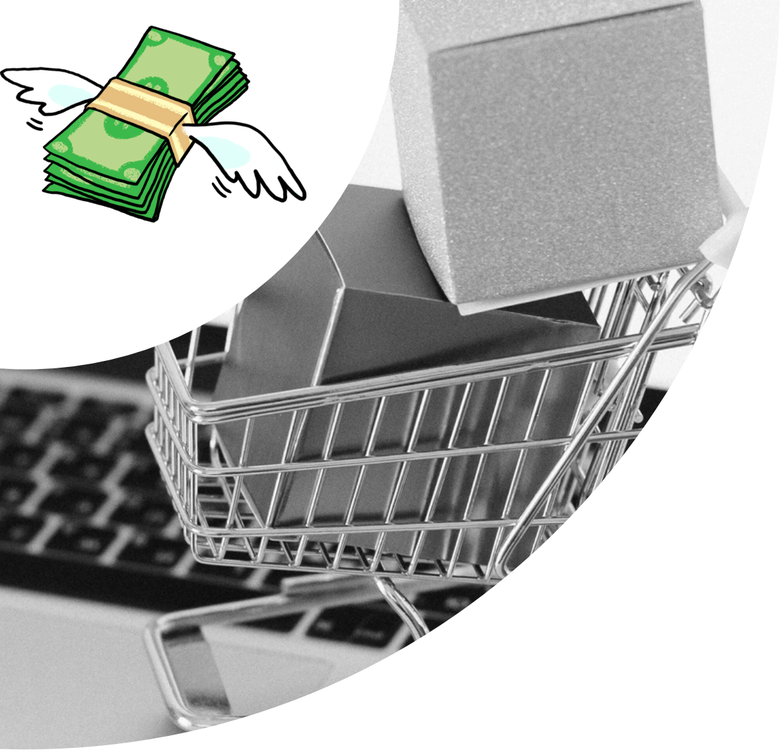
ANALYSIS of Bank Loan Data

Group 2

Introduction

- Loan defaults represent a significant risk for banks, leading to financial losses.
- Identifying customers with a high likelihood of default is crucial for minimizing these risks.
- This analysis focuses on predicting future loan defaults by exploring common characteristics of customers.



Objective

Identifying the customers who have high chance of failing to pay in the future.

Description of the Dataset

1500 records

12 variables

2 categorical variables
10 Numerical 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

Preprocessing

No Missing Values
No Duplicate Records

The dataset was divided into training (80%) and test (20%) sets
The training set - 1,200 observations
The test set - 300 observations.

The ncust and customer ID variables were omitted for better analysis.

Spotting Outliers

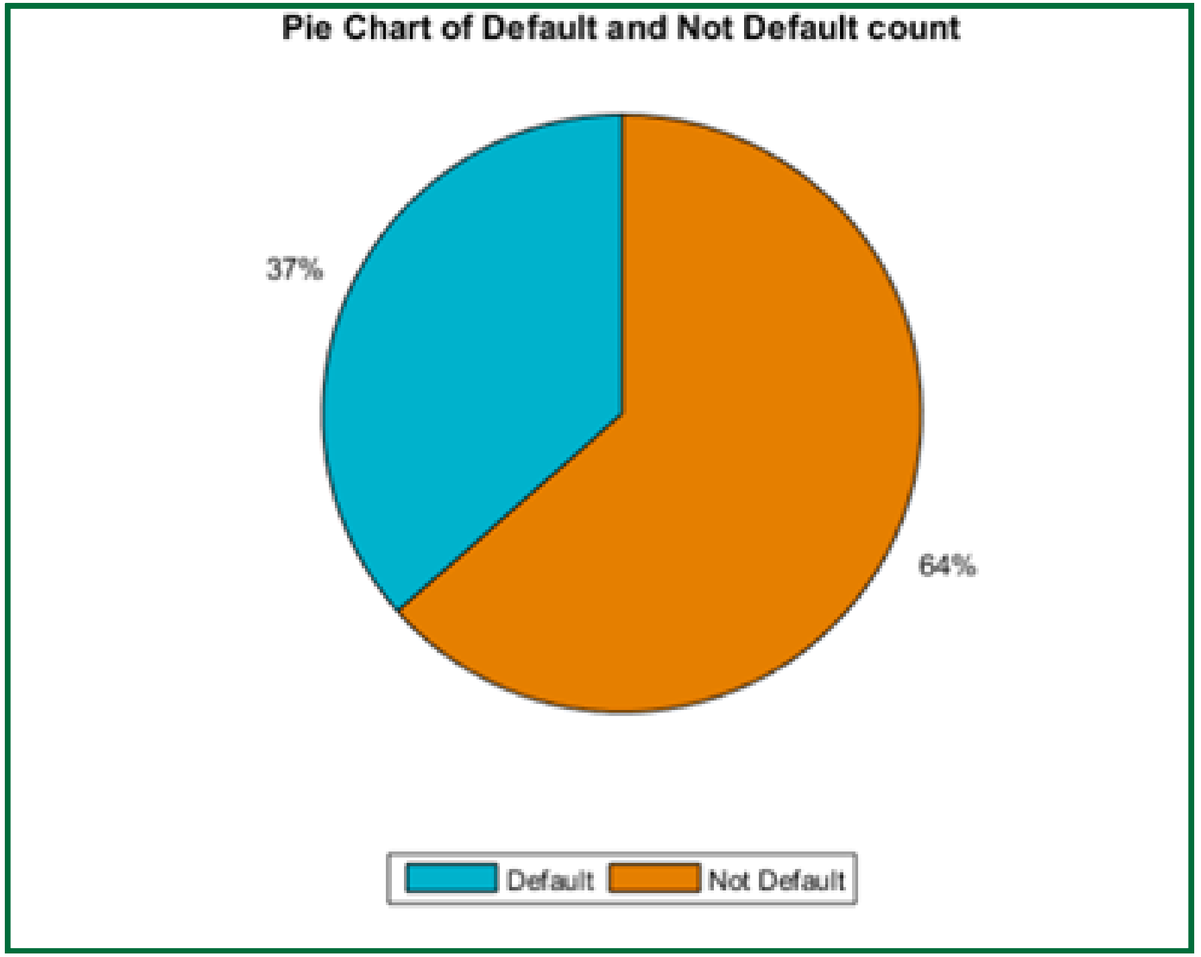
Detecting outliers in multivariate data

The technique : **Mahalanobis Distance**
Mahalanobis Distance (MD) is a way to measure how far a data point is from the center (mean) of a dataset while considering correlations between variables.

The Mahalanobis distance analysis identified **100 outliers** out of 1200 training dataset.(8.33%).

Descriptive Analysis

Default Status



Most of the individuals in the dataset are without previous defaults

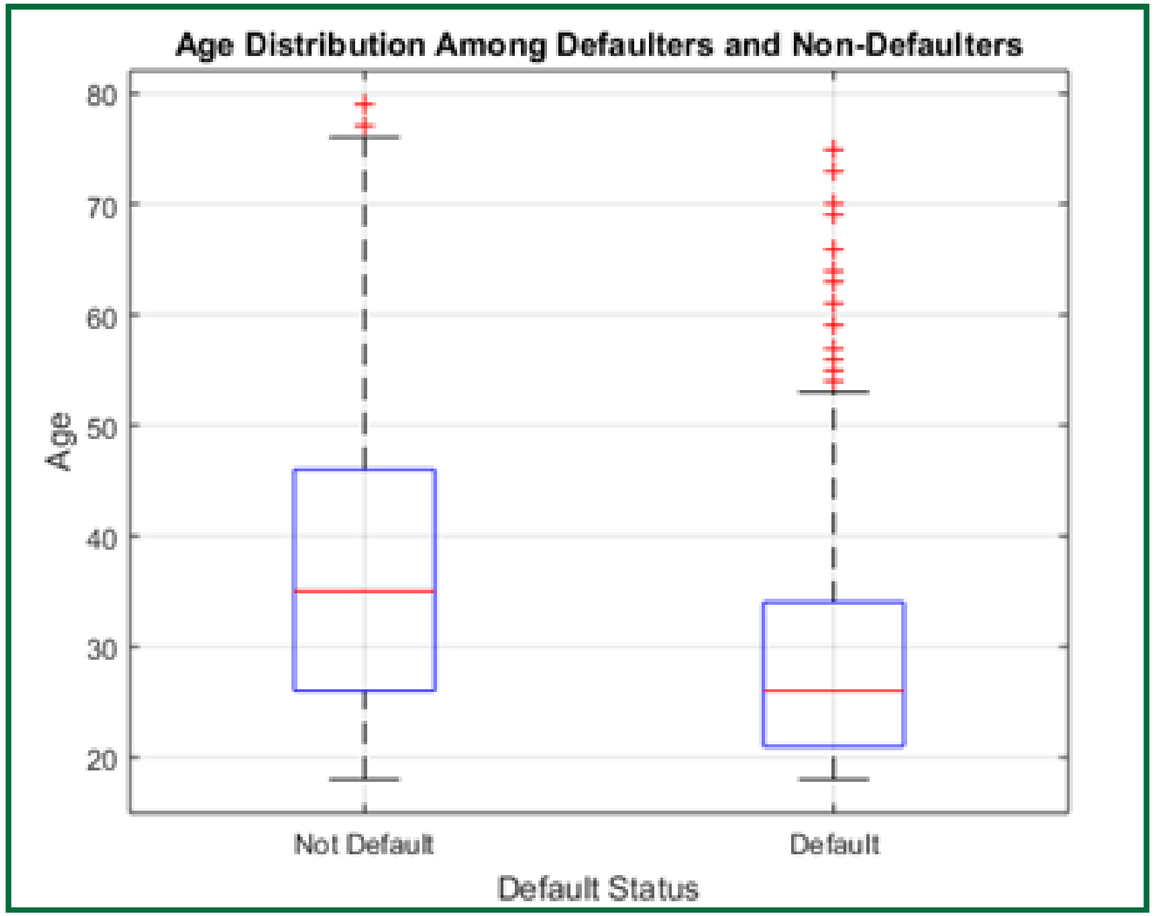
Age vs Default Status

- Median age of defaulters is lower than that of non-defaulters

It means Younger individuals have more risk to be default.

