ANALYSIS OF THE

BANK LOAN DATA

REPORT

GROUP 02

Kavindi Chamathka	s16367
Lakmini Thamodya	s16340
Maheesha Sewmini	s16349
Poornima Tharangani	s16368
Kavina Dias	s16200

Abstract

Banks frequently face the challenge of loan defaults, which is an unavoidable issue. If loans are not repaid, banks experience financial losses. To minimize this problem, they aim to keep default rates as low as possible. In that case, here, a thorough descriptive analysis and advanced analysis were conducted using MATLAB software on the bank loan dataset. This analysis aims to identify potential future default customers on bank loans, analyzing factors like credit history, income, employment duration, debt-to-income ratio, other debts, etc.

Table of Contents

Abstract 1
Table of Contents 1
Objective of the Analysis 1
Description of the Dataset 1
Data Pre-Processing 2
Descriptive Data Analysis 2
Advanced Data Analysis 7
Conclusion 10
Appendix 10
References 10

Objective of the Analysis

The analysis focuses on identifying the customers who have a high chance of failing to pay (default) in the future. We aim to explore the factors that are associated with customers who have defaulted on their payments in the past. Also, we will apply clustering techniques to segment customers into homogeneous groups, focusing on identifying those with a higher likelihood of default in the future.

Description of the Dataset

This dataset consists of 1500 records of 12 variables.

Variable No	Variable Name	Variable descriptions	Description of categories
1	branch	Branch	
2	ncust	Number of customers	
3	customer	Customer ID	
4	age	Age in years	

5	ed	Level of education	1 – Did not complete high school 2 – High school degree 3 – Some college 4 – College degree 5 – Post/Under-graduate degree
6	employ	Years with current employer	
7	address	Years at current address	
8	income	Household income in thousands	
9	debtinc	Debt to income ratio (x100)	
10	creddebt	Credit card debt in thousands	
11	othdebt	Other debt in thousands	
12	default	Previously defaulted	0 – No 1 - Yes

Table 1

Data Pre-Processing

We did not find any missing values or duplicate observations in the Bank Loan dataset. Therefore, there was no need to remove or impute missing data. The **ncust (number of customers)** and **customer ID** variables were omitted to improve the effectiveness of the analysis. Next, the dataset was split into **training (80%)** and **test (20%)** sets. The training set contained **1,200 observations**, while the test set had **300 observations**. Analysis was performed using the training set.

Descriptive Data Analysis

The objective of **identifying customers who have a high chance of defaulting** (failing to pay in the future) is a fundamental problem in **credit risk analysis**. The goal is to analyze the **patterns and characteristics of customers who previously defaulted** and use this information to determine **risk factors** associated with defaulting in the future.

Identifying Key Risk Factors

Analyze how different variables are associated with default:

- Does **low income** increase the risk of default?
- Does **higher debt-to-income ratio** indicate financial stress?
- Are **younger or older customers** more likely to default?
- Do **certain education levels** have higher default rates?

This helps in developing **risk profiles** for future prediction.

Default Status

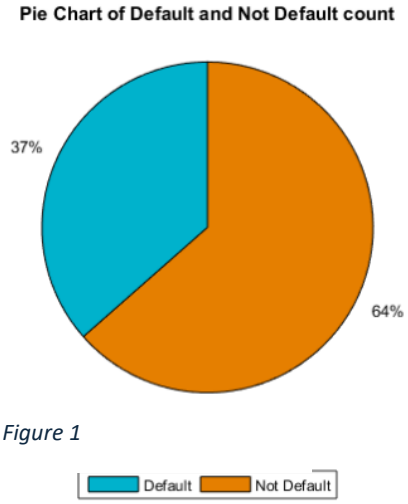


Figure 1

Demographic Behavior

Age vs Default Status

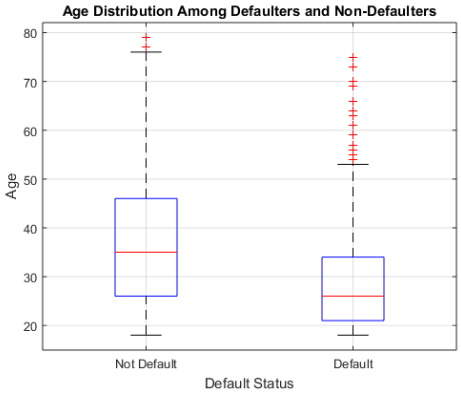


Figure 2

Default Status has been selected as the target variable. As in the figure 1 Pie chart, No of individuals who has not previously defaulted (Default=0) is more than who has defaulted.

