LOAN DEFAULT PREDICTION

Group 4 Final Project

By:

Karthik Badiganti, Sandhya Cheepurupalli,
Abhinay Marineni, Sai Sree Pulimamidi

Abstract:

Loan defaulting happens when a person misses payments for a specified period. Numerous factors, including financial difficulty, job loss, or unforeseen bills, might cause this. Due to the potential loss of income and decline in the value of their loan portfolios, loan defaults are a serious issue for lenders. Defaults may also have a negative effect on the borrower's credit rating, making it more difficult for them to qualify for new loans or credit in the future. Companies regularly utilize data mining tools to assist them in identifying customers who are having loan defaults and are likely to leave. The goal of the project is to analyze past data and understand if there are any customers with loan defaults. Also, the focus of the project is to project if any customer has the probability of facing loan default in the future. This would help the lender to act accordingly.

Introduction:

Lenders may suffer serious consequences if a loan defaults. A borrower has not repaid a debt per the terms and circumstances set forth when they are said to be in default. As a result, the lender can sustain financial losses as the loan's principal, interest, and other fees won't be repaid. Defaulting on a loan can harm the lender's reputation in addition to causing financial losses and operational difficulties. Customers may stop trusting the lender's capacity to control risk, which could result in lost sales and a decrease in market share. Therefore, it is the lender's best advantage to develop effective risk management and credit evaluation rules and procedures to reduce the risk of loan default. To recognize uncertain loan applications, machine learning can be used. Machine learning is essential to the financial ecosystem in several areas, including risk management, asset management, and loan approval. The process financial organizations use to evaluate customer loan applications. The main goal is to reduce investment risk while optimizing the return on investment. This involves figuring out the probability that clients will stop making payments on their loans and calculating the loss that could result from doing so. Profit opportunities are also considered, taking into aspects like interest rates, loan amounts, and terms of length.

Overview and Exploratory Analysis:

The given banking dataset at first was so complicated it consisted of a large amount of data in which the names of clients were not given; instead, it was assigned in the form of ID numbers and unique variables. The data contains 80,000 rows and 763 columns, as shown in *Figure 1*.

```
str(data)
```

```
'data.frame':   80000 obs. of  763 variables:
 $ X   : int  78539 61541 76531 22066 45589 99812 43515 1429 32243 73119 ...
 $ id  : int  78539 61541 76531 22066 45589 99812 43515 1429 32243 73119 ...
 $ f1  : int  120 154 126 127 162 131 122 128 131 131 ...
 $ f3  : num  0.146 0.349 0.976 0.951 0.518 ...
 $ f4  : int  2200 4200 1500 3100 3500 2200 1600 1600 1500 3500 ...
 $ f5  : int  4 4 10 7 7 7 16 4 17 16 ...
 $ f6  : int  76878 76635 7399 14448 80502 2459 78053 75464 14629 9836 ...
 $ f7  : int  8703 2843 437 2681 3840 4312 1555 3947 8766 4948 ...
 $ f8  : int  724 3253 1453 2136 130 3390 92 5802 312 1058 ...
 $ f9  : num  119 158 123 129 160 ...
 $ f13 : int  14 4 15 11 8 10 12 12 14 10 ...
 $ f14 : num  0.742 0.724 0.655 0.803 0.861 ...
 $ f15 : num  0.718 0.724 0.644 0.746 0.861 ...
```

Figure 1: Total Number of Variables Initially

Data Cleaning:

In the process of preparing the data, we cleaned the data. Firstly, we observed the structures of the dataset to check which types of variables were present in it. The dataset consists of numeric variables. There were a total of 762 columns and one loss column. All the columns are masked.

For all the columns, we checked the null values and the percentage of null values. We found that the column with the highest percentage of null values is 17.8%. Then we removed the variables with zero variance and high correlation. Variables with zero variance have the same value for all the samples and therefore do not provide any information to the model. Removing them will reduce the dimensionality of the data and improve the efficiency of the model. Hence, zero variance variables were removed.

Exploring target variables:

After removing the variables with zero variance and high correlation, we found that more than 90% did not have loan defaults, as shown in *Figure 2*. This means that the dataset is highly imbalanced towards the non-default class. Imbalanced datasets can lead to biased and inaccurate models. This is because most classification algorithms are designed to optimize the overall accuracy or error rate, which the majority class can dominate if the dataset is highly imbalanced.