More outliers in the default group

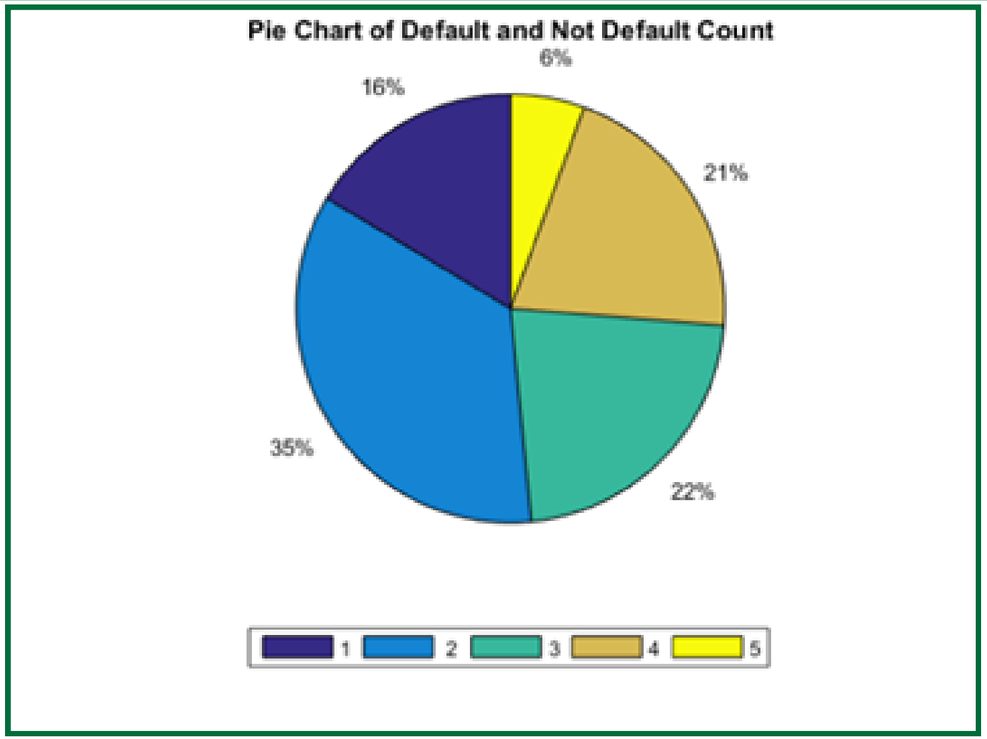
Very old individuals are disproportionately likely to default.

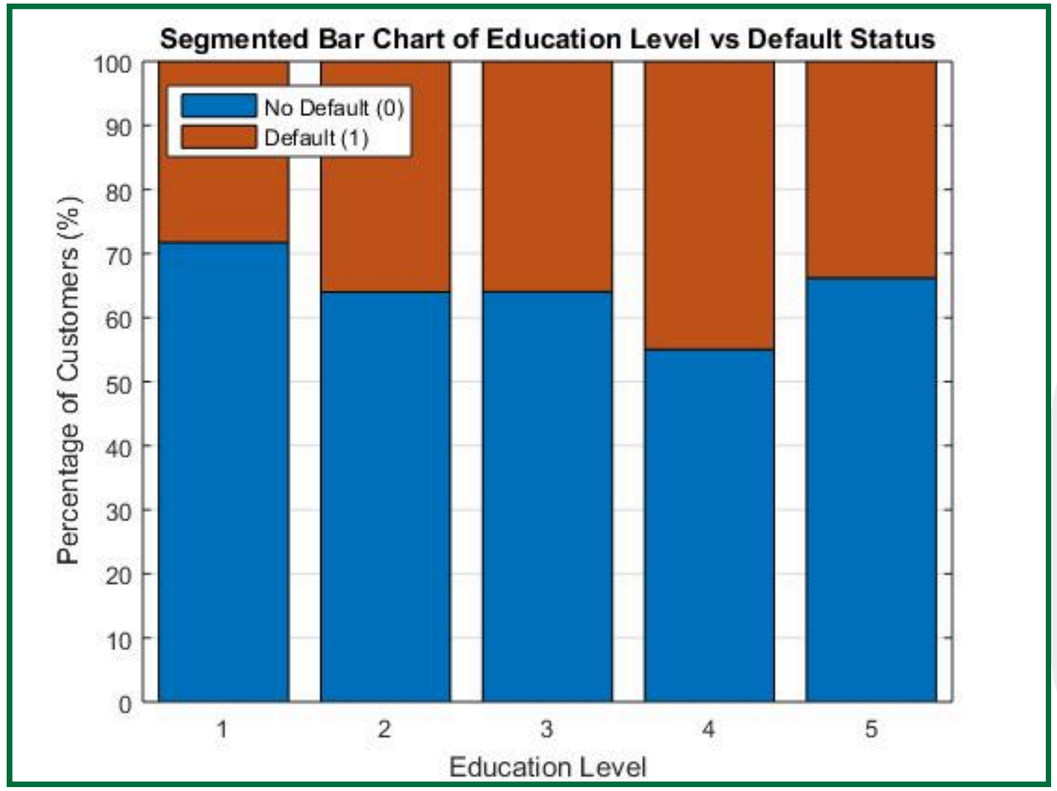


Edu Level vs Default Status

Most individuals did not complete high school, resulting in an imbalance across education levels.

It is difficult to determine the exact relationship between education level and default status.





The percentage of individuals who have not defaulted is similar across all education levels.

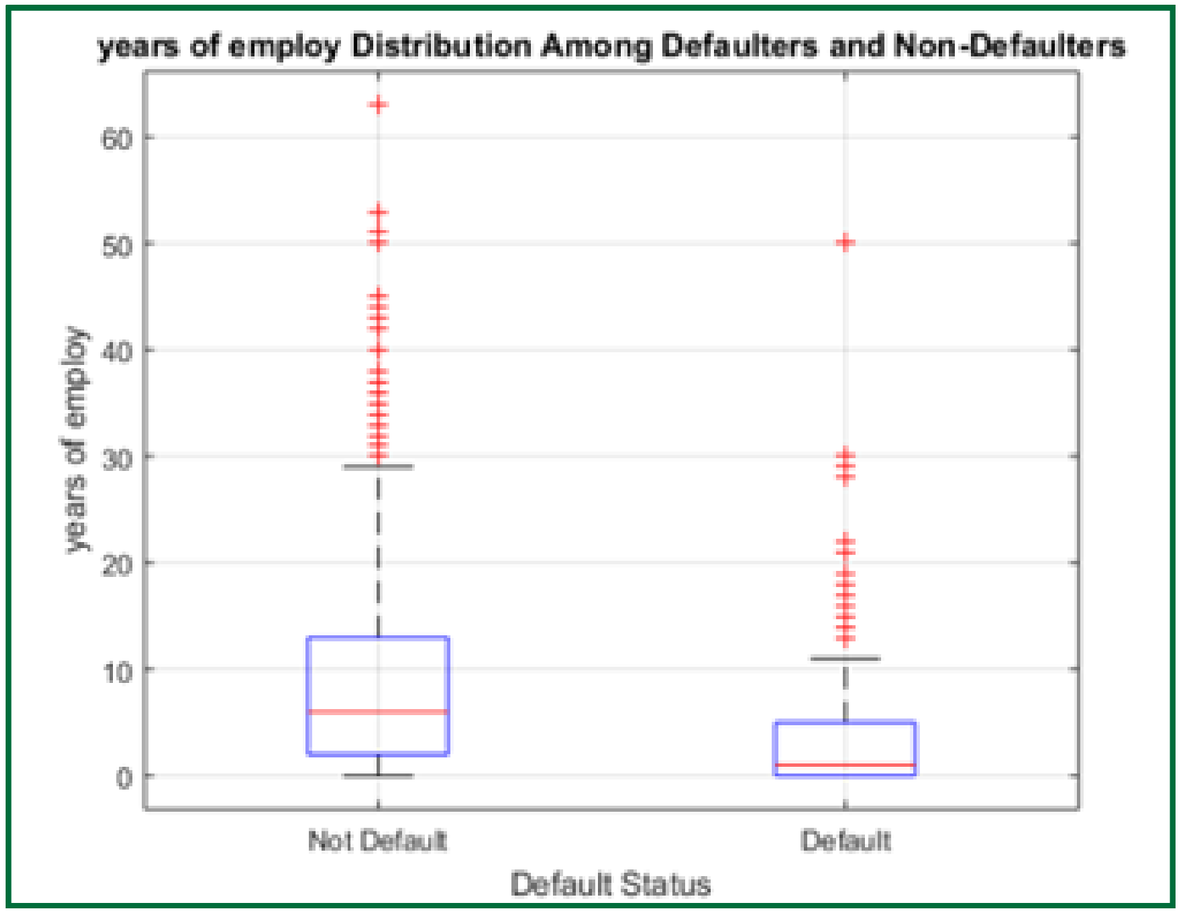
🔎 **Possible Explanation**

This may due to the Imbalance of the data

Years with current employer vs Default Status

- Median years of employment for the default group is at the minimum level of the non-default group.

  People who default tend to have shorter job tenure than those who do not default.
- Shorter employment duration may indicate **job instability** or **lower financial security**, increasing the risk of default.



The presence of many outliers

Some individuals in both groups have very high employment years but still default.

