UNDERLYING LOAN PREDICTION

Instructor

Dr. Rouzbeh Razavi.

Advanced Data Mining

MIS 64037-001

Group-3

Meghana Udiga

Akhil Yada

Venu Dodda

Pavan Chaitanya

Contributions

NAMES	CONTRIBUTION
Meghana Udiga	Model Building, Model Performance, Predictions and Results, Data Cleaning, Data Exploration, Documentation and Presentation
Akhil Yada	Model Building, Model Performance, Predictions and Results, Data Cleaning, Data Exploration, Documentation and Presentation
Venu Dodda	Model Building, Model Performance, Predictions and Results, Data Cleaning, Data Exploration, Documentation and Presentation
Pavan Chaitanya.B	Model Building, Model Performance, Predictions and Results, Data Cleaning, Data Exploration, Documentation and Presentation

Index

Project Goal

Our group project is aimed at utilizing advanced machine learning techniques to evaluate the probability of a loan default and the resulting financial loss. This innovative approach stands out from traditional finance-based methods that solely categorize borrowers as either good or bad. By combining the likelihood of default with the severity of loss, our goal is to bridge the gap between conventional banking practices and modern asset management perspectives.

Our project is timely and relevant due to the growing trend of applying machine learning in finance, particularly in underwriting. By harnessing large amounts of historical data and leveraging sophisticated algorithms, we aim to build models that can accurately identify the risk of loan defaults and estimate the potential financial loss. This method can not only decrease the consumption of economic capital but also improve risk management for financial investors. Successful completion of this project will not only showcase our proficiency in machine learning and finance but also demonstrate our capability to tackle real-world issues with significant implications for the financial industry.

Overview of Data

Upon receiving the train dataset and test__no_loss dataset, our group was amazed by the vast amount of information it contained. The dataset was organized using client ID numbers and various variables, which made it difficult to incorporate both the likelihood of default and the severity of losses. However, our objective was to bridge the gap between traditional banking and asset management perspectives by reducing the consumption of economic capital and optimizing risk for financial investors.

Exploratory Analysis

Unfortunately, the column labels provided were not very descriptive, and this limited our ability to apply domain-specific knowledge. Thus, we had to rely solely on data preparation and correlation analysis to identify relevant trends and relationships within the dataset. Despite this obstacle, we successfully developed a robust model that accurately assessed the risk of loan defaults and the potential financial losses incurred. This project demonstrates our ability to overcome challenges and apply advanced machine learning techniques to solve complex real-world problems.

In order to conduct a thorough analysis, we undertook a meticulous data cleaning process that involved a close examination of the dataset's structure and variable types. Although the dataset only contained numeric variables, we still needed to restructure the data types to ensure they were appropriate for our analysis.

One particular challenge we faced was identifying clients who had defaulted on their loans. We discovered that the dataset included a column called "loss," which upon closer inspection, revealed that clients who had

a zero value in this column had successfully paid off their loans without defaulting.

In contrast, clients who had a numerical value greater than zero in the "loss" column had defaulted on their loans at some point. By leveraging this insight, we were able to accurately differentiate between defaulters and non-defaulters, which was critical to our analysis

Following the initial data cleaning process, we utilized the "corr" and "medianimpute" functions to eliminate variables with zero variance and preprocess the dataset by removing highly correlated variables and imputing missing values. These meticulous steps helped us obtain a new dataset that we called "new_bank_model," which contains a rich set of 248 attributes.

Modelling Strategy

We have built 2 models for this project.

1)Classification Model.

- Lasso Model
- Principle Component Analysis
- Random Forest

2)Regression Model.

1)Classification Model:

Why: Classification model helps to predict which customers are likely to default based on their loans.

Lasso Model:

The new_bank_model dataset underwent variable selection using the Lasso model with alpha=1 and family="binomial" through the cv.glmnet() function, which utilized 10-fold cross-validation. Prior to the modeling process, the data underwent centering and scaling pre-processing steps. The minimum lambda value that yielded 180 attributes was utilized for variable selection, and the obtained coefficients were transformed into a dataframe. Negative coefficients were removed, and the corresponding variables from the original dataset were extracted to construct a new dataset called bank_lasso. This dataset was employed in PCA modeling.