Based on the results in Figure 2, the age range of the non-defaulted group (20–35) is wider than that of the defaulted group (27–45). The median age of individuals who defaulted is lower than that of those who did not default. This suggests that younger individuals are at a higher risk of defaulting. Therefore, age may be significantly associated with default status.

Level of education vs Default Status

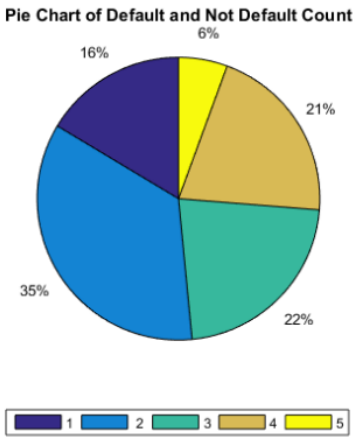


Figure 3

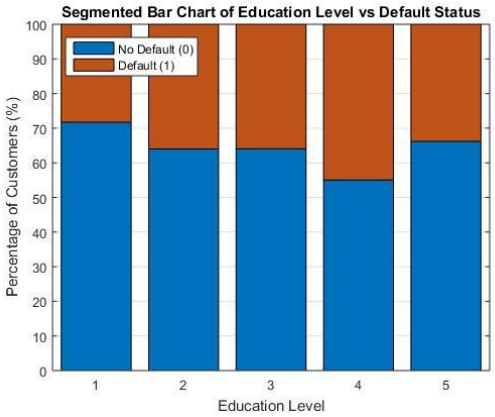


Figure 4

In the given dataset, most individuals did not complete high school, resulting in an imbalance across education levels. Due to these unequal proportions, it is difficult to determine the exact relationship between education level and default status. The segmented bar chart shows the distribution of defaulted and non-defaulted customers across different education levels. Overall, the percentage of individuals who have not defaulted is similar across all education levels.

Years with current employer vs Default Status

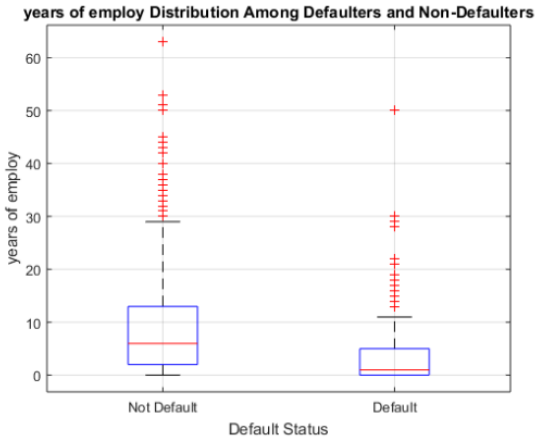


Figure 5

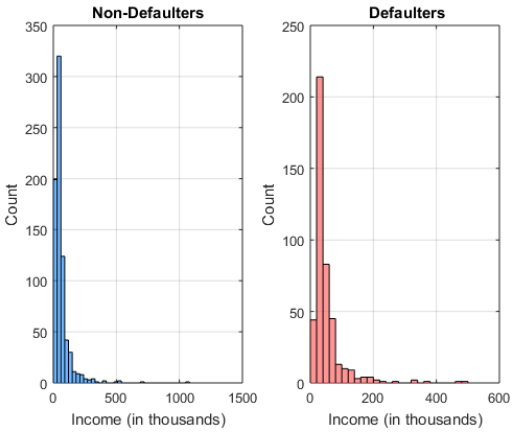


Figure 6

According to the boxplots in Figure 5, the range of years with the current employer is wider for the non-default group compared to the default group. This suggests that individuals with fewer years at their current job have a higher risk of defaulting on loans.