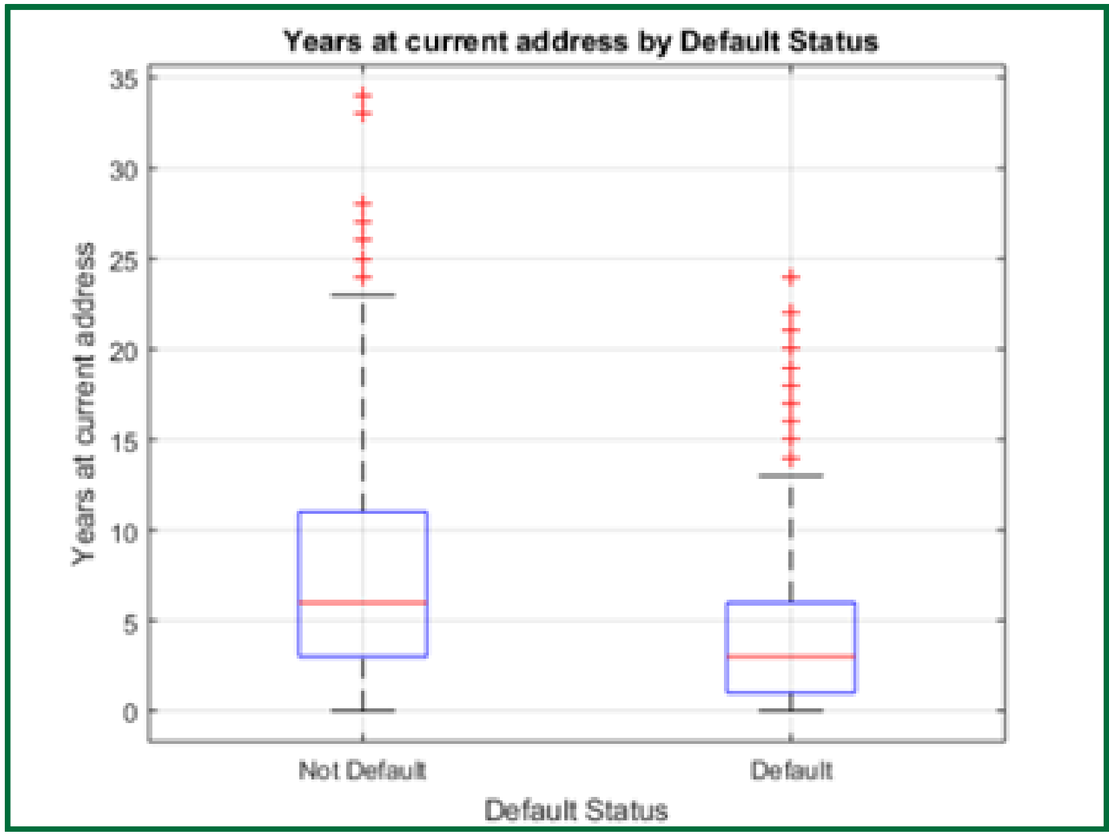
🔎 Possible Explanation

High-income individuals may still default due to other financial burdens...

Years at current address vs Default Status

- The analysis shows that the Individuals who have lived at their current address for fewer years tend to default more often.

Residential stability might be linked to financial stability, as frequent movers may have **unstable financial situations.**

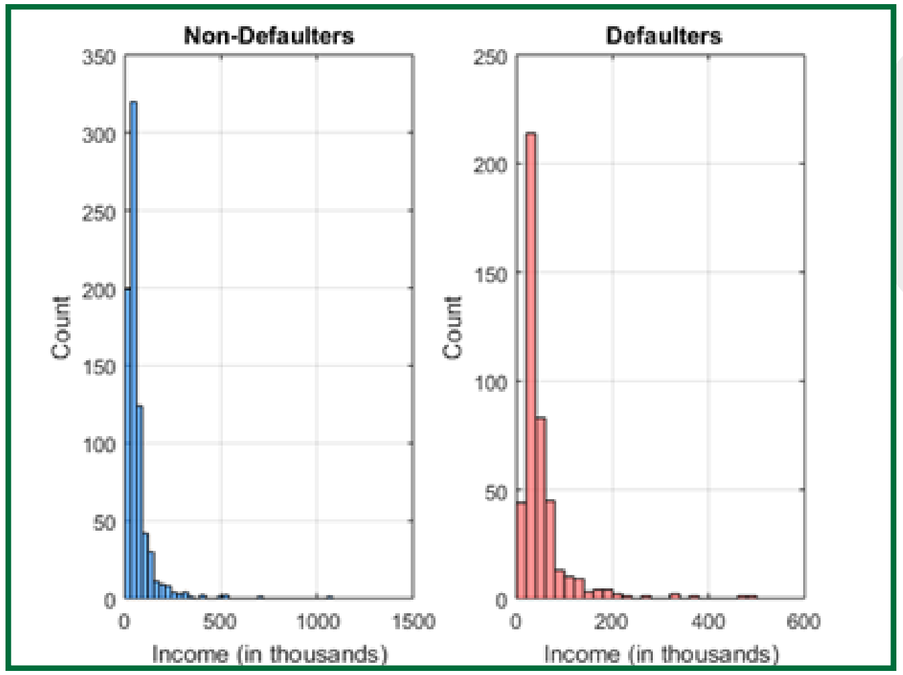


Outliers and Variability

The spread of defaulters is narrower

There is less variation in their years at the current address compared to non-defaulters.

Income vs Default Status



- The two histograms not tells a clear difference between two default groups. Both are right-skewed