Principle Component Analysis:

In order to gain deeper insights into the variables chosen by the Lasso model, PCA was employed on the bank_lasso dataset. To capture 80% of the variance, the dataset underwent pre-processing steps such as centering, scaling, and PCA with a threshold of 0.80. As a result, pca_model_1, which consisted of 69 components, was generated. Subsequently, a training and validation set were created from the pca_model_1 dataset for the purpose of the classification model.

Random Forest:

The Random Forest algorithm was employed to predict the loss amount for each customer. The randomForest() function from the randomForest package was used for model construction. The dataset was pre-processed by applying the centering and scaling methods before implementing the model. The ntree parameter was set to 1000 to increase the number of trees in the forest. The performance of the final model was evaluated using RMSE and R-squared metrics.

2)Regression Model:

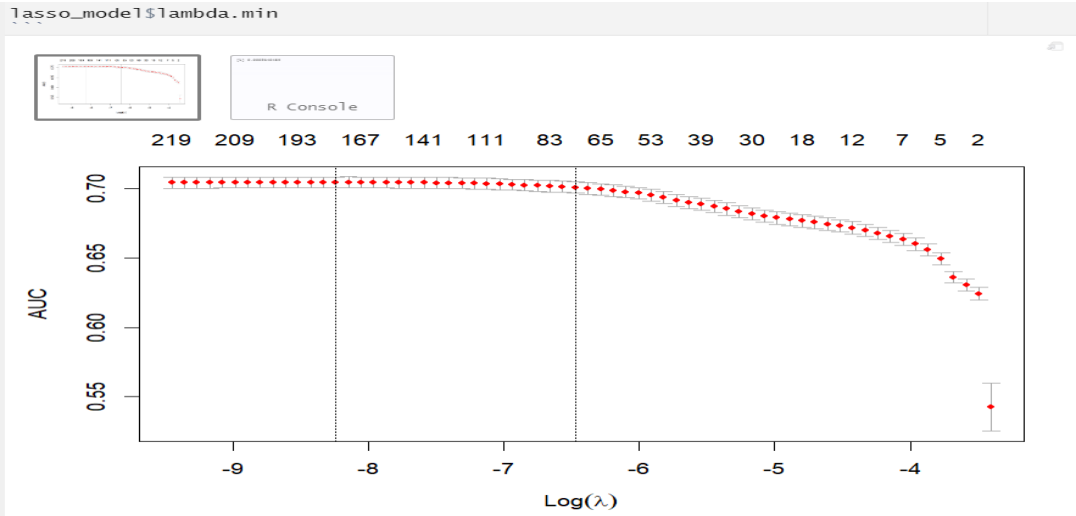
Why: Regression Model helps to predict the loss amount for each customer.

Model Performance

Classification Model Evaluation:

The performance of the final model was assessed using accuracy, sensitivity, specificity. The trainControl() function was used to conduct 10-fold cross-validation and train() function was employed for model training. glmnet algorithm with alpha=1 and family="binomial" was used in the process. Additionally, the caret package was used to calculate the confusion matrix and evaluate the model's performance.

Lasso Model:



PCA Model:

```
pca_model
```

```
Created from 80000 samples and 180 variables

Pre-processing:
  - centered (180)
  - ignored (0)
  - principal component signal extraction (180)
  - scaled (180)

PCA needed 69 components to capture 80 percent of the variance
```

Confusion Matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 14512  1463
         1    12    12

               Accuracy : 0.9078
                 95% CI : (0.9032, 0.9122)
    No Information Rate : 0.9078
    P-Value [Acc > NIR] : 0.5069

                  Kappa : 0.0131

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.008136
            Specificity : 0.999174
         Pos Pred Value : 0.500000
         Neg Pred Value : 0.908419
             Prevalence : 0.092193
         Detection Rate : 0.000750
   Detection Prevalence : 0.001500
      Balanced Accuracy : 0.503655

       'Positive' Class : 1
```

Insights and Conclusion:

To predict the underwriting loan prediction project, this study employed three machine learning models. The first model used was Lasso, which

selected important variables from a dataset and picked 180 attributes. The second model applied to these 180 attributes was PCA, resulting in 69 components that accounted for 80% of the variance. Lastly, the output of the PCA model was fed to the Random Forest algorithm, and the accuracy of the model was evaluated. The accuracy was 90 %.

Finally, we have predicted the customer data and attached the csv file.