Years at current address vs Default Status

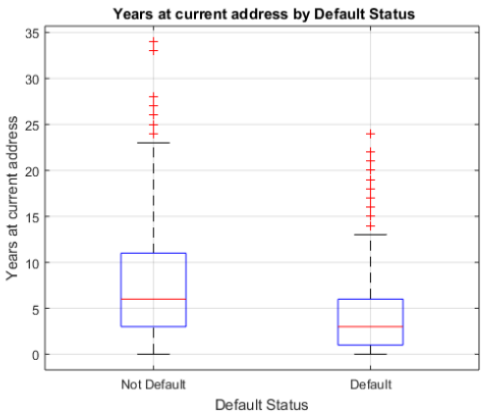


Figure 7

Figure 7 illustrates the relationship between default status and years at the current address. The analysis shows that individuals who have lived at their current address for a longer period have fewer previous defaults. In contrast, those who have recently moved are more likely to default on loans.

Financial behavior - Income vs Default Status

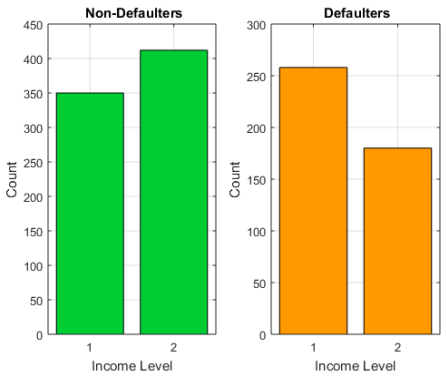


Figure 8

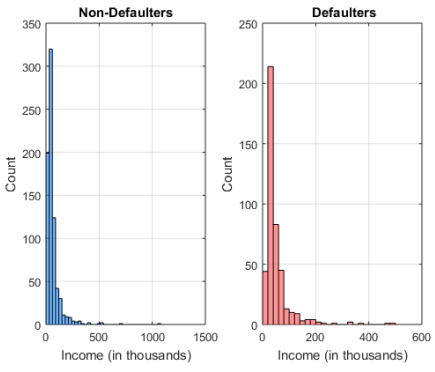


Figure 9

The figure8 shows how the income is distributed with two groups non-defaults and defaults. We cannot get a clear idea from the two histograms (figure 9). So, we have checked how the income level distributed in two groups. It is clear that Most of the individuals in low income level has high risk to be defaulted.

Debt to income ratio vs Default Status

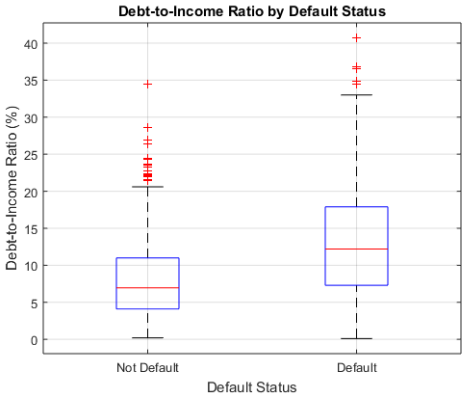


Figure 10

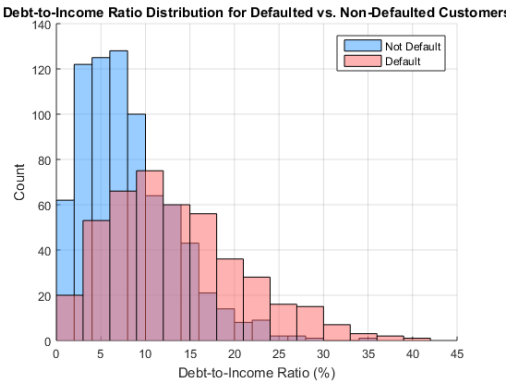


Figure 11

In the figure 10, individuals with previous defaults have wider range of debt to income ratio. When we examine the median measure, most of individuals with high debt to income ratio has previous defaults.