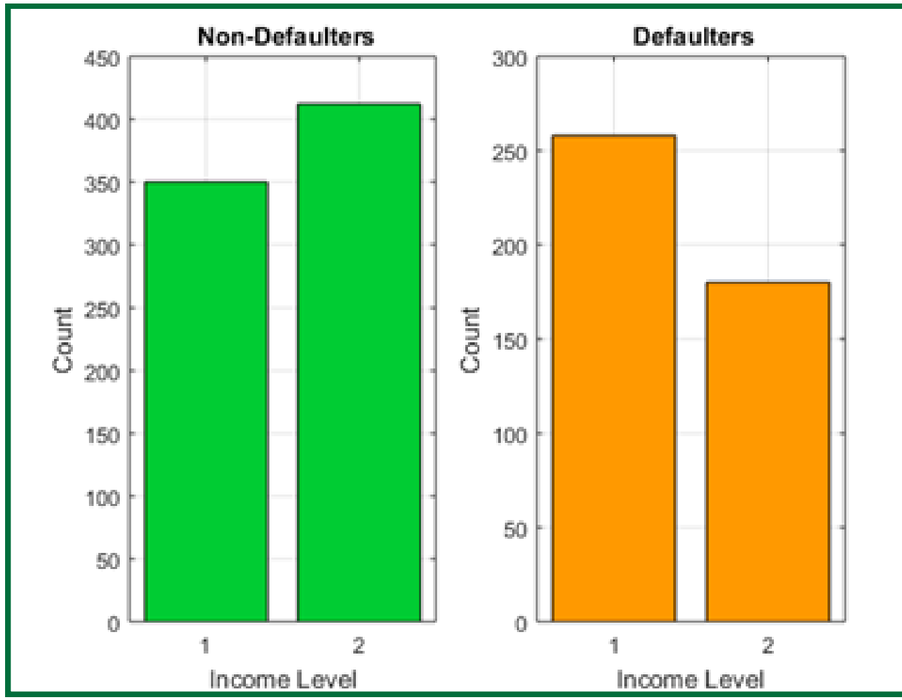
Income Level

Income<=40 000- Income level 1

Low income

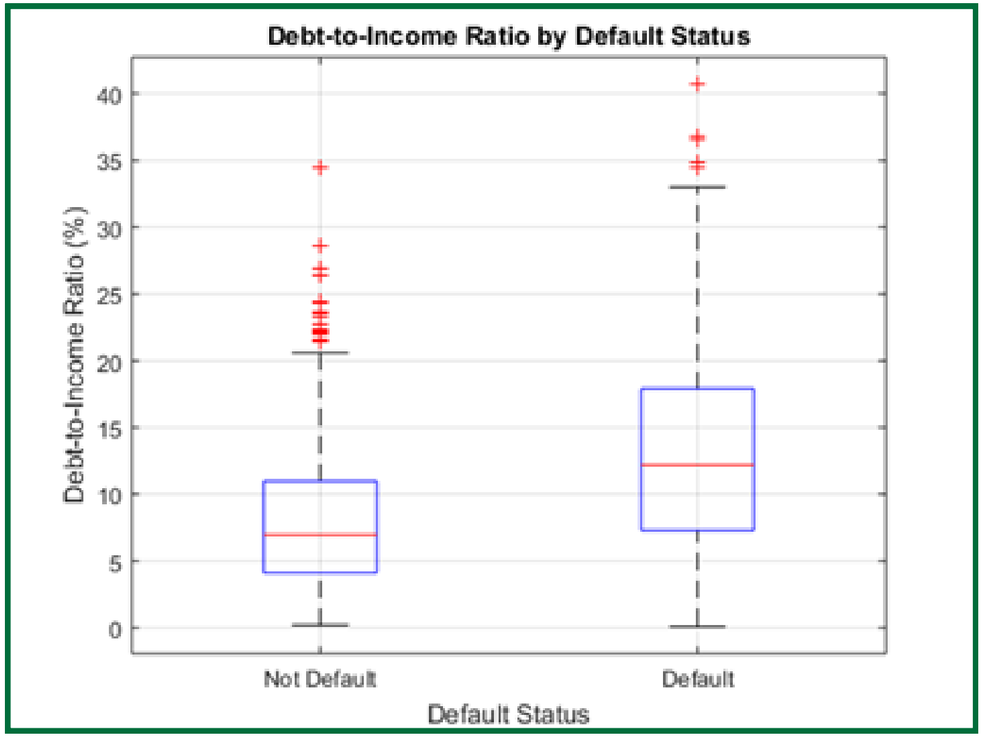
Income>40000-Income level 2

High Income



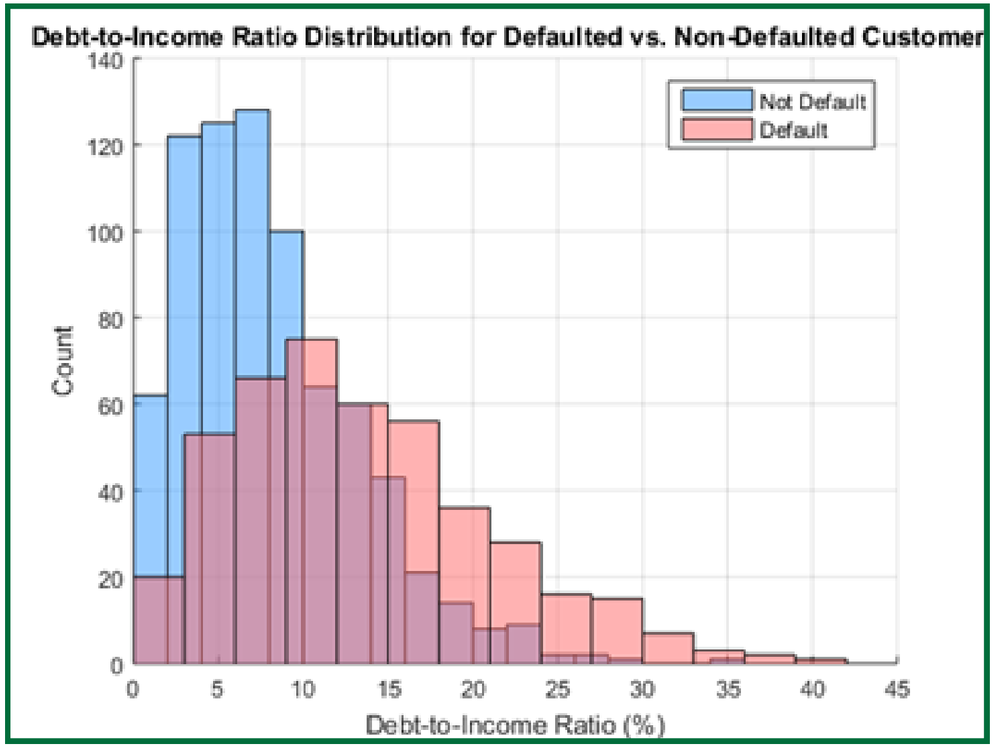
- Most of defaulters are in low income level . This suggests that the **individuals with low income may have high Risk to be default**.

Debt to income ratio vs Default Status



Debt to Income Ratio

The Debt-to-Income (DTI) ratio is a financial metric that compares an individual's total monthly debt payments to their gross monthly income.
It is used by lenders to assess a borrower's ability to manage monthly payments and repay debts.



- Higher DTI is associated with a higher probability of default

  They are dedicating a larger portion of their income to debt payments, leaving less financial flexibility.

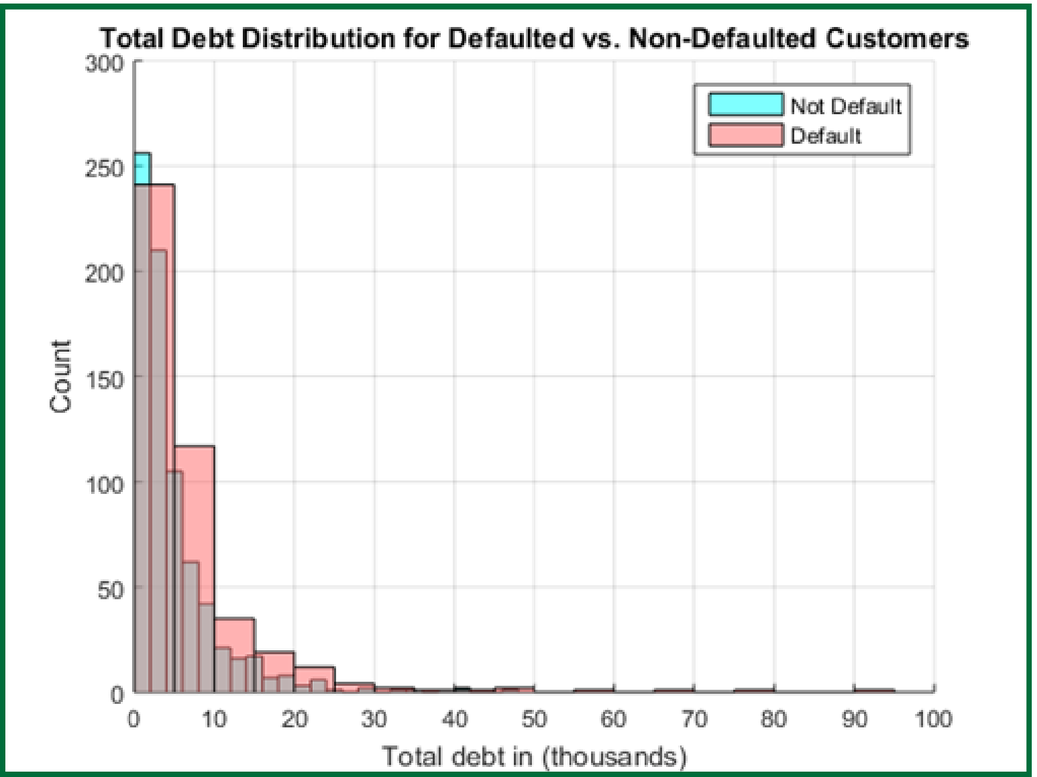
- Lower DTI is more common in the non-default group

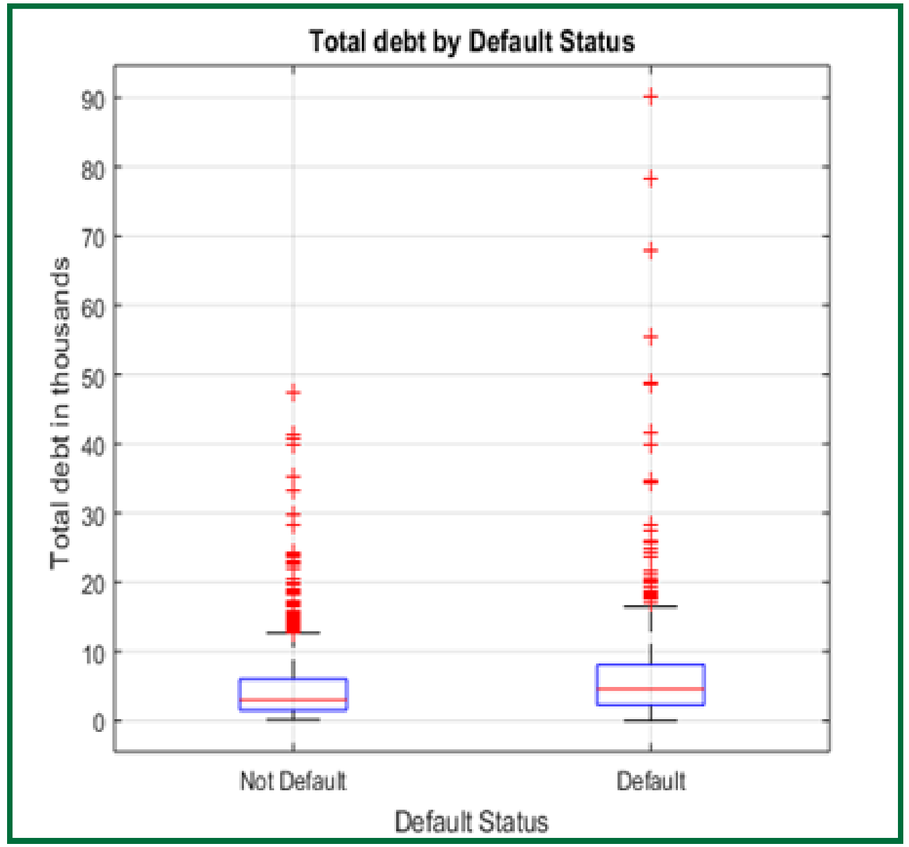
  They have enough disposable income to meet debt obligations comfortably.

Total Debt vs Default Status

Total Debt = Creddebt + Othdebt

The difference is not substantial.
Total debt alone is not a strong differentiator between the two group



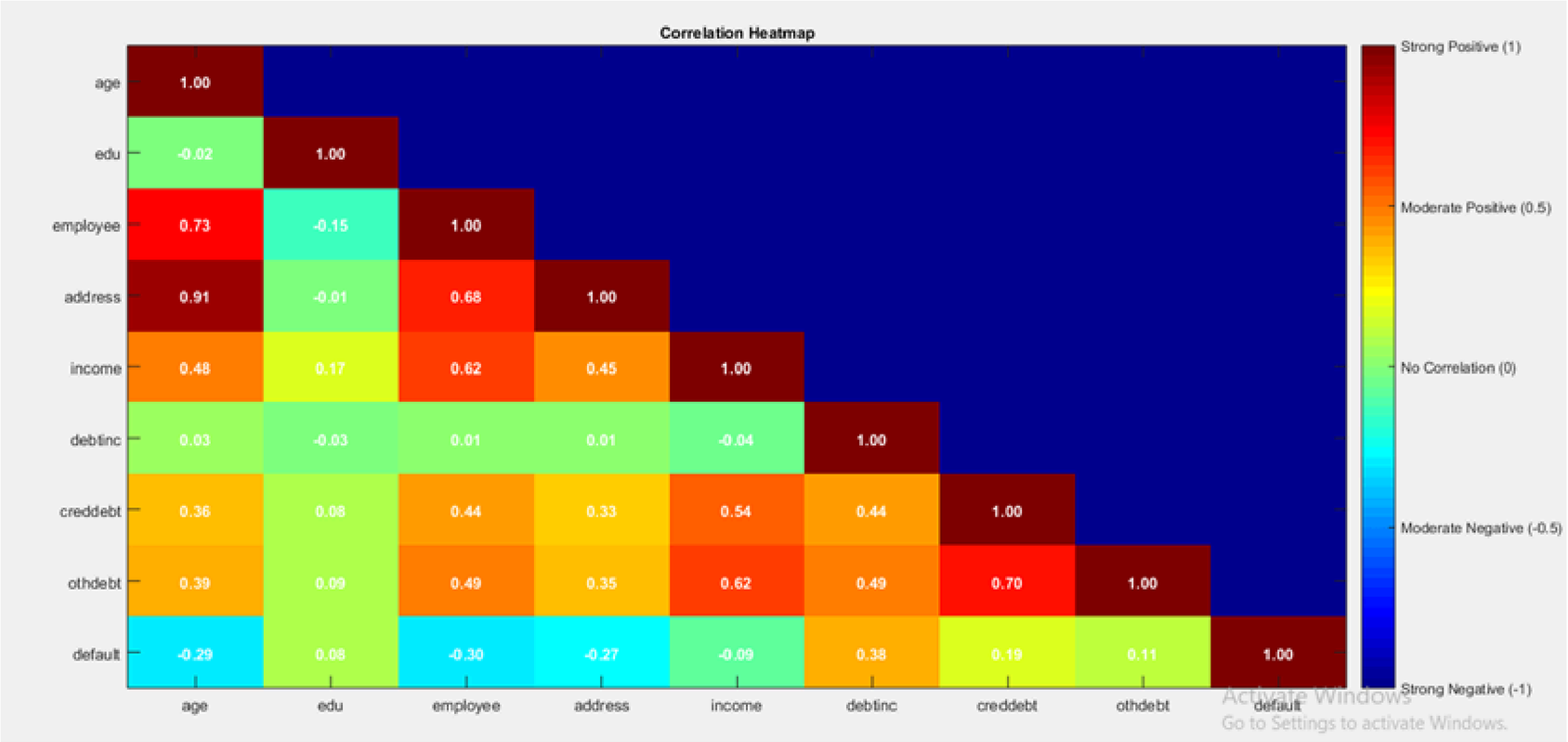


Presence of Many Outliers in Both Groups

some individuals carry extremely high debt regardless of default status.

some high-debt individuals manage their finances well, while others struggle.

