ADVANCE DATA MINING & PREDICTIVE ANALYTICS

"PREDICTIVE MODELING FOR LOAN DEFAULT RISK ASSESSMENT"

GROUP-10 PROJECT REPORT

Submitted by:

Pravalika Girneni

Suraj Gadapa

Sravan sobhani

Pradeep reddy kethu

Team contribution:

Team members	Contribution
Pravalika Girneni	Data Exploration / Preprocessing / Probability of Default Model
Suraj Gadapa	Drafting of Project Report and PowerPoint Presentation
Sravan Shobani	Data Exploration / Preprocessing / Models / Prediction Files
Pradeep Reddy Kethu	Drafting of Project Report and PowerPoint Presentation

Index

Project Goal

The main objective of this project is to create a predictive model that can estimate the probability of a loan defaulting and the possible loss that can result from one. The model will consider both the likelihood of default and the potential severity of the losses brought on by default, in contrast to conventional finance-based techniques.

By minimizing risk to the financial investor, the project seeks to close the gap between traditional banking and asset management. The model will offer a more thorough and precise evaluation of the risk associated with a loan by taking into account both the likelihood and severity of default.

the project aims to improve decision-making processes for loan underwriters and asset managers, enabling them to make more informed and strategic decisions that minimize risk and maximize returns. The project may also have broader implications for the financial industry as a whole, potentially leading to more sophisticated and effective risk management practices.

Overview of data

The dataset user ID numbers and numerous factors to identify the clients rather than their real names. Due to our goal of projecting both the possibility of default and the severity of losses, one problem we ran into was the inadequate labeling of columns. This method combines asset management, which aims to minimize risk to the financial investor, with traditional banking, which places an emphasis on conserving economic capital. We were forced to rely only on data preparation and correlation analysis because the anonymous column titles precluded us from using any domain expertise. This necessitated the creation of the dataset.

The target variable in the dataset is called "loss" and it shows how much money the bank would lose if a loan went into default. Only the 'train_v3.csv' dataset contains the 'loss' variable; the 'test no_lossv3.csv' dataset does not. Since the goal is to forecast the "loss" variable for the test set, this is the case.

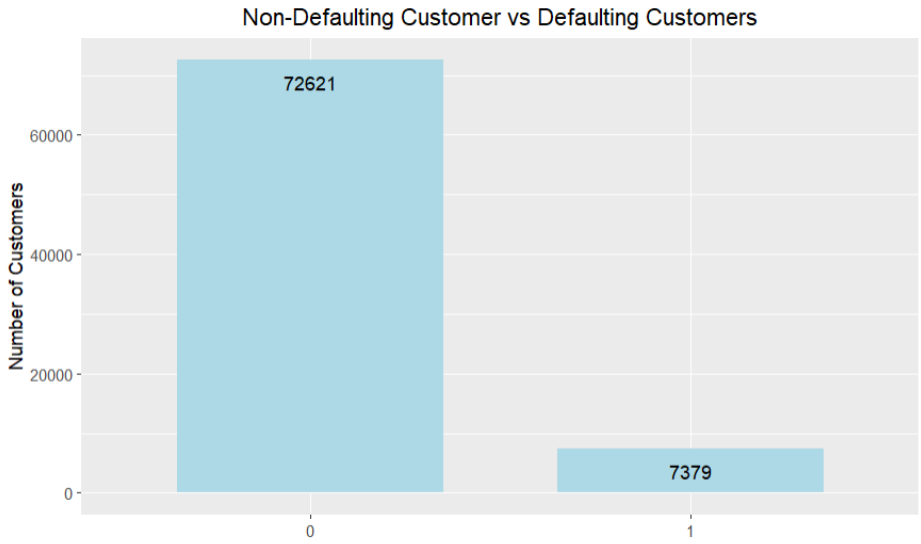
The information offers a complete picture of loan applications overall, and it may be used to forecast default rates and calculate the bank's prospective losses.

Data exploration analysis

We have done a number of data preparation operations and visualizations on a bank dataset.

- The code's initial operation is to build a "default" column from the "loss" column. The "default" column is set to 1, indicating that the client has defaulted, if the "loss" value is higher than 0. If not, 0 is entered in the "default" column. The "loss" column is also scaled by a factor of 100.

