RIT Business Analytics Competition Spring 2022

Rochester Institute of Technology
Tron Schell - tys6181@rit.edu
Matt Hochman - mbh1188@rit.edu

Executive Summary

Through an in-depth analysis of past loan portfolios, we believe Small Capital Bank (SCB) can improve portfolio profitability, borrower selection, and reduce bias. This report outlines the methodologies, tools, and techniques used to provide SCB with the recommendations necessary to improve its business.

SCB is concerned with the profitability of its loans and wishes to have an in-depth understanding of its portfolio health. Our analysis utilized loan profitability ratios and regression models to find important variables affecting profitability. SCB's 2017-2018 portfolio was not profitable, with and without considering loans with a current status. SCB has the potential to increase profits by using an improved default prediction model, specifically, one that we created and outlined in this report.

Our model was created using XGBoost, an advanced machine learning library designed to be highly efficient, flexible, and portable ("XGBoost documentation", 2021). Through optimization and hyperparameter tuning, our model yielded an F1 score of 0.89, proving to be accurate at predicting the default probability of prospective borrowers in SCB's 2019 portfolio. By using our predictive model, SCB can improve portfolio profitability by avoiding customers with a high chance of default.

It is important that SCB wants to explore the possibility of bias in its processes as bias can affect profitability and raise ethical concerns. It was found that SCB does have two variables, age, and gender, that can affect portfolio bias. It was found that age had a significant impact on our model. In order to reduce bias and improve our model, our report recommends solutions such as changes to data collection or altering the weight certain variables have on our model to reduce bias.

Data Preparation

Dataset 1 (2022-dataset1) contains 877,986 loans of 877,986 borrowers from SCB's 2017-2018 portfolio. There are 47 variables, which we split into three categories: Credit/Financial, Loan Details, and Personal Information. One of these variables, "loan_status" is our target variable for default prediction. There are three types of loan status: current, paid, and default; which needed to be converted to a binary variable to be used within our prediction model. Loans with a current status were omitted as they didn't aid the model in predicting defaults.

Dataset 2 (2022-dataset2) contains 495,242 prospective borrowers from SCB's 2019 portfolio. Compared to Dataset 1, there are 39 variables. The missing variables concern details of loans that are currently accepted by SCB, such as "interest_rate" and "loan_status". Dataset 2 contains prospective loans, therefore these details are not present.

In order to analyze SCB's portfolio, we applied the Loan Profitability Formula (Formula 1) to all the loans in Dataset 1. Additionally, for defaulted loans, we calculated the time in weeks that passed until they defaulted.

To prepare Dataset 1 to be used as a training set for a predictive model, several actions were taken. In addition to omitting loans with the current status, all variables not present in Dataset 2 were removed. Another variable "employment_length" contained symbols that had to be removed to ensure compatibility with our XGBoost machine learning model, and all other categorical variables needed to be converted to numerical columns through the use of "One Hot Encoding" Once Dataset 1 was sufficiently cleaned, there was a consideration to remove outliers, but it was decided against due to XGBoost's lack of sensitivity to such outliers.

Data Analysis

To begin our exploratory data analysis, we used both a Tableau dashboard (Figure 6) and Python to examine the distribution of Dataset 1 by "loan_status". Here we found that there is a class imbalance present between paid and defaulted loans, with 26% of the loans being defaults. Due to

the class imbalance, it was clear that our default prediction models would have to be scored based on F1 score, rather than accuracy.

We found through analyzing the distributions of Credit/Financial variables that paid and defaulted loans have near-identical Fico scores (Figure 1). The means of high and low Fico scores for paid and defaulted borrowers in the same proximity, with only a difference of ten points. Additionally, variables considered red flags for lending, such as "default_12months" and "collections_12months" were checked. It was found that a majority of borrowers had a value of 0 in these variables. A second distribution was produced, omitting all 0 values to see if there was a difference in those that did have defaults or collections within the last twelve months. Again, borrowers that paid and defaulted on their loans were found to be near-identical, with most borrowers having only 1 default or collection within the last twelve months (Figure 2). These observations informed us that SCB's current loan approval process does appear to avoid unappealing borrowers. We also took note that this may have an effect on our default prediction model and what variables are considered impactful to it.