Correlation Heatmap



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

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

Summary of results

- Lower Default Risk (Negative Association):
  - Older age
  - Stable job
  - Long address duration
  - Higher income reduce default chances.

- Higher Default Risk (Positive Association):
  - More debt,
  - Higher debt-to-income ratio
  - Education level slightly increase default risk.

Stable income & employment lower risk, while high debt increases it.

Advanced Analysis

Finding the best model

we check the accuracy by training and test errors

- PNN(Probability neural Network)
- KNN
- Classification Trees

PNN (Probability Neural Network)

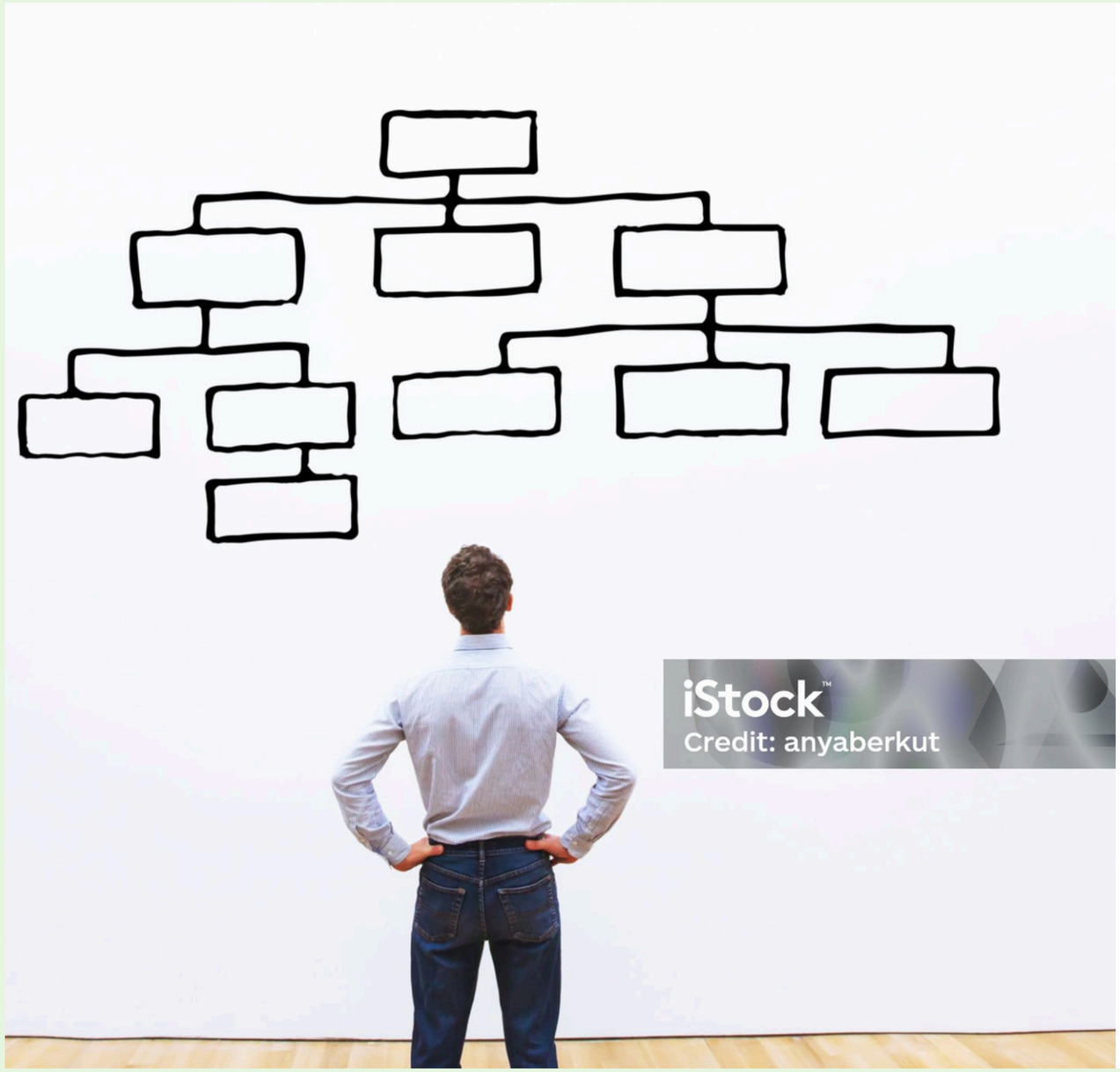
A probabilistic neural network (PNN) is a sort of feedforward neural network based on radial basis function used to handle classification and pattern recognition problems. Error is the same across all spread parameter values that we used.

Error rate => 36.67%
Accuracy =>63.33%

KNN

The KNN algorithm is a distance based powerful classification algorithm.

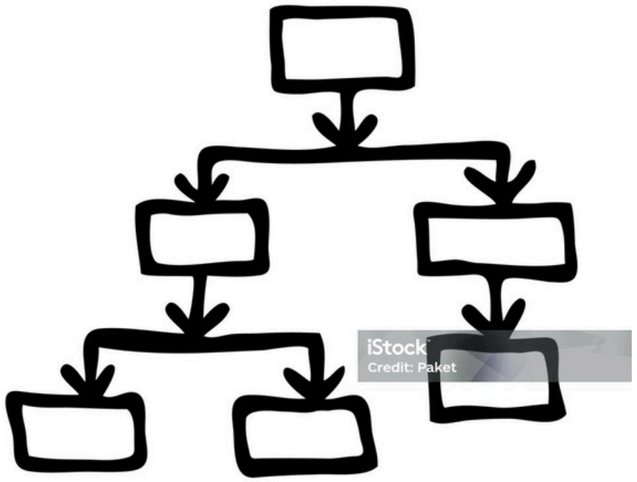
Optimum K => 7
Accuracy => 62.00%



Identifying Important Variables and Prediction

Decision Trees

An Overview



- A supervised learning algorithm used for classification and regression.
- Splits data into branches based on decision rules derived from feature values.

Root Node

The first split based on the most important feature.

Internal Nodes

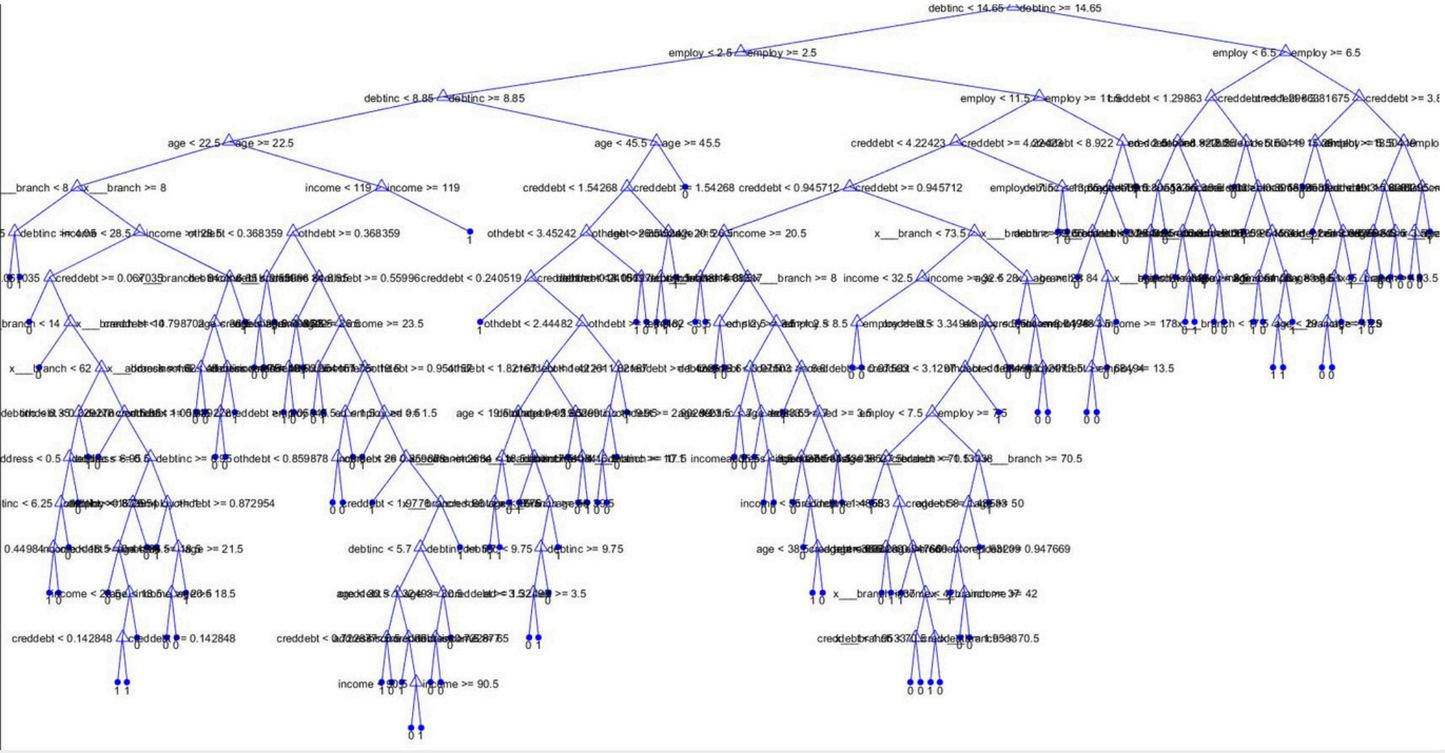
Splits that create further branches.