In the overlapped histogram in figure 11, it is clear that the debt to income ratio distribution for individuals with previous defaults is lower than who have not previous defaults. Average debt to income ratio is low in group with previous defaults. Individuals with low debt to income has more like hood to be defaulted. So, debt to income may be significantly associated with default status.

Total Debt vs Default Status

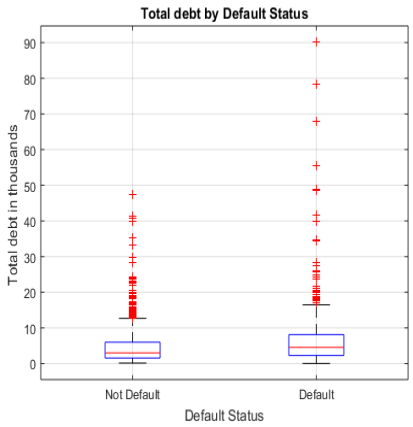


Figure 12

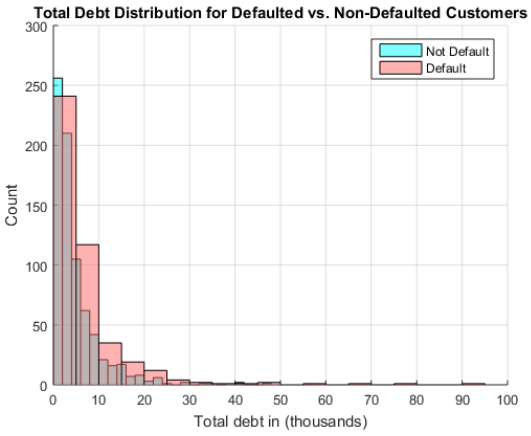


Figure 13

Figure 12 shows no clear difference in total debit between individuals with and without a previous default. Similarly, the histogram does not indicate any noticeable distinction between the two groups.

Correlation heatmap

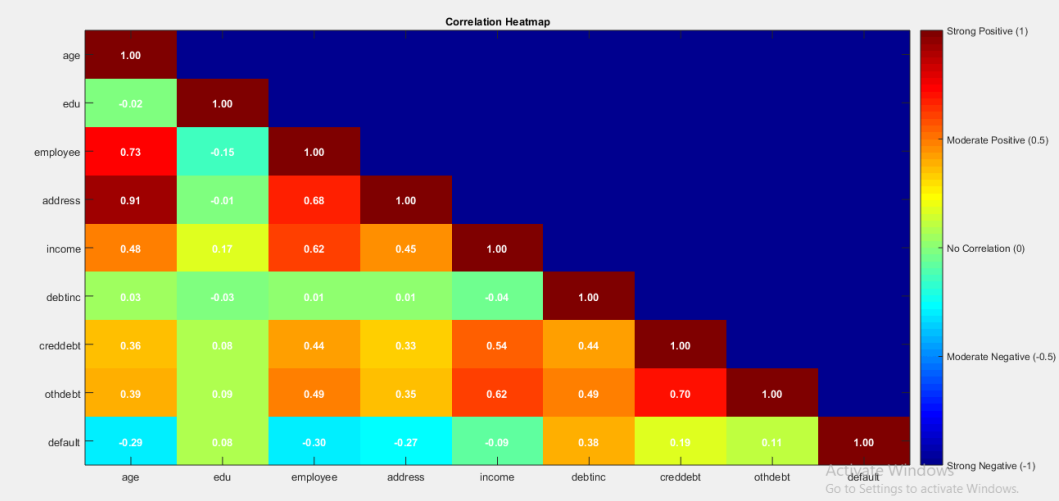


Figure 14

Figure 14 (the heatmap) shows the correlation between different predictor variables and the response variable (default status). It helps identify multicollinearity and how each predictor variable is related to the response variable. The **blue color** represents negative correlations, while the **red color** indicates strong positive correlations. The heatmap highlights a high correlation of **0.91 between address and age**, suggesting possible multicollinearity between these two variables.