Profitability

When analyzing SCB's 2017-2018 portfolio health, we focused on the amount of time it took for loans to default and loan profitability ratios. It was found that the average time to default was about 15 months for 36-month loans and 16 months for 60-month loans (Table 1). Overall, the average time to default for all loan terms was also 15 months (Table 1). It is important to know this because the longer a defaulted loan is active, there will be positive impact on the loan's profitability ratio. The entire portfolio of paid and defaulted loans has a mean profitability ratio of 0.93 (Table 2). This can be broken down into a mean profitability ratio of 1.12 for paid loans and 0.40 for defaulted loans (Table 2). Due to 36 and 60 month defaulted loans having an average length of 15 and 16 months until default, they have a considerate difference in profitability. 36-month defaulted loans have a mean profitability ratio of 0.46 compared to 0.30 for 60-month loans (Table 2). 60-month loans on average do not have enough time to accrue enough principal and interest to obtain a higher profitability ratio.

We constructed a linear regression using the loan profitability ratio, "loan_profit", as our target variable to be explained by variables from the Credit/Financial, Loan Details, and Personal Information categories. Two different regression models were constructed, for paid and default loan status. After several rounds of removing statistically insignificant variables, it was found that certain variables remained the most important to each model.

For paid loans, It was found that "loan_interest_rate" was the most statistically significant. This can be explained due to the fact that most paid loans will reach a ratio of 1.0 because they have paid back the principal of the loan. Any subsequent profit will come solely from the interest rate. There are other statistically significant variables, such as "bankruptcy_record" and "marital_status", but they are considerably less significant than "loan_interest_rate".

For defaulted loans, "marital_status" and "loan_interest_rate" were the most significant variables. It must be noted that "loan_term" was omitted from the regression as we already determined that it was the most important variable to defaulted loans' profitability. That being said, it makes sense that "loan_interest_rate" would be one of the most important variables for defaulted loans; a loan with a higher interest rate will be more profitable than that of a lower interest rate. Besides "martial_status", other variables such as "annual_income", "derogatory_record", and "bankruptcy_record" were found to be statistically significant. These can all be considered indicators of different levels of wealth and could possibly contribute to a loan's profitability.

Predicting Defaults

To understand the factors behind loan defaults, a predictive model was to be constructed. The goal of the model is to accurately predict whether or not a borrower will default on their loan using the variables present in both Datasets 1 and 2. Three models were considered: Decision Tree, ADABoost, and XGBoost (Figure 3). The models varied in complexity and ability to provide robust results. After testing all three models, it was found that XGBoost yielded the greatest F1 score when tested on Dataset 2, leading to its selection. Our first XGBoost model produced an F1 score of 0.16, which was determined to be too low to be an effective tool for prediction. This score was improved over time through hyperparameter tuning, and feature selection, eventually yielding an F1 score of 0.89.

It was found that the most important variable was "age" (Figure 7). This immediately raised concerns as a borrower's age should not be an indicator of default. At first, variables in the Credit/Financial category were presumed to be the most important as they indicate wealth, financial stability, and credit. To investigate this further, we analyzed the loan status, paid or default, of different age generations such as Generation Z, Generation Y, etc. What we found was that people ages that fell in the Generation Z category had 68% less favorable labels in loan status and Generation Y had 39% less favorable labels in loan status. However, Generation X had 28% more favorable labels and this trend continues until people's ages of Post War with 26% more favorable labels (Figure 5).

Another observation made was the difference in importance among states. California had the highest level of importance compared to much lower levels of other states. By looking at the distribution of borrowers across states, it was found that in Dataset 1, SCB had a vast majority of its borrowers from California, followed by Texas, New York, and Florida. There were 1.84 times more borrowers from California than Texas. It was also found that SCB had no borrowers from Iowa and Colorado.

Results

After our thorough analysis of SCB's loan portfolios, several key findings were made about SCB's profitability and the presence of bias in its portfolios.

In regards to SCB's 2017-2018 portfolio, it was found that it was just below profitable. With a profitability ratio of 0.93 between paid and defaulted loans, SCB appears close to making up for its losses. In terms of cash, $6.81 billion has been lent to borrowers, while $6.44 billion has been returned. In SCB's 2017-2018 portfolio, if all of the current status loans for this portfolio were to default, SCB's profitability would drastically decrease to 0.63 and SCB would be $2 billion short of breaking even. The most important factors for profitability are ensuring loan repayment and interest rates. Once a loan has been repaid it has a guaranteed profitability ratio of 1.0. From there interest rates will determine profitability. Defaulted loans will not be able to reach a 1.0+ ratio so it is important to either avoid borrowers that have a high chance of default or consider higher interest rates for these riskier loans. All defaults on average occurred 15 months into the loan, it is possible to increase profit on the loan by having a higher interest rate for that period. Additionally, for loans with an increased chance of default, the loan term is important. 36-month loans will be more profitable than 60-month loans when they default. This is due to the fact that a larger percentage of the 36-month loan's term has been completed. SCB should avoid risky 60-month loans in order to increase its profitability.