Leaf Nodes

Final decision points

The Original Tree

- Prune Criterion ‘Impurity’ was used – Target Variable was Imbalanced



Observations

Train Error - 0.0917
CV Error - 0.3292
Test Error - 0.3533

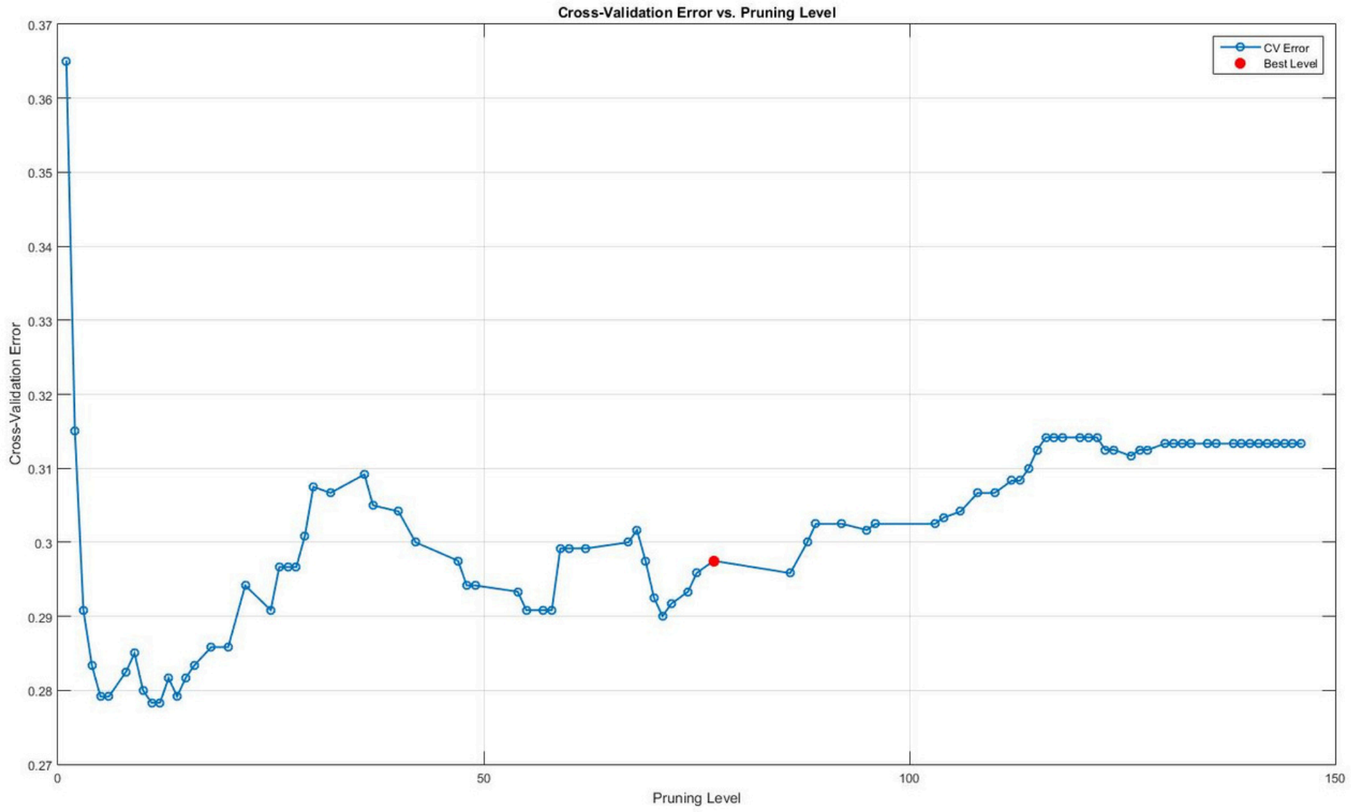
Complex and Deep Tree

Most likely to be Overfitted

Improvement

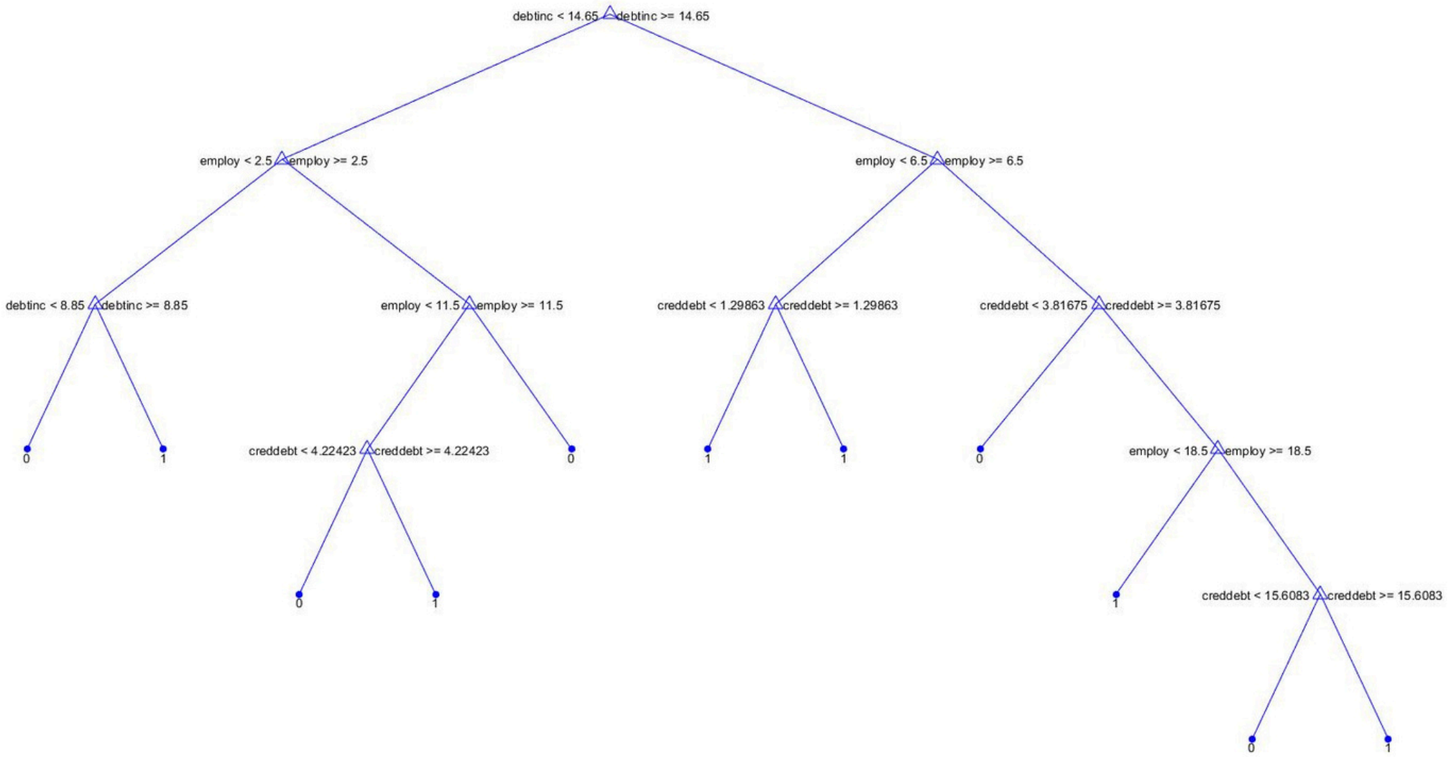
Consider a cross validation approach to get the best pruned tree

Pruning Process



Process	10-Fold Cross Validation
Selection Criteria	1 SE deviation Rule
Optimal prune level	79