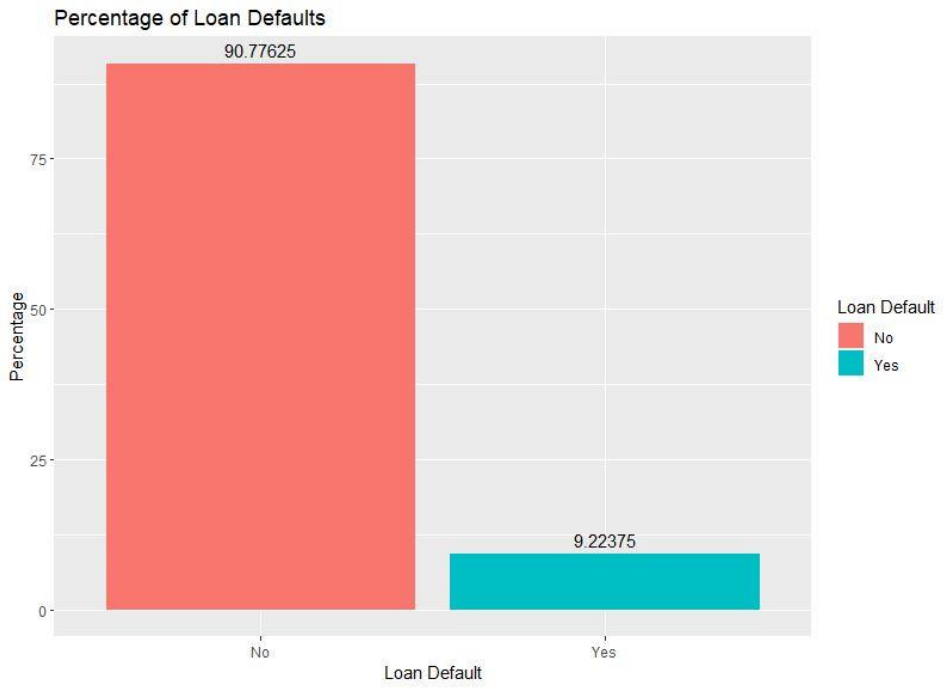

Figure 2: Percentage of People with Loan Defaults

In exploring the dataset and target variables, we observed that among the loan defaulters there is a high percentage of loan loss with 1,2,3 losses. They occupy the major share in the loan defaults data. It can be also observed that even though they have been defaulted we assume that they are paying the loan as early as possible with very little loss.

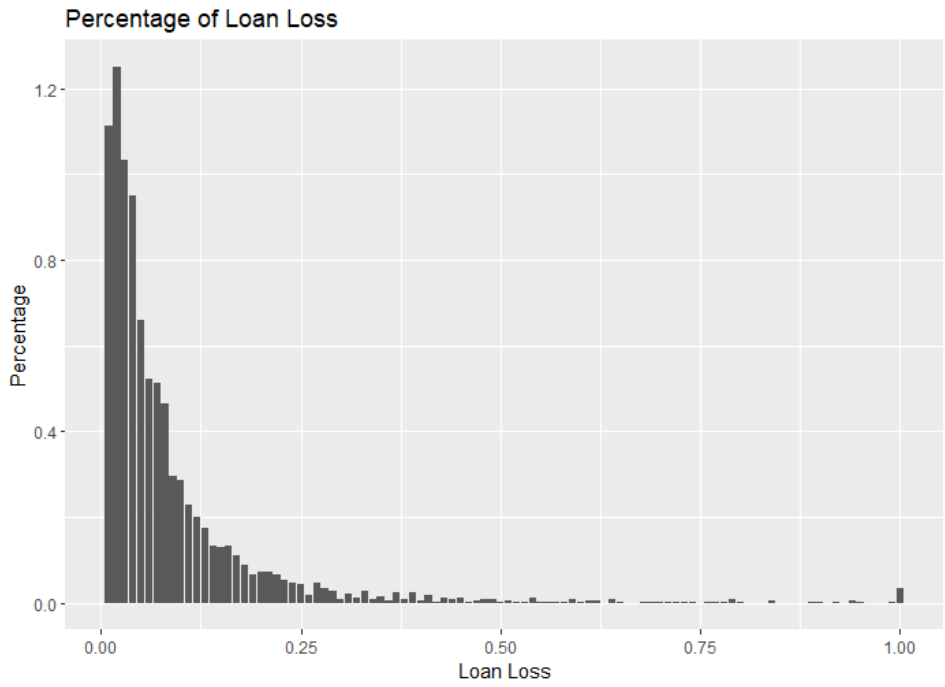

Figure 3: Percentage of Loan Loss

Model Building Strategy:

During the model-building strategy phase, to build an effective model, it's divided into two models, one to predict the loan default and the other one to predict the loss for the defaulters. To achieve this, the data set was cleaned and reduced in size. This process, known as feature selection, involved removing zero variance variables and highly correlated variables, reducing from 763 independent variables to 248 independent variables. Missing values were imputed using the median imputation method before running through a regularization model to further refine the data set. This approach helped us to identify the most critical variables for the model-building process.