The best way to maximize profits is to use our model for predicting defaults. By accurately predicting the possibility of default from a borrower, SCB can avoid taking on unnecessary risk, possibly damaging profits. Currently, the model strictly assigns borrowers with probabilities of default over 50% as a borrower who will default. Our model can be adjusted to allow different

probabilities of default to be accepted. This way, SCB can adjust the model to represent the amount of risk they are willing to take on.

Consideration of Bias

According to the Equal Employment Opportunity Commission (EEOC), there are eight protected classes including race, color, national origin, religion, sex, age, disability, and genetic information, within the United States ("Legal Information Institute", 2020). Of which, age and sex apply to the SCB dataset.

The current data collection methods that are being used are acceptable, considering that only two out of the eight protected classes are collected, but can be improved upon. Upon some analysis using the mean difference in outcomes from the privileged and unprivileged classes like age generations, and gender specificity, we found some instances of inherent bias that the machine learning model will learn. Because Age was such a highly determinant feature within the model, we want to avoid the model learning any instance of bias within the age category, of which we found that the lower a person's age is, the less likely that they would receive a loan from the model and the opposite for people that were older. Generation Z had 68% less favorable loan status labels and in comparison, people from Generation X had 28% more favorable loan status labels (Figure 5). Our suggestions for mitigating these biases are twofold (Figure 4). Our first suggestion would be to completely eliminate the protected classes, age, and sex from the dataset and to instead use different metrics to determine loan worthiness. Some examples could be the use of collateral, of which SCB can check if the collateral includes a depreciating asset and the total amount of collateral.

If SCB cannot change its data collection methods, then another solution would be to use a tool like IBM's AiFairness360 in order to transform the data to make the model fairer during training. There are three ways in which the AiFairness360 open source library can be utilized with a particular dataset, pre-processing, in-processing, and post-processing, with in-processing being the most effective ("Trusted AI", 2021). Our suggestion would be to first see if reweighting the protected and unprotected classes has a significant effect on mean outcomes and model performance (Kamiran & Calders, 2011). We suggest this because this can easily be implemented as a function, in which the data can always be run through before model training and is easily explainable to shareholders and/or employees.

Bibliography

1. *XGBoost documentation*. XGBoost Documentation (2021) - xgboost 1.5.2 documentation. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/stable/

2. Legal Information Institute. (2020). *Protected characteristic*. Legal Information Institute. Retrieved from https://www.law.cornell.edu/wex/protected_characteristic

3. Trusted-AI. (2021, March 4). *Trusted-ai/AIF360: A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models.* GitHub. Retrieved from https://github.com/Trusted-AI/AIF360

4. Kamiran, F., & Calders, T. (2011, December 3). *Data preprocessing techniques for classification without discrimination - knowledge and information systems*. SpringerLink. Retrieved from https://link.springer.com/article/10.1007/s10115-011-0463-8

Appendix

Formula 1

$$Loan\ Profitability\ = \frac{\Sigma\ Loan\ Collect - \Sigma\ Loan\ Loss}{Loan\ Amount}$$

Table 1

Time to Default	
% Defaulted (Default / Default + Paid)	26%
Average Months to Default (All)	15
Average Months to Default (36 Month)	15
Average Months to Default (60 Month)	16

Table 2

Profitability Ratios	
Mean of All Loans	0.62
Mean of Paid Loans	1.12
Mean of Defaulted Loans	0.40
36 Month Loans	0.46
60 Month Loans	0.30
Mean of Current Loans	0.25
Mean of Paid & Defaulted Loans	0.93