Correlation between predictor variables and default Status

We conduct spearman correlation test and all the variables be significant at both 0.05 and 0.01 levels. Here are the results,

Variable	Spearman Correlation coefficient
Age	-0.2958
Edu level	0.0806
Employ	-0.3345
Address	-0.2917
Income	-0.1451
Debt to income	0.3603
Total debt	0.1692

Table 2

- Age (-0.2958): Age has mild negative association with default status
- Employ (-0.3345): Years with current employer has moderate negative association with status.
- Address (-0.2917): Years at current address has low negative association.
- Income (-0.1451): It has negative association, means when income increased the risk of be default decreased.

All the above factors have negative association with default status. Means when increasing these factors like hood of being defaulted decreased.

- Edu level (0.0806): Education level has very small positive association with default status.
- Debt to income (0.3603):it has moderate linear relationship with default status.
- Total Debit (0.1692): It has small linear association

Advanced Data Analysis

Classification Tree

The decision tree was implemented in order to identify the variables that were associated with previously default status. Firstly, a decision tree was fitted to the train set considering the prune criterion 'Impurity' as the train set was imbalanced and 'Impurity' approach considers class proportions therefore it accounts for how mixed the classes are in each node, helping the tree maintain sensitivity to minority classes.

Original Tree Without Pruning

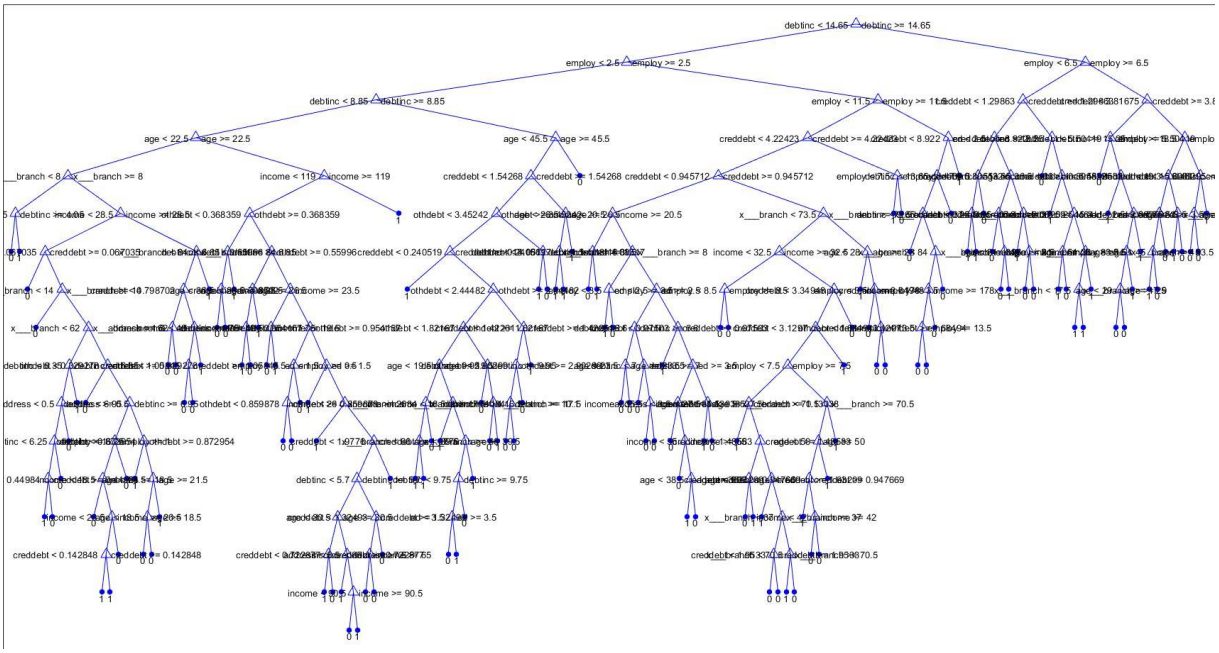


Figure 15

The train error was 0.0917 but the cross-validation error was 0.3292 and the test error was 0.3533. The above values suggest that overfitting might have occurred and looking at the complexity of the tree, a deep tree with many leaves tends to overfit the training data, as it tries to capture the noise in the data and does not generalize well to new data. This means that this tree's accuracy on the test set would be less. Therefore, cross validation loss was used to prune the above tree to get the most optimal tree. By looking at the plot of cross validation error vs prune level it was identified by the algorithm itself that the most prune level occurs at the 79 prune level under the 1 SE rule from kfoldloss algorithm.