Classification Model for Predicting Loan Default:

To improve the accuracy and accessibility of the prediction of people's default rates, the Lasso regression analysis approach was used in this project for both variable selection and regularization. Lasso regression simplifies the model and prevents overfitting by reducing the number of variables. The resulting model is more interpretable and has better predictive power, making it useful for financial organizations. The 248 independent variables were used as inputs to the model. The primary objective of this project was to help the bank determine whether to approve or reject a loan application based on a customer's default history. By leveraging Lasso regression, the bank could identify the most critical variables for predicting default rates, thereby increasing the accuracy of loan approval decisions. The Lasso regression model is widely used in finance and banking due to its ability to handle high-dimensional data sets and select significant variables that contribute to the prediction of the model's overall accuracy.

For the optimal lambda value of 0.0002897 the resulted number of independent variables are 173 as shown in *Figure 4*.

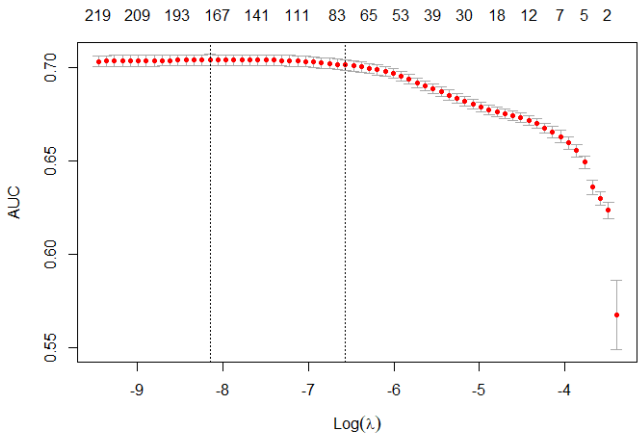

Figure 4: Lasso Regression showing Optimal Lambda

After careful consideration, we proceeded with the first dashed line representing the minimum lambda value. By selecting the minimum lambda value, we figured out a balance between reducing the number of variables while maintaining a high level of prediction accuracy. This approach allowed the team to identify the most important variables for predicting default rates while minimizing the risk of overfitting the model.

Principal Component Analysis (PCA) was used to thereafter reduce the number of variables. PCA is a widely used statistical technique for reducing the dimensionality of large datasets while retaining as much of the original variability in the data as possible. PCA helps to reduce the dimensionality of large datasets while retaining the most important information. This can lead to more efficient and faster analysis of large datasets, making it easier to identify patterns and trends. It can help reduce multicollinearity's effects in datasets where two or more variables are highly correlated. By reducing the dimensionality of the dataset, PCA can remove redundant information and reduce the effects of multicollinearity, making it easier to identify the most significant factors. After deploying PCA, the total number of variables was further reduced to 67 for a 80% threshold variance.

The first vertical dashed line in the graph represents the minimum lambda value. In contrast, the second dashed line represents the lambda value within one standard deviation, which could further reduce the variables.

We decided to use the random forest approach to predict the probability that each customer will default. Before using the Random Forest algorithm, we partitioned the data into 80% and 20%. Where 80% is used for training and 20% for validation. After partitioning, we used the Random Forest algorithm, which creates multiple decision trees and outputs the average value for each customer's probability of default. Random forest can achieve high accuracy by combining the results of multiple decision trees. This makes it particularly useful for complex datasets where other models struggle to achieve high accuracy. The model is less prone to overfitting compared to other models. This is because each decision tree in the forest is trained on a different subset of the data and uses different random subsets of features. We used the "randomForest" package, which allows for further hyperparameter tuning, which has a faster processing time and allows us to attempt more iterations, although it has fewer hyper-tuning parameters.

As in Figure 5, the area under the curve (AUC) as 0.6097 in the ROC curve below.

