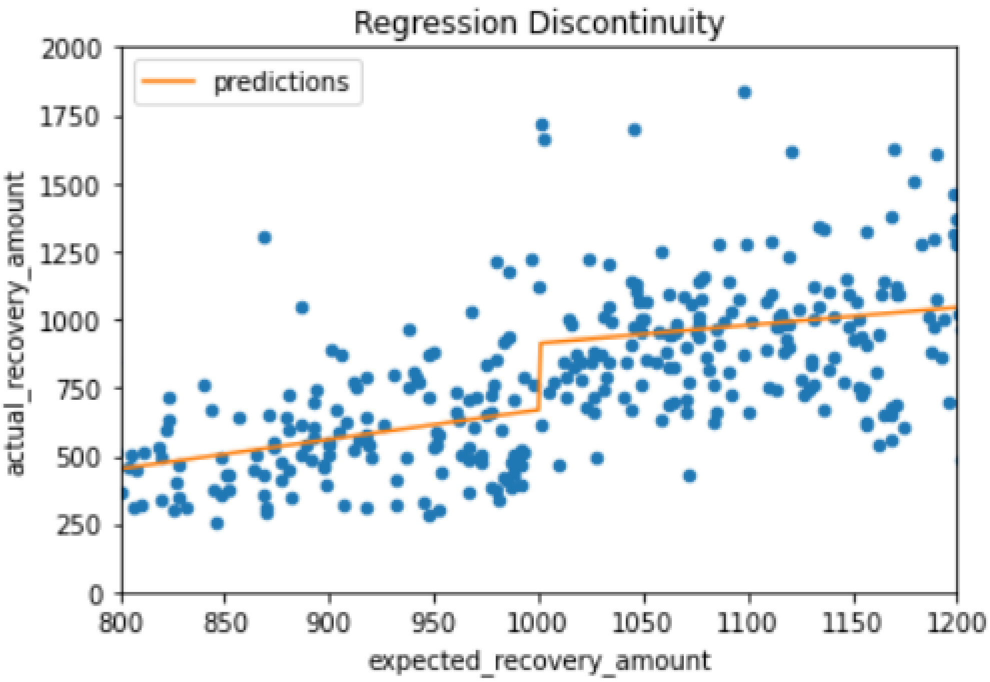
Regression Discontinuity

In order to find the true causal effect that the level 1 recovery strategy has on the actual recovery amount, we can instead use a regression discontinuity design. In order to do this, there should be no evidence of manipulation of data near the threshold, which we have assumed to be true. To start, we removed the id column from the data and subsetted the observations to only include recovery strategies 0 and 1. Before we run a regression to find the effect of a level 1 recovery strategy, we need to create an indicator variable to determine if an observation has passed the threshold or not. Also, we need to create a running variable to account for the observations distance from the threshold. Finally, we must subset the data to only include observations that have an expected recovery amount between 800 to 1200. Now we are ready to create our regression where we predict the actual recovery amount using the following features: whether the observation has crossed the threshold, the distance from the threshold, and the interaction between crossing the threshold and distance from the threshold.



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	669.5706	37.068	18.063	0.000	596.642	742.500
threshold	244.0563	53.578	4.555	0.000	138.644	349.469
running_var	1.0837	0.365	2.966	0.003	0.365	1.802
running_var:threshold	-0.4303	0.502	-0.858	0.392	-1.417	0.557

The main coefficient of interest from our regression summary is the one corresponding to threshold, since it gives us the causal effect of moving from recovery strategy 0 to recovery strategy 1. We can see that by spending the additional $50 for recovery strategy 1 increases the actual recovery amount by $244 on average. This can be seen as the vertical jump in the regression plot below.



Conclusion

As evidenced by our exploratory analysis, OLS regression, and RDD model, upgrading a customer from Level 0 to Level 1 results in a sufficient increase in recovered funds to justify the additional expenditure. To reiterate, this means that the level upgrade was accompanied by an additional recovery amount greater than $50. In the regression, the switch from level 0 to level 1 recovery strategy **estimates** a jump of 217 dollars which is much greater than the 50-dollar investment of the upgraded recovery strategy. This conclusion is also supported by the regression discontinuity model which shows a jump of 244 dollars at the threshold. This again is much greater than the 50-dollar jump for the recovery strategy so overall, the bank should choose to upgrade the level of recovery.

For future steps, it would be helpful to analyze the other levels that extend beyond level 1. We discovered that the bank's strategy performs well at the $1000 threshold, but how about $5000 or $10,000? Our conclusion may be only partially useful in the grand scheme of things for the bank.