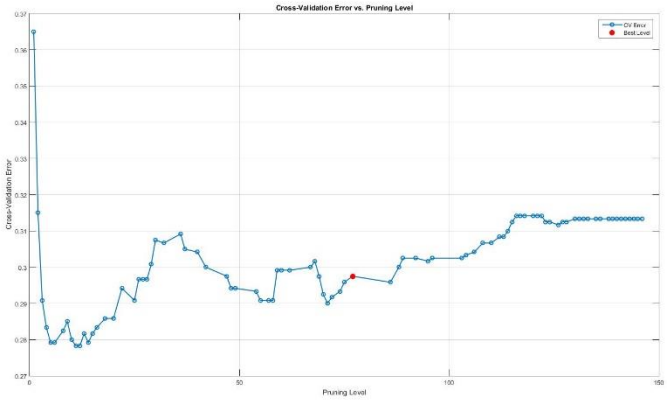


Figure 16

The new tree after pruning,

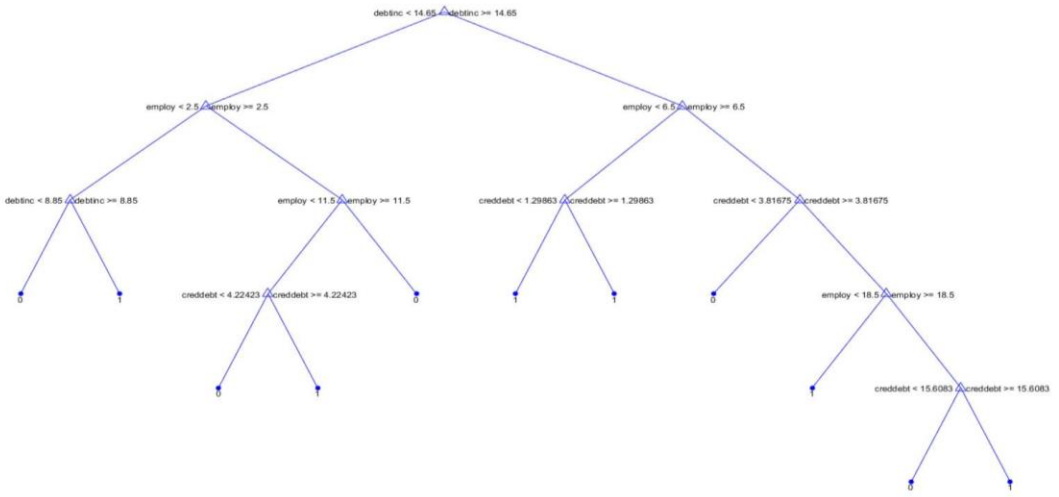


Figure 17

The training error was 0.2274 but the cross-validation error was 0.32 and the test error was 0.2767. Then, by using the variable importance plot the important predictors were selected. In this dataset the most important variables are:

- employ
- debtinc
- creddebt

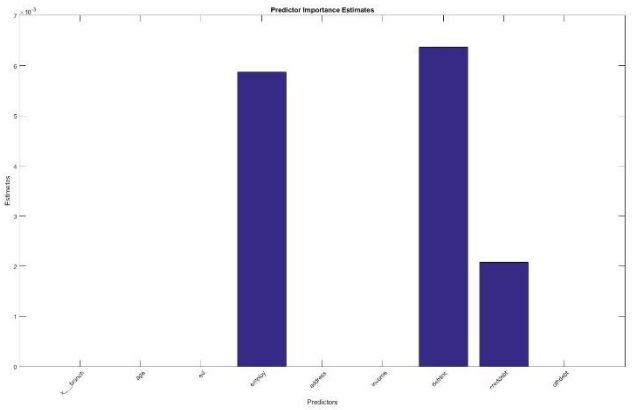


Figure 18

PNN (Probability Neural Network)