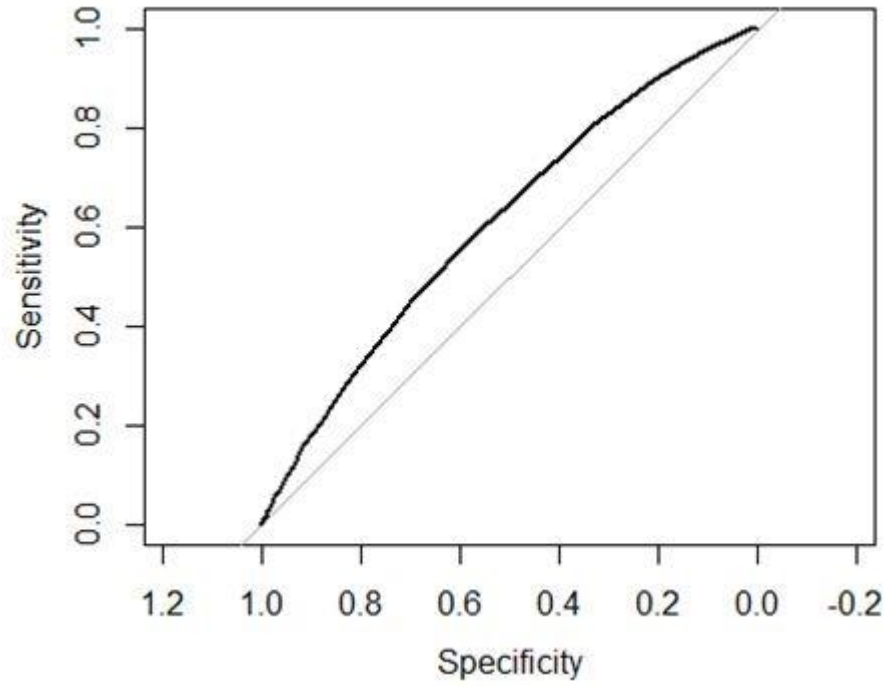

Figure 5: ROC Curve

Regression model for Predicting Loss from Defaulters

To build the Loss Given Default model, the team used regression analysis on data from only defaulted customers. We split the original loss column into default and loss, normalized the values, and then performed feature selection by removing near-zero variance and highly correlated variables. After performing the variable reduced from 762 to 253.

We used Lasso regression to reduce the number of variables. After this, the number of variables decreased to 112 from 253. Figure 5 shows that when we reduce the variables in the dataset to 112 variables using feature selection, the lambda value is 0.000731.

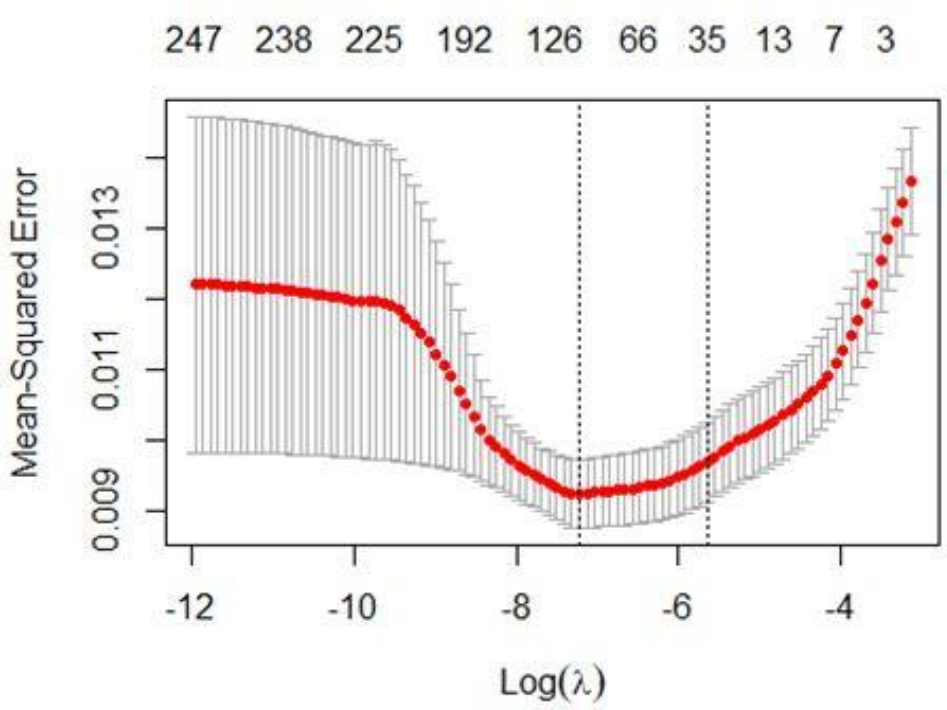

Figure 6: Lasso Regression for and Reducing Number of Variables

Ridge regression was used to build a model for predicting Loss Given Default. All 112, as shown in Figure 6 variables, were used in this model, and the objective function was to minimize the mean absolute error (MAE). After Ridge regression, the MAE is 0.05137, and lambda is 0.04582.