Results of the Pruning



Observations

Train Error - 0.2274
CV Error - 0.32
Test Error - 0.2767

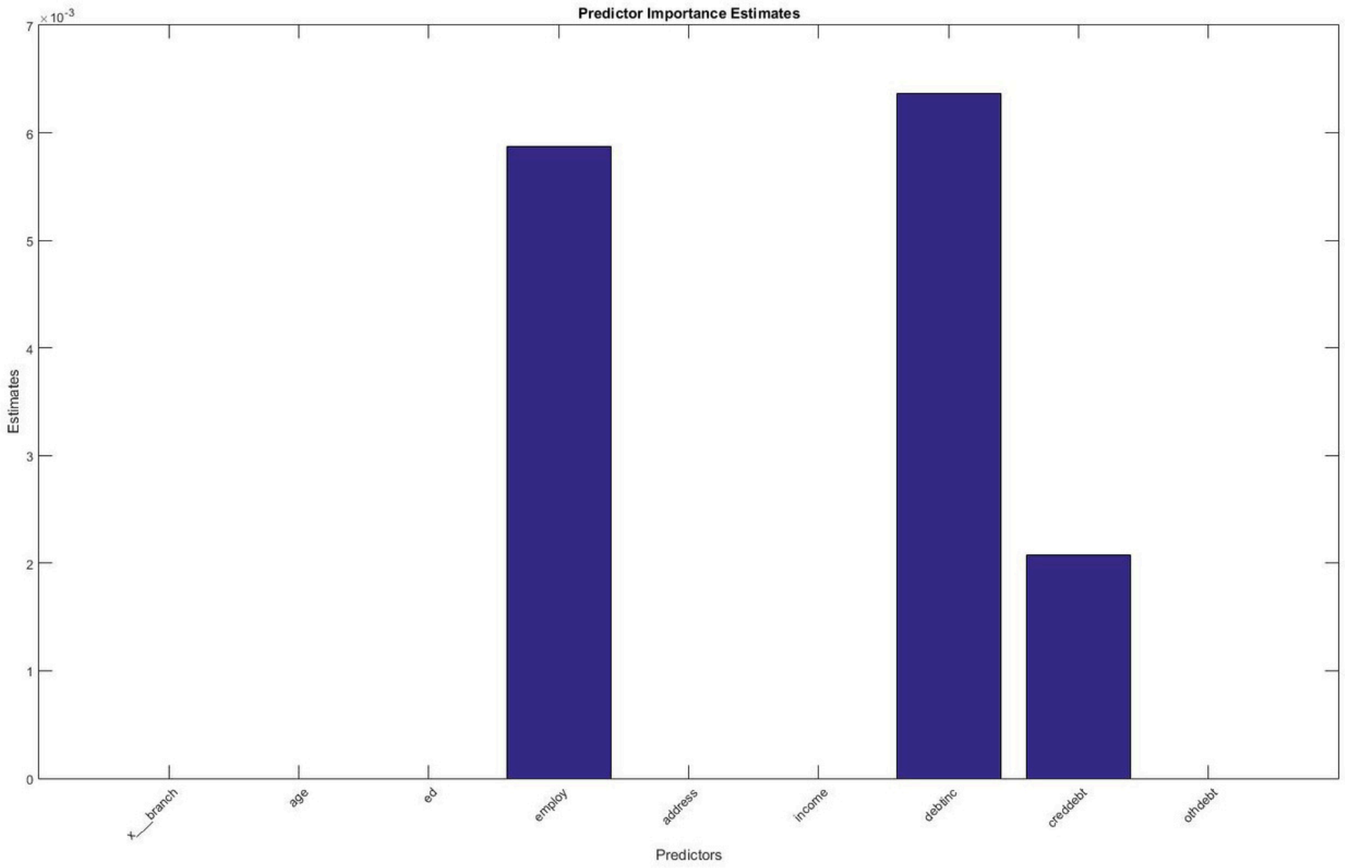
Has 9 levels now instead of 88 levels

Conclusion

Test Error has Reduced

Tree is simple and less likely to overfit

Results of the Pruning



Identifying the most important features

The most important features are

- **employ**
- **debtinc**
- **creddebt**

Conclusion

Important feature reduction from 8 to 3

Identify Homogeneous Groups

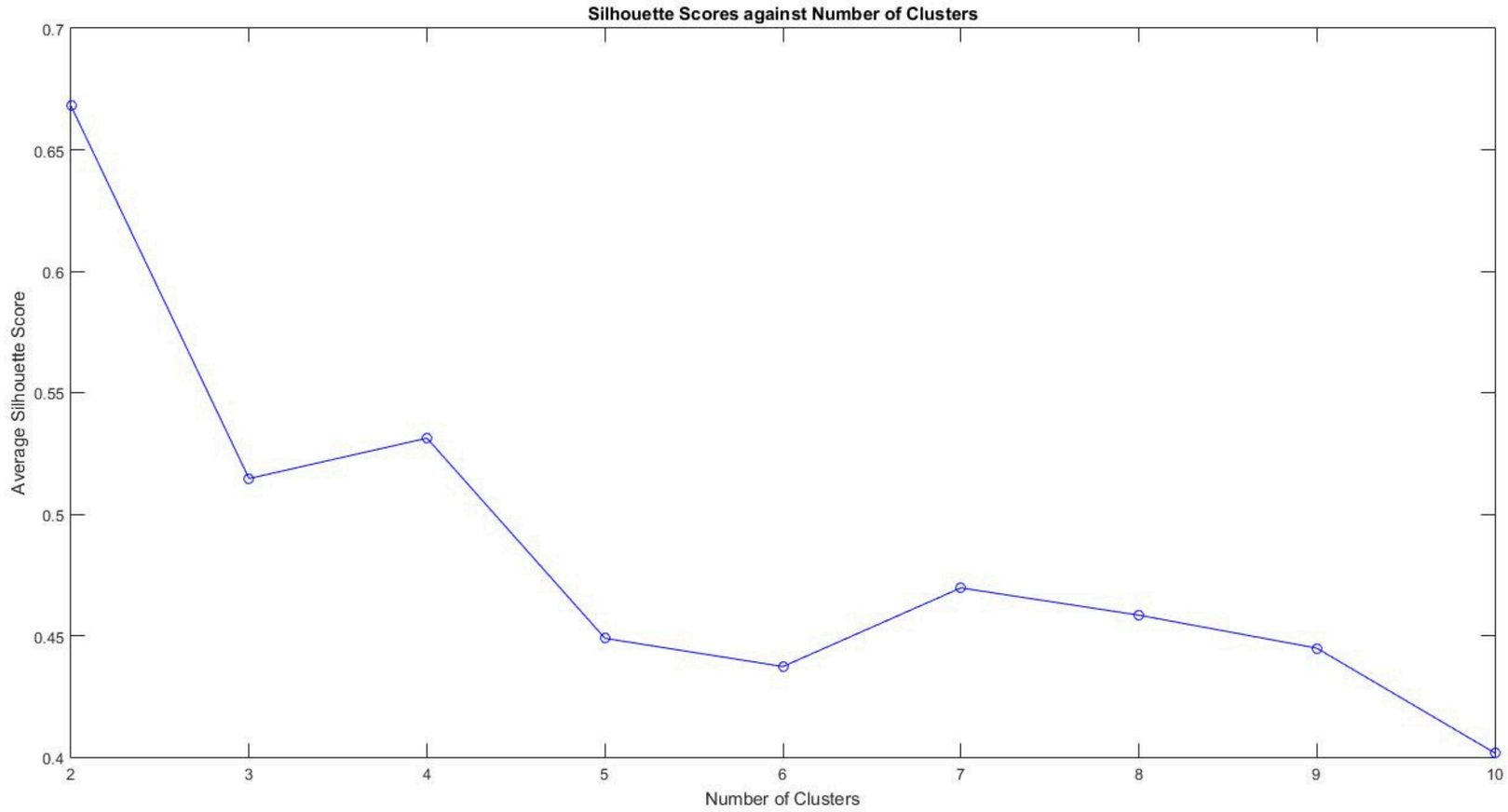
K-Means Clustering...

K-Means Clustering

Find optimal number of clusters using average silhouette score.

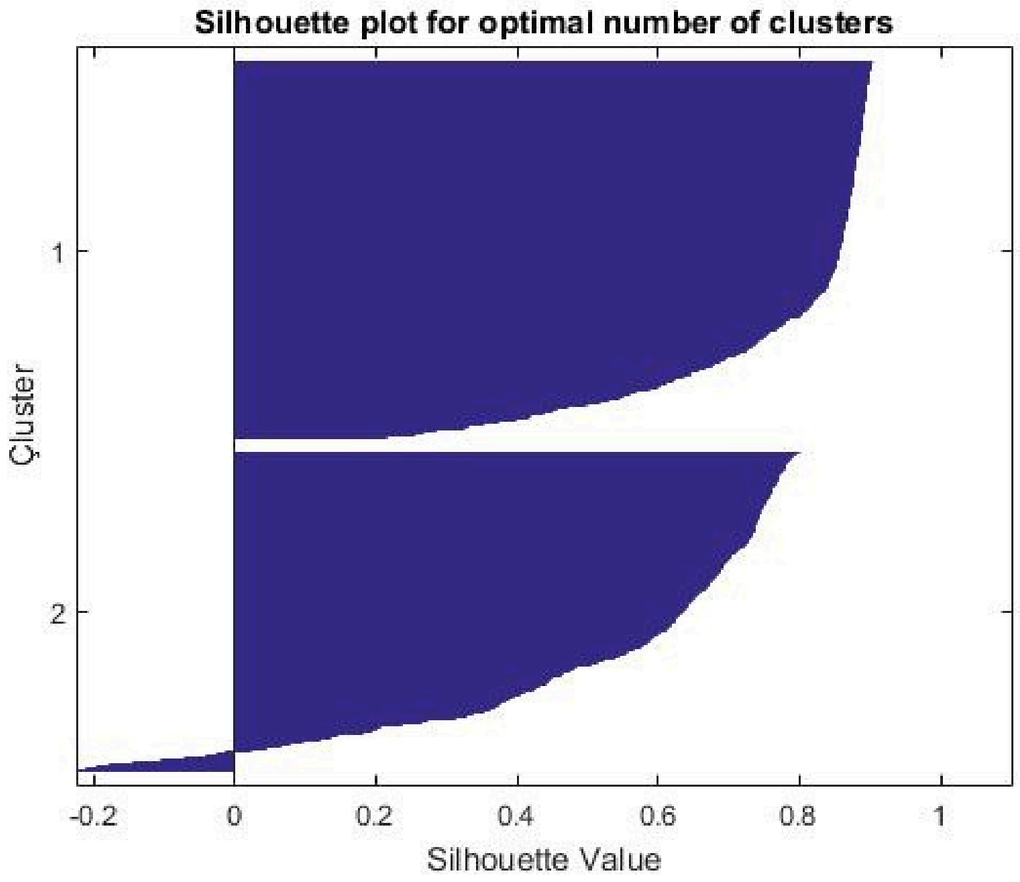
- Optimal k = 2
- With average silhouette score of 0.6680

🔎 **Number of clusters is 2**



K-Means Clustering

Silhouette Plot for the two clusters



- Measure how well each point fits within its assigned cluster compared to other clusters.