- We then searches the dataset for any missing values. The "rowMeans" function is used to compute the percentage of missing values in each row. It is discovered that the minimum and greatest percentages of missing values are 0% and 47%, respectively.
- We are doing dataset visualization after determining whether any values are missing. To compare the proportion of defaulting and non-defaulting consumers, a bar chart is made. The graph demonstrates that there are more non-defaulting clients than defaulters.



- After that, the dataset must be preprocessed by having zero-variance and strongly correlated variables removed. To find variables with zero variance, one uses the "nearZeroVar" function, and to exclude highly correlated variables, one uses the "corr" method in the "preProcess" function. Using the "medianImpute" method in the "preProcess" function, the values that are lacking are imputed. The final dataset, called "new_bank_model," comprises 248 properties.

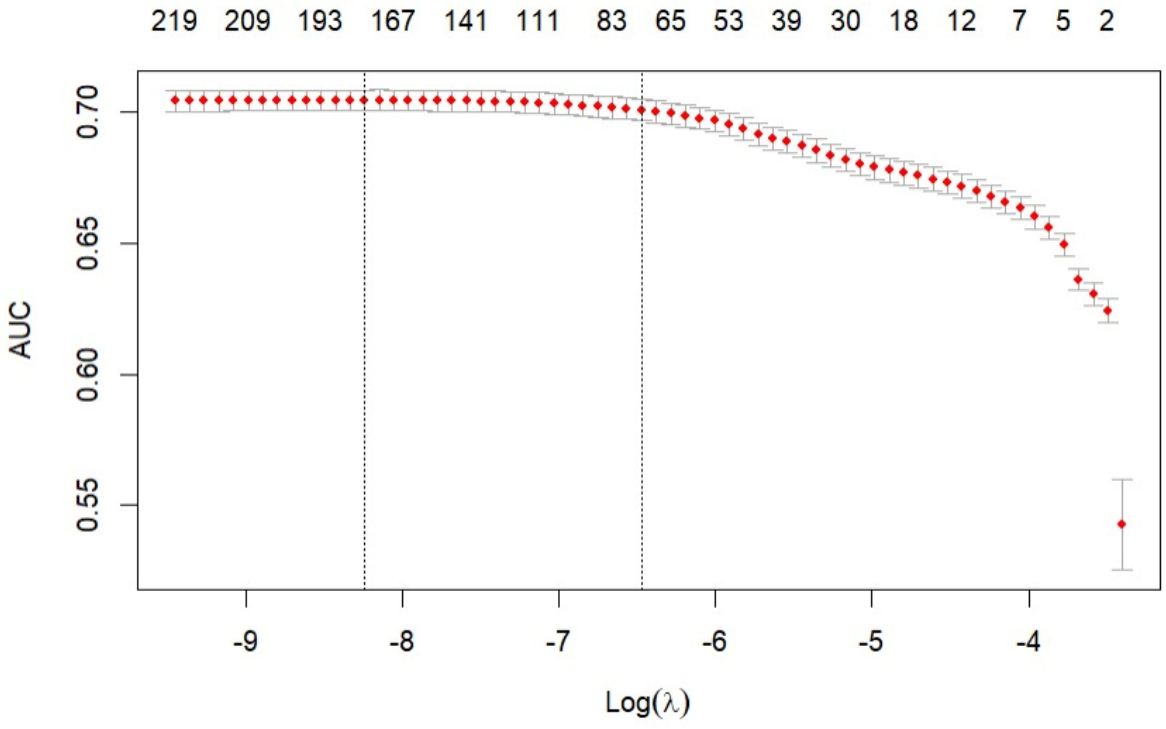
Model strategy

To estimate the likelihood that consumers would default on their loans, we are using a classification model. It selects variables using Lasso and Principle Component Analysis (PCA), trains a random forest model on the variables, and then uses the learned model on a test dataset.

In order to increase the precision and understandability of default rate projections for the bank, we used the LASSO regression analysis technique in this project to undertake variable selection and regularization. 248 variables served as the model's initial input. The regression method known as LASSO (Least Absolute Shrinkage and Selection Operator) penalizes the absolute magnitude of regression coefficients. According to their history of defaults, the bank will decide whether to allow or reject each customer in the framework of this initiative.