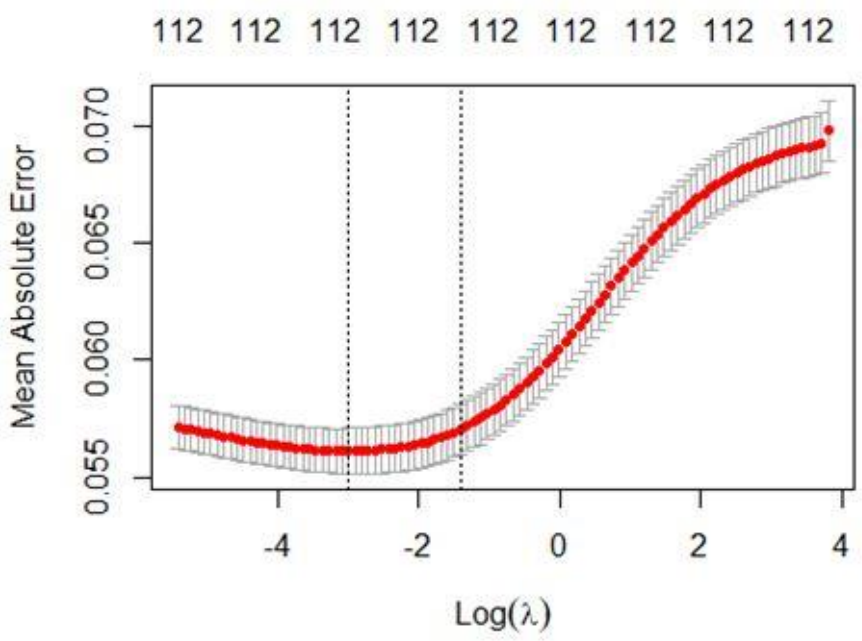

Figure 7: Ridge Regression for Loss Given Default

Estimation of the model's performance

Random Forest for Predicting Loan Default

For measuring the model's performance, we used AUC. This method helps to make True positive and false positive probabilities from the prediction model. The "ntree" and "mtry" variables could be hyper-tuned to further enhance the model by using the "Random Forest" package. A "ntree" value of 500 and a "mtry" value of 10 were chosen as the model's foundation, and this combination produced an AUC result of 0.578.

Ridge Regression for Predicting Loss from Defaulters

Ridge regression was utilized to compute LGD, and the model's performance was measured using Mean Absolute Error (MAE) as a metric. The LGD model produced an MAE of 0.05137, and we experimented with different Lambda values to identify the best one. After analyzing the results, we determined that the most effective Lambda value was 0.04582.

Insights

- The use of Lasso regression and random forest models to forecast loan default rates has shown to be successful and efficient when dealing with high-dimensional datasets.
- The number of independent variables was reduced to 173 using lasso regression, which also helped to regularize the variables. This improved the model's interpretability.
- The random forest technique predicted the probabilities of loan default using multiple decision trees, which produced a strong classification model. It proved capable of managing complicated datasets and reducing overfitting problems.
- The number of variables was further reduced to 67 using Principal Component Analysis (PCA), maintaining a sizable percentage of the variation in the dataset.
- With a low Mean Absolute Error (MAE) of 0.05137, the Ridge regression model for Loss Given Default (LGD) demonstrated a reliable assessment of possible losses in the event of loan default.

Conclusion

In conclusion, this work used Lasso regression and random forest models to forecast loan default rates, providing useful information for lenders to evaluate risk. The important variables influencing default rates were successfully chosen using the Lasso regression method, which also helped to simplify the model and make it easier to understand. By utilizing numerous decision trees, the random forest method proved its reliability and accuracy in predicting the likelihood of loan default. These models help lenders make well-informed decisions on loan approval by giving them a thorough grasp of the variables affecting loan defaults. These prediction models can help lenders better identify high-risk borrowers and reduce the possible financial losses brought on by loan defaults. Lenders can do this by integrating these models into their risk management systems.