Figure 1

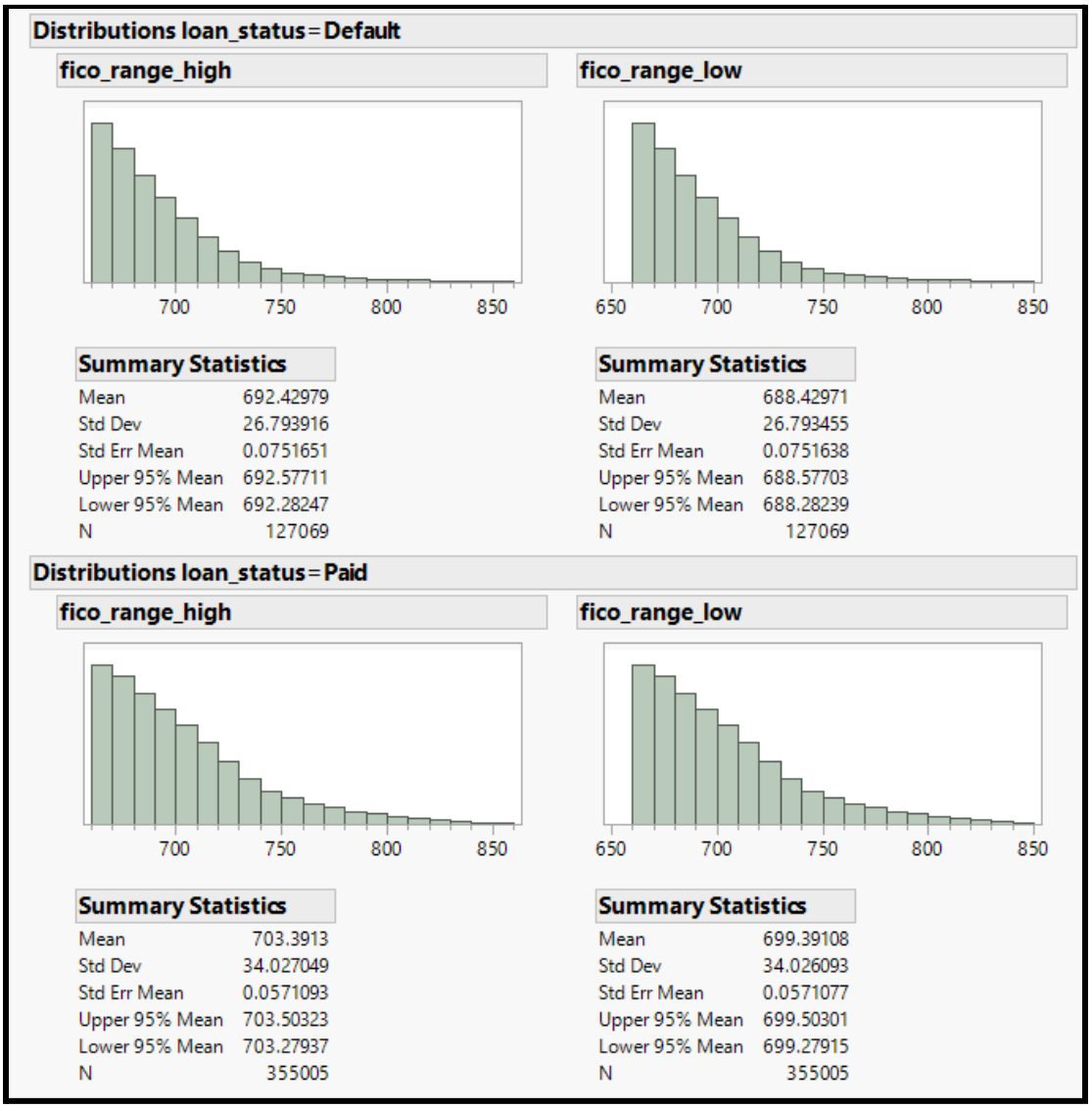

Distributions loan_status = Default
fico_range_high
700
750
800
850
Summary Statistics
Mean 692.42979
Std Dev 26.793916
Std Err Mean 0.0751651
Upper 95% Mean 692.57711
Lower 95% Mean 692.28247
N 127069
fico_range_low
650
700
750
800
850
Summary Statistics
Mean 688.42971
Std Dev 26.793455
Std Err Mean 0.0751638
Upper 95% Mean 688.57703
Lower 95% Mean 688.28239
N 127069
Distributions loan_status = Paid
fico_range_high
700
750
800
850
Summary Statistics
Mean 703.3913
Std Dev 34.027049
Std Err Mean 0.0571093
Upper 95% Mean 703.50323
Lower 95% Mean 703.27937
N 355005
fico_range_low
650
700
750
800
850
Summary Statistics
Mean 699.39108
Std Dev 34.026093
Std Err Mean 0.0571077
Upper 95% Mean 699.50301
Lower 95% Mean 699.27915
N 355005

