# Team 2 Project Proposal

**Team members:** Glenda Pinho, Mollika Tahsin, Subhransu Nanda, and Somdath Rampersad

The dataset chosen for our project is from the U.S. Small Business Administration (SBA). The SBA is a government agency that assists small enterprises in the U.S. credit market by providing loans. For our project, we are analyzing whether an SBA loan should be approved or denied based on the provided SBA loan data. The dataset being used for our project analysis can be found at <https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied>. The dataset comprises of 899,164 instances and 27 variables. Table 1 below summarizes some of the characteristics of the data. There are 15 categorical and 12 numeric variables.

*Table 1 - Summary of Variables*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Classification** | **Number of Missing Values** |
| LoanNr\_ChkDgt | Identifier Primary key | Categorical | 0 |
| Name | Borrower name | Categorical | 14 |
| City | Borrower city | Categorical | 30 |
| State | Borrower state | Categorical | 14 |
| Zip | Borrower zip code | Categorical | 0 |
| Bank | Bank name | Categorical | 1559 |
| BankState | Bank state | Categorical | 1566 |
| NAICS | North American industry classification system code | Categorical | 0 |
| ApprovalDate | Date SBA commitment issued | Numerical | 0 |
| ApprovalFY | Fiscal year of commitment | Categorical | 0 |
| Term | Loan term in months | Numerical | 0 |
| NoEmp | Number of business employees | Numerical | 0 |
| NewExist | 1 = Existing business,  2 = New business | Categorical | 136 |
| CreateJob | Number of jobs created | Numerical | 0 |
| RetainedJob | Number of jobs retained | Numerical | 0 |
| FranchiseCode | Franchise code,  (00000 or 00001) = No franchise | Categorical | 0 |
| UrbanRural | 1 = Urban,  2 = rural,  0 = undefined | Categorical | 0 |
| RevLineCr | Revolving line of credit:  Y = Yes, N = No | Categorical | 4528 |
| LowDoc | LowDoc Loan Program:  Y = Yes, N = No | Categorical | 2582 |
| ChgOffDate | The date when a loan is declared to be in default | Numerical | 736465 |
| DisbursementDate | Disbursement date | Numerical | 2368 |
| DisbursementGross | Amount disbursed | Numerical | 0 |
| BalanceGross | Gross amount outstanding | Numerical | 0 |
| MIS\_Status | Loan status: charged off = CHGOFF,  Paid in full = PIF | Categorical | 1997 |
| ChgOffPrinGr | Charged-off amount | Numerical | 0 |
| GrAppv | Gross amount of loan approved by bank | Numerical | 0 |
| SBA\_Appv | SBA’s guaranteed amount of approved loan | Numerical | 0 |

The question we seek to answer is “should the loan be approved?” To arrive at our answer, we must first deal with missing values, anomalies, and imbalances in the data. In figure 1 below, we plotted the distributions of several categorical variables. From the plots, we can see that there are variables with missing or invalid values, and imbalances in the distributions.

Our target variable, the loan status, has an imbalanced distribution. We believe that the imbalance may be a natural occurrence given that the dataset comprises approved loans which were already considered to be creditworthy. To overcome the class imbalance, we are proposing to take a random sample, of equal number of instances, from each category (PIF, CHGOFF) to formulate our dataset for analysis. We will use a 70/30 split of this dataset for our training and validation data. Given the large sample size of data, we believe this approach would be feasible.

To handle the missing and invalid values, we are proposing two approaches. The first approach is to drop all instances that have missing or invalid values. The second approach is to impute all missing and invalid values with the mode or mean, depending on the variable classification.

A picture containing logo

Description automatically generated

*Figure 1- Distribution Plots*

Once we have cleaned the dataset, we will attempt to narrow down the number of variables using three feature selection methods; backward, forward, and stepwise. We will select the variables that are the most statistically significant.

After selecting the most statistically significant variables, we will use ensemble learning to make our final prediction. The models we will consider for our ensemble learning are Logistic Regression and Naïve Bayes.