A Lasso model is first run in the pipeline to choose variables from the "new_bank_model" dataset. The remaining columns are utilized as predictors, and the "default" column serves as the target variable. The coefficients of less significant predictors are effectively reduced to zero by the regularization penalty that the Lasso model applies to the predictor coefficients. With the "cv.glmnet" function, 10-fold cross-validation is used to find the ideal level of regularization. The lambda value with the lowest cross-validation error is chosen, and the "coef" function is used to extract the coefficients that correspond to this lambda value. The final coefficients are put into a dataframe and ordered by decreasing magnitude.

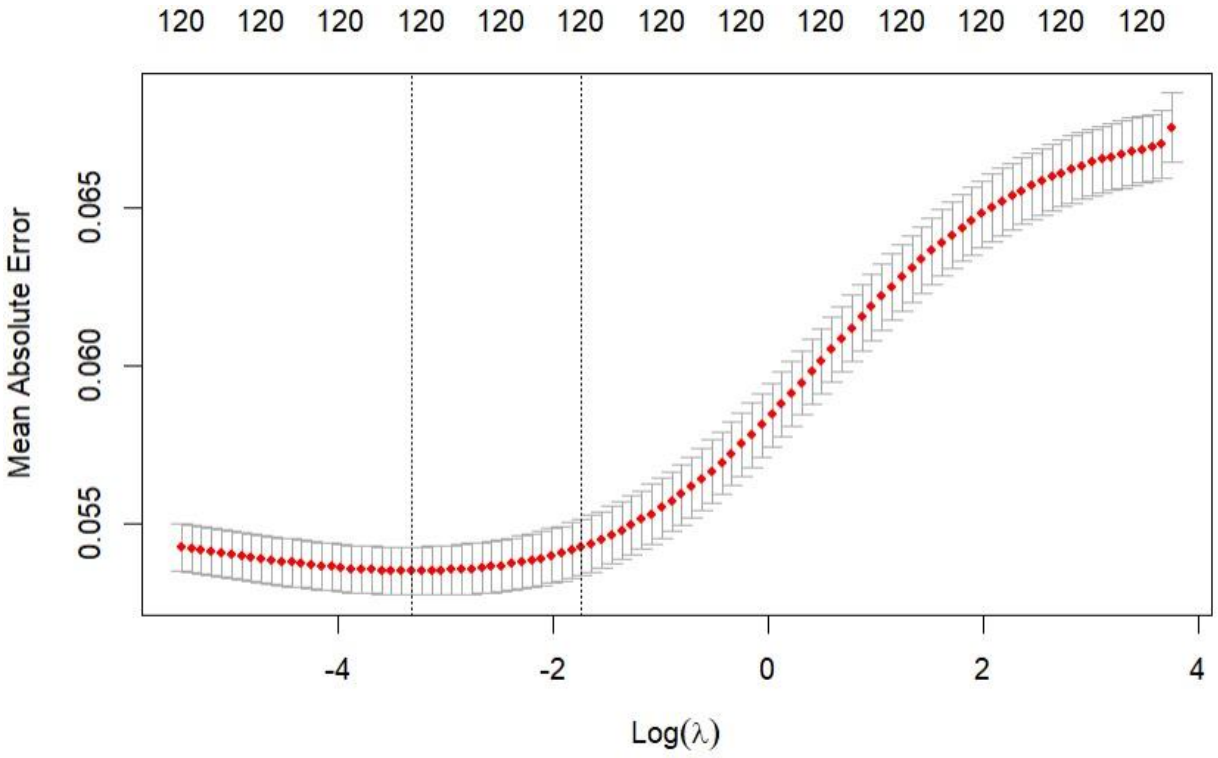
The "abs" function is used to make the negative coefficients positive, and the dataframe is rearranged without the intercept column. The "default" column is added to the vector created from the resulting dataframe of variable names. Finally, a new dataset called "bank_lasso" is created by using the "select" function to extract the columns from "new_bank_model" that contain the variable names that have been chosen.