Figure 2

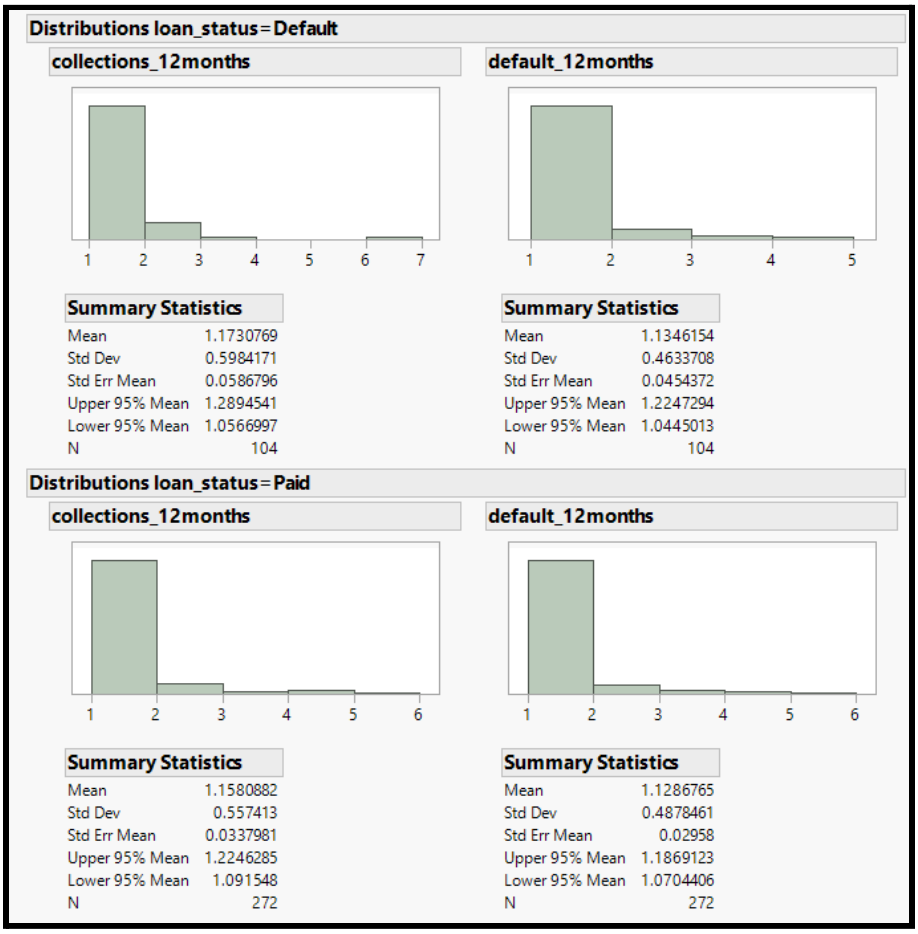


Figure 3

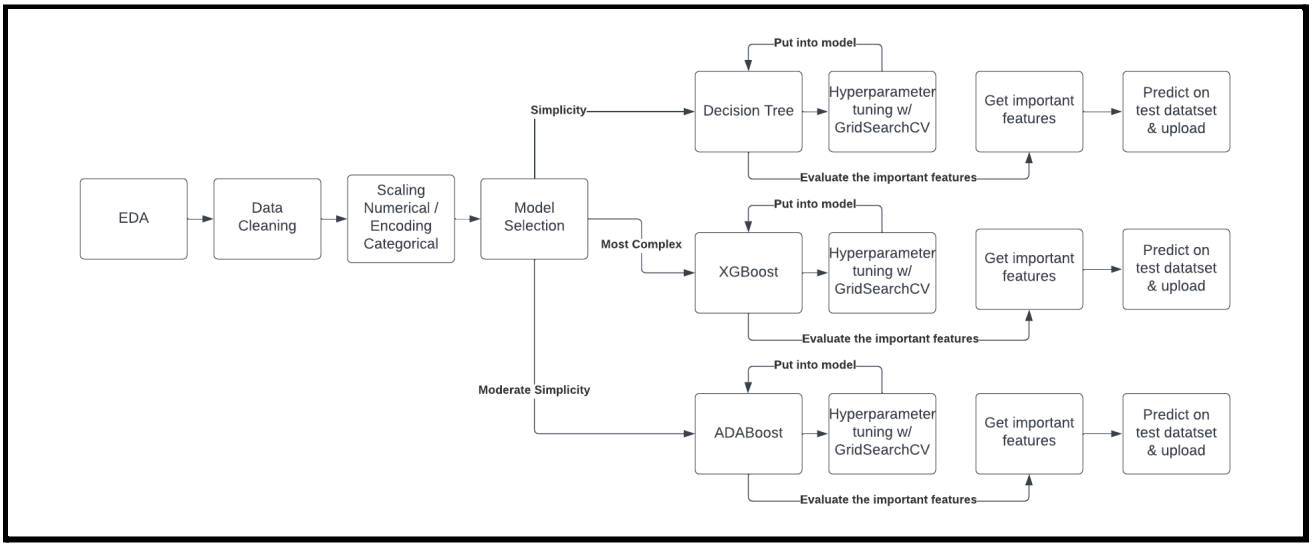


Figure 4

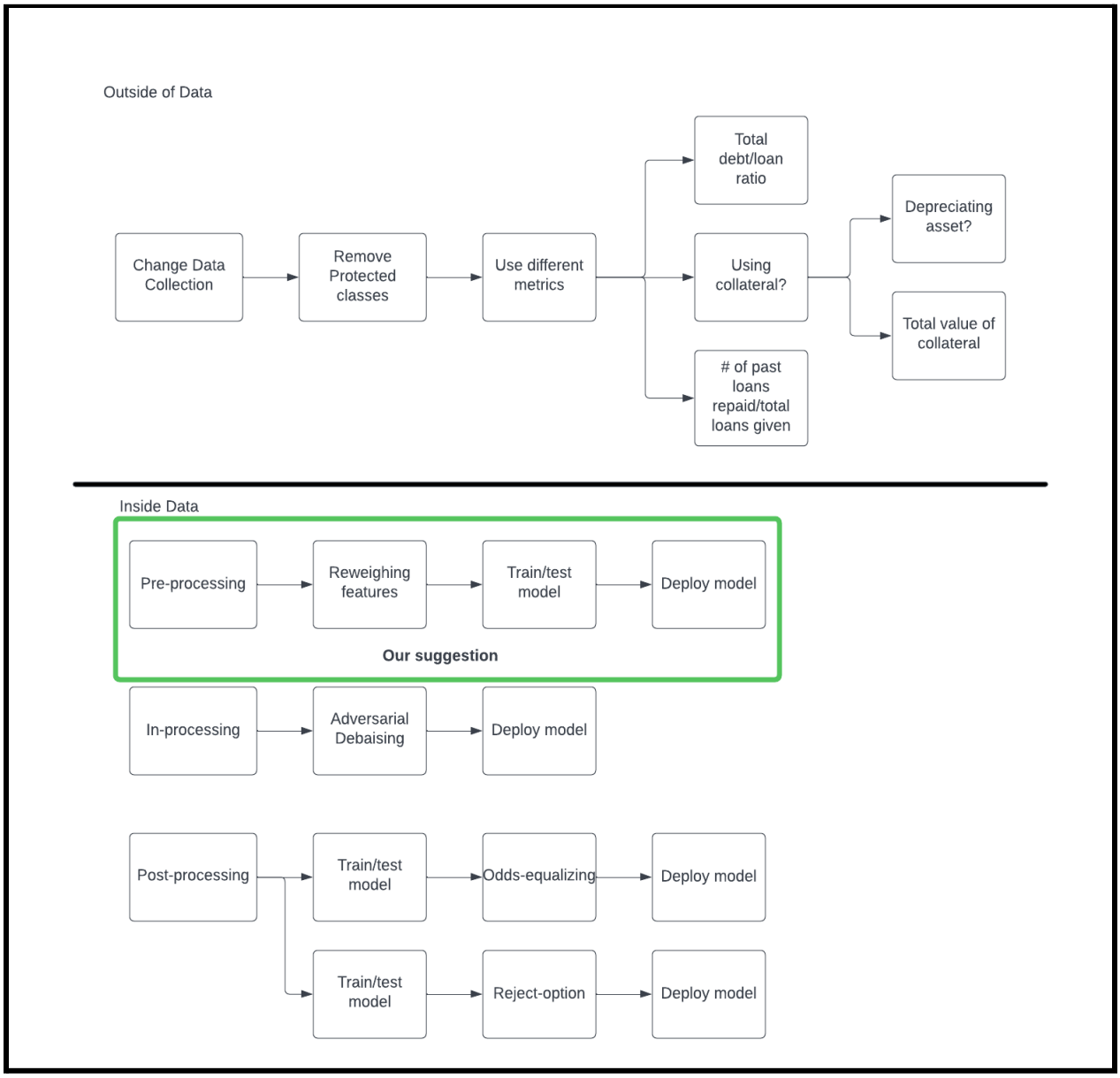

Outside of Data
Change Data Collection
Remove Protected classes
Use different metrics
Total debt/loan ratio
Using collateral?
# of past loans repaid/total loans given
Depreciating asset?
Total value of collateral
Inside Data
Pre-processing
Reweighing features
Train/test model
Deploy model
Our suggestion
In-processing
Adversarial Debaising
Deploy model
Post-processing
Train/test model
Odds-equalizing
Deploy model
Train/test model
Reject-option
Deploy model