A probabilistic neural network (PNN) is a sort of feedforward neural network based on radial basis function used to handle classification and pattern recognition problems. Error is the same across all spread parameter values that we used. So we can get accuracy as 63.33% and error rate as 36.67%

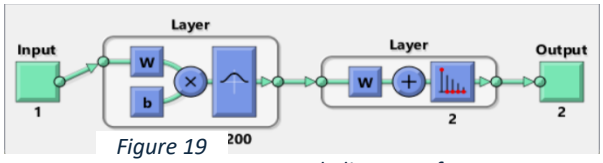


Figure 19 Network diagram of PNN

KNN

The KNN algorithm is a distance based powerful classification algorithm. The optimal K value is 7. Through this optimal model can be obtained and it gives accuracy as 62.00%

Cluster Analysis

Grouping customers into homogeneous clusters helps to identify shared characteristics and highlight the segment with a higher likelihood of default. In this study, cluster analysis was conducted to classify customers based on their financial and demographic attributes. The k-means clustering algorithm was applied, using the silhouette scores across different values of k to determine the optimal cluster structure.

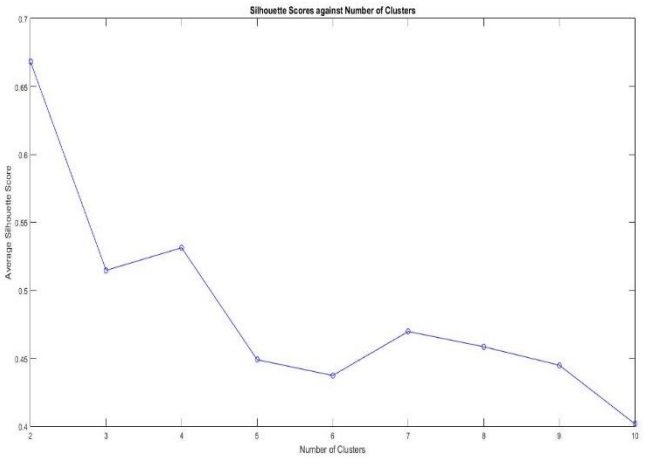


Figure 20

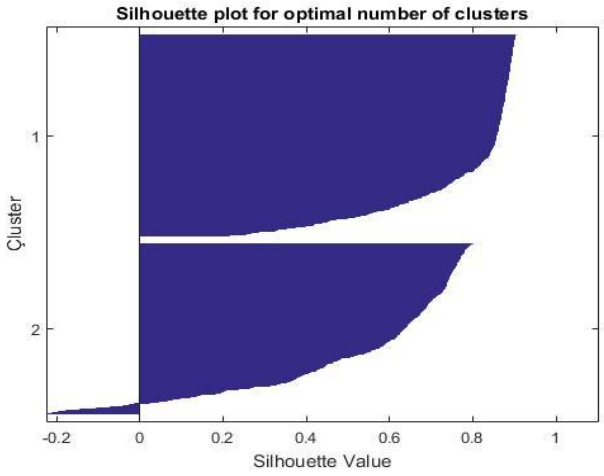


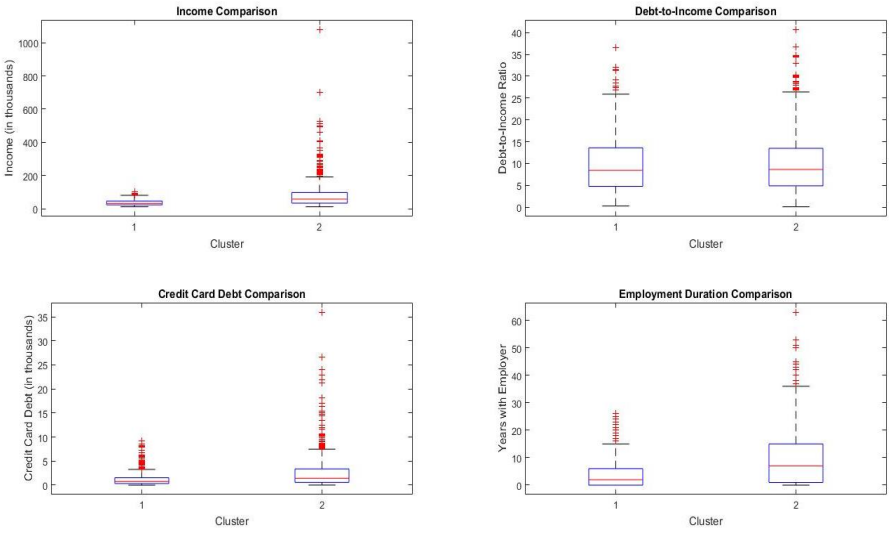
Figure 21

Figure 20 suggested that the optimal value of k is 2, as it achieved the highest silhouette value which is 0.6680. Therefore, the optimal number of clusters was determined to be two. In this Silhouette plot in Figure 21, for the two clusters, most data points have positive silhouette values, giving a well-formed clustering structure, though a few points in cluster two indicate some misclassifications. Overall, the clustering quality appears strong, supporting the decision to use two clusters.

Default Percentages	
Cluster 01	40.31%
Cluster 02	32.00%

Table 4

When the default rates were calculated in each cluster, it was found that cluster 1 had the highest percentage of default rates, indicating that this cluster might consist of a higher risk group. The characteristics of the claims within cluster 1 were then compared to identify any common features that may help identify the customers who have a high chance of failing to pay.



The boxplots here compare financial and demographic characteristics such as income, debt-to-income ratio, credit card debt and employment duration across two clusters, giving the distinct differences. Cluster 1 shows lower income, shorter employment compared to cluster 2 suggesting that cluster 1 is more financially vulnerable and at higher risk of default. But the debt-to-income ratio is quite similar in both clusters and cluster 1 has lower credit card debt.