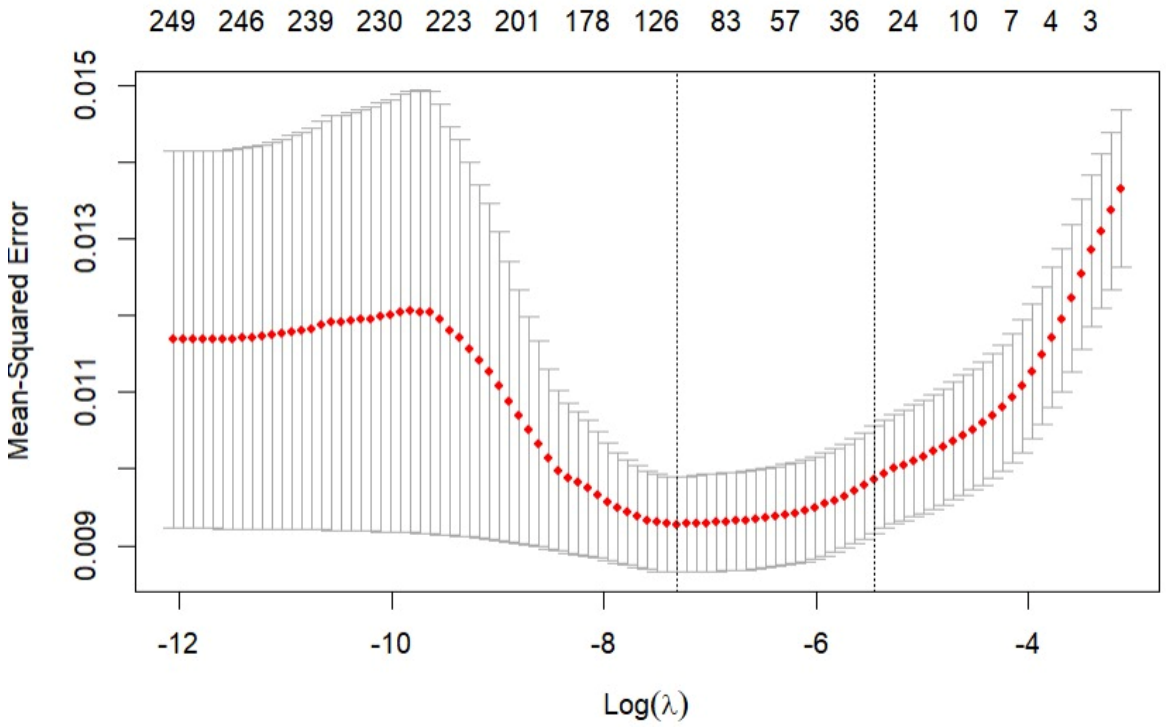


Model performance:

Ridge Regression for Loss Given Default(LGD Model)

We used ridge regression for the calculation of LGD and MAE(Mean Absolute Error) as a metric to measure the performance of the model.MAE for the LGD Model is 0.05, and the number of Lambda values tried were 100. Out of the 100 lambda values worked the best lambda was selected as 0.0348.





Conclusion and insight

Based on the "loss" column, a new column called "default" has been generated, with a value of 1 if the loss is more than 0, 0 otherwise. The dataset's missing values have been examined, and the highest percentage of missing values was discovered to be 47%. Using the "corr" and "medianimpute" procedures, zero-variance variables have been eliminated, strongly correlated variables have been handled, and missing values have been imputed. A new dataset containing 248 properties is produced after processing.

Lasso and PCA, two variable selection techniques, have also been employed to categorize the number of defaulting consumers. Out of the 248 attributes, the Lasso model returned a total of 180 attributes, and the coefficients were transformed into a data frame. The intercept columns returned by the Lasso model have been eliminated, and the data frame has been rearranged in decreasing order. A "default" column has been added to the data frame, and the data frame has been transformed into a vector. The original data set "new_bank_model"'s attributes have been chosen using the coefficients from the Lasso model.

In order to ensure that 80% of the variation is captured, the variables were further processed using PCA with a threshold limit of 0.80. PCA identified 69 components that represented 80% of the total. The PCA model now includes the "default" column from the previous model. From the

values produced by the PCA model, a train set and validation set have been created, and the "default" column has been transformed into a factor in both sets. Using the train set and PCA, a classification model has been created that will categorize the number of consumers who are defaulting.

References:

2017. An Introduction to XGBoost R package, Retrieve from:

https://www.r-bloggers.com/an-introduction-to-xgboost-r-package/amp/

2018. R-Random Forest, Retrieve from:

https://www.tutorialspoint.com/r/r_random_forest.htm

NCSS Statistical Software, Ridge Regression, Retrieve from:

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

2018, Introduction to Principal Components and FactorAnalysis, Retrieve from:

ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf

Stephanie, 2015, Lasso Regression: Simple Definition, Retrieve from,

https://www.statisticshowto.datasciencecentral.com/lasso-regression/