Figure 5

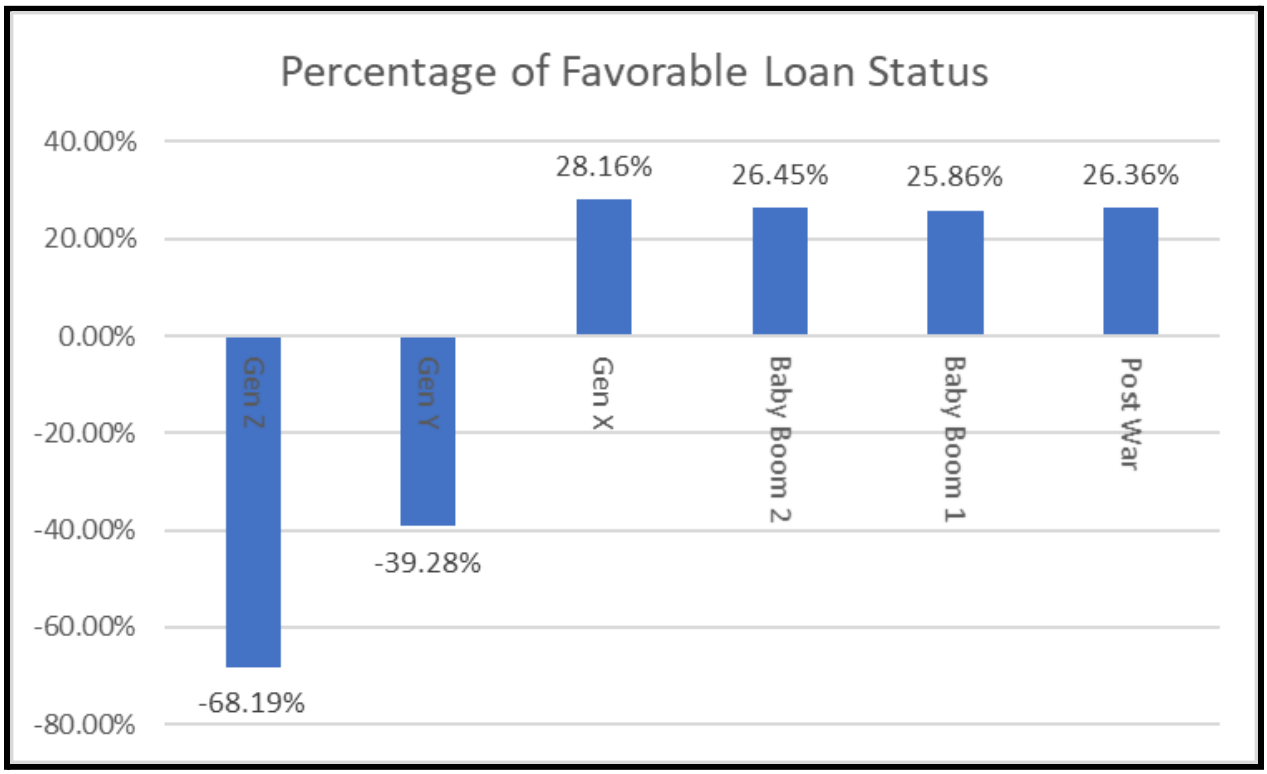


Figure 6

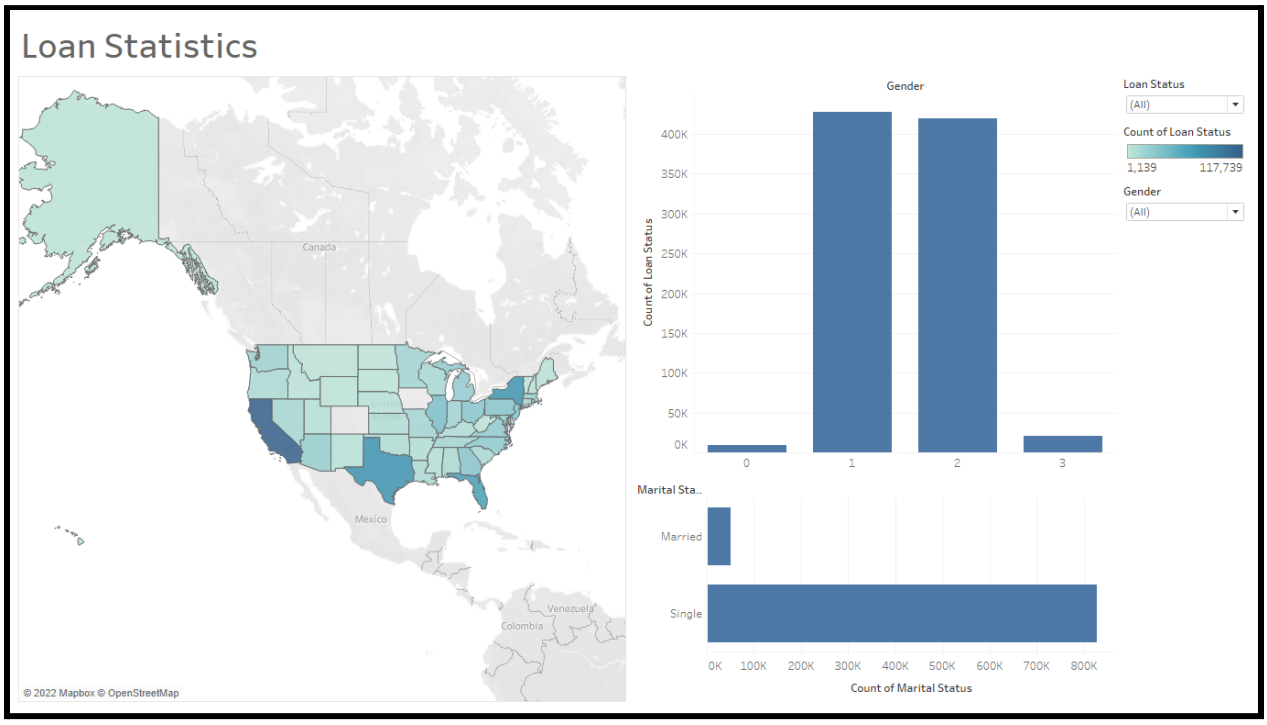
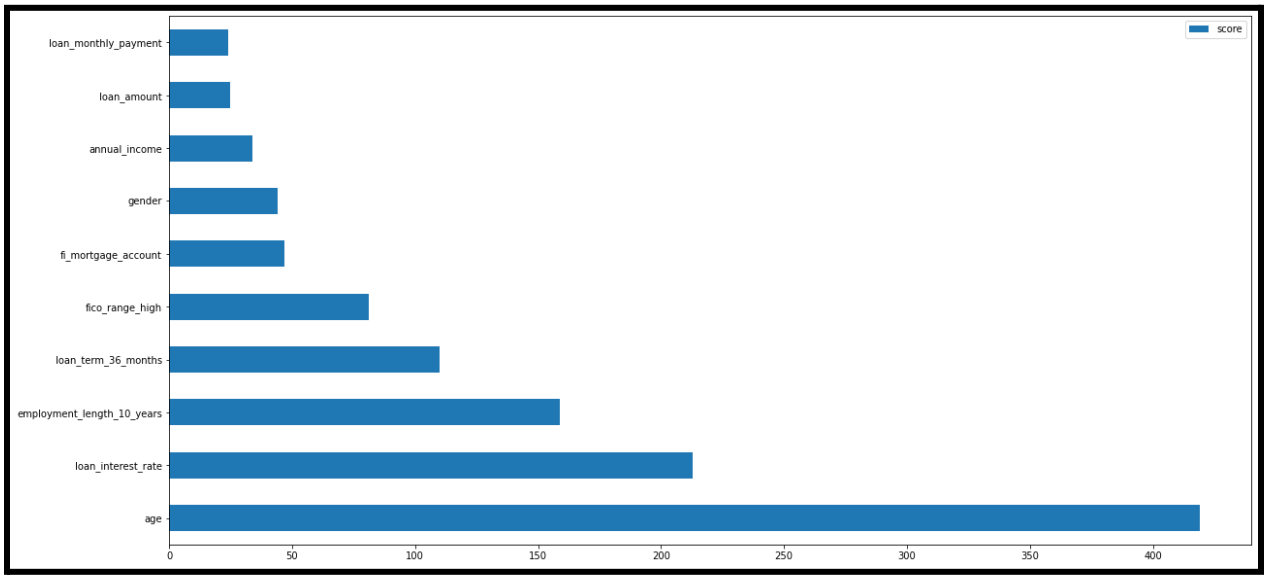


Figure 7


score
loan_monthly_payment
loan_amount
annual_income
gender
fi_mortgage_account
fico_range_high
loan_term_36_months
employment_length_10_years
loan_interest_rate
age
0
50
100
150
200
250
300
350
400