Figure 22

Variable	Cluster 01 Median	Cluster 02 Median
age	28	38
address	3	7

othdebt	1.7684	3.2898

Table 5

This table compares the median values for the other different variables between the two clusters. Here, cluster 1 may indicate a greater default risk with less financial stability. Therefore, the identified common characteristics of cluster 1 were, lower age, lower number of years with current employer, lower number of years at current address, lower income, lower credit card debt and other debts.

Comparing Common Characteristics using median

	age	employ	adderess	income	credebt	othdebt
Cluster 01	28	2	3	33	0.77324	1.7684
Previously Defaulted Customers	26	1	3	35	1.3682	2.8529

Table 6

When comparing the common characteristics identified in cluster 01 with the customers who previously defaulted, age, employ,address and income reveal quite similarity but credebt and othdebt show littlebit lower values in cluster 01.

Conclusion

This analysis provided valuable insights into the factors influencing bank loan defaults. By examining key demographic and financial variables, we identified patterns that distinguish high-risk customers from low-risk ones. Our findings suggest that younger individuals, those with lower income levels, shorter employment durations, and higher debt-to-income ratios are more likely to default on loans. Advanced analytical techniques, including Probabilistic Neural Networks (PNN), K-Nearest Neighbors (KNN), decision trees and clustering methods, further reinforced these findings. The clustering analysis revealed two distinct customer segments, with one group exhibiting significantly higher default risk. Overall, the insights from this study can aid financial institutions in refining their loan approval processes, developing targeted risk mitigation strategies, and improving credit risk management. By focusing on high-risk customer segments, banks can implement proactive measures to reduce default rates and maintain financial stability.

Appendix

The dataset and all matlab codes used are available in google drive:

https://drive.google.com/drive/folders/1LItInzU5_gZsnjt74wHXMsZqIfkxfhHh?usp=drive_link

References

https://www.mathworks.com/help/stats/k-means-clustering.html

https://en.wikipedia.org/wiki/Mahalanobis_distance

https://www.mathworks.com/matlabcentral/fileexchange/3596-pearson-chi-square-hypothesis-test

https://www.mathworks.com/help/stats/improving-classification-trees-and-regression-trees.html