Cluster 01 – 40.31%
Cluster 02 – 32.00%

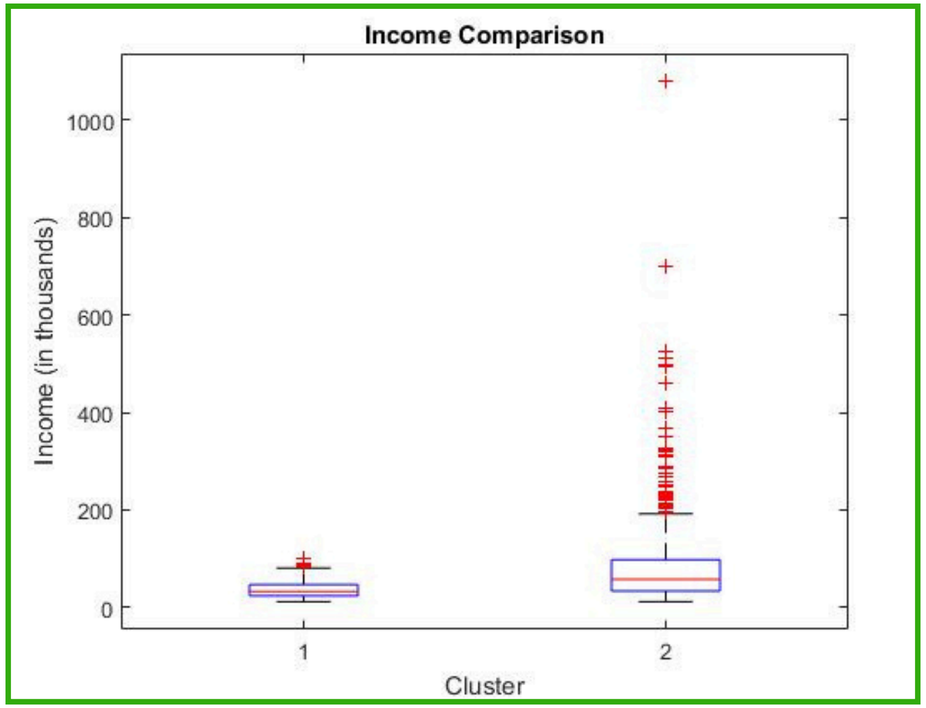
- The group of customers having a higher chance to default in the future was identified as;

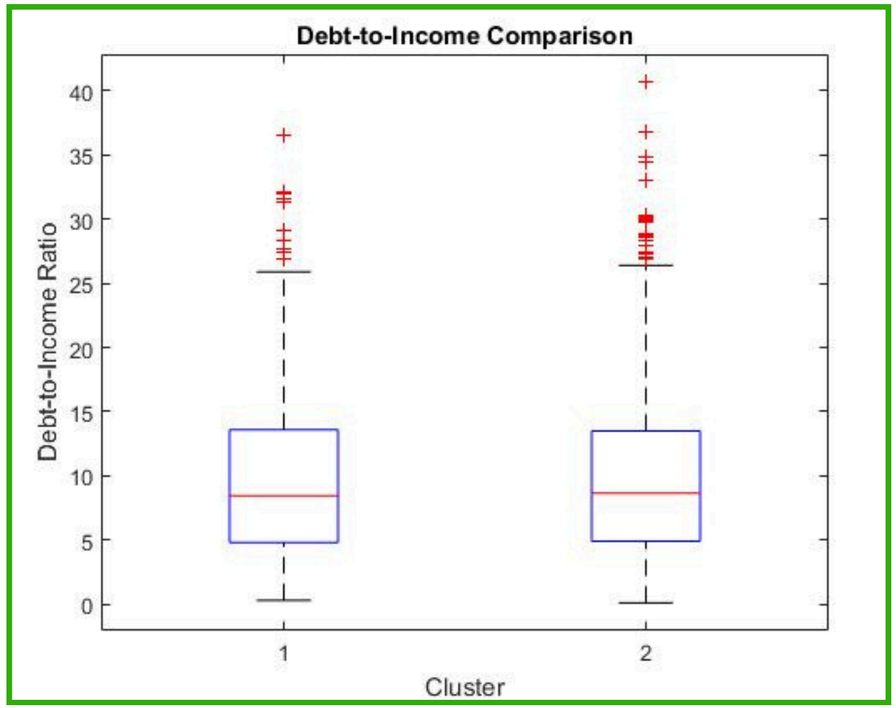
🔎 **Cluster 01**

Common Characteristics in Cluster 01

Income Comparison

- Median income is lower among the customers in cluster 01.
- High end outliers in cluster 02 suggest having a much higher income compared to cluster 01.

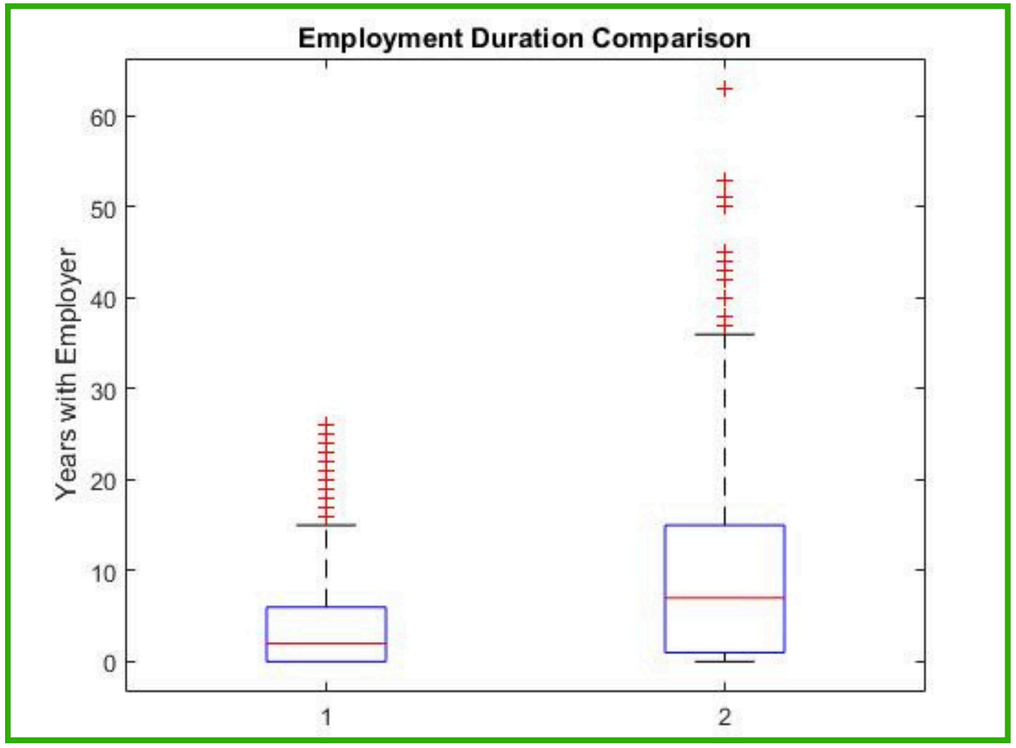




Debt-to-Income Comparison

- Median of debt-to-income ratio is quite similar in both clusters.
- Both are having high end outliers.

Common Characteristics in Cluster 01

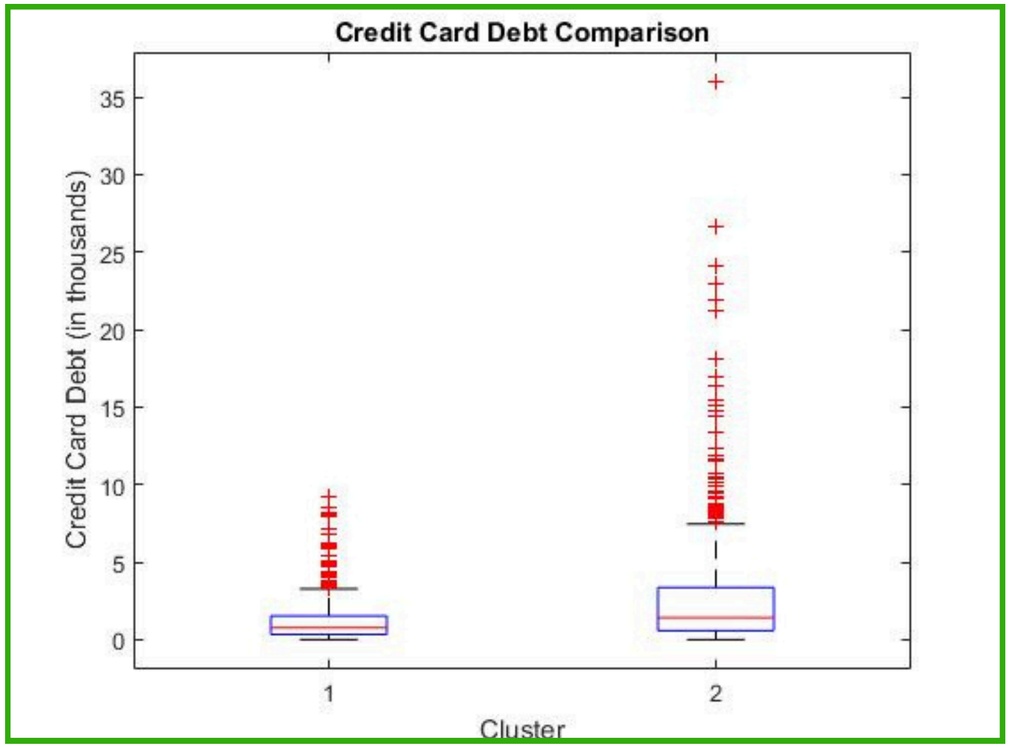


Employment Duration Comparison

- Median number of years with employer is lower among the customers in cluster 01.
- High end outliers in cluster 02 suggest a higher number of years compared to cluster 01.

Credit Card Debt Comparison

- Median of credit card debt is lower in cluster 01.
- Fewer extreme outliers compared to Cluster 02, indicates a relatively lower financial burden.



Common Characteristics in Cluster 01

Variables	Cluster 01	Cluster 02
Age	**28**	**38**
Years at current addresss	**3**	**7**
Other debts	**1.7684**	**3.2898**

- Lower median age of the customers in cluster 01 indicates it consists of generally younger people.
- Median number of years at current address is lesser than cluster 02 showing that they have moved residences more frequently.
- Other debts is also lower for the customers in cluster 01, may be they are more cautious about money.

Common Characteristics

- Lower age
- Lower number of years with current employer
- Lower number of years at current address
- Lower income
- Lower credit card debt
- Lower amount of other debts.

Comparison of Common Characteristics

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

- Age, Years with current employer, Years at current address and Income show quite similar median values for both groups.
- Credit card debt and debt-to-income ratio is higher of the previously default customers.

🔎 **This suggests that defaults are more likely caused by how people handle their money and debt rather than their income or job stability.**

THANK YOU!

Group 2


Kavindi Chamathka **s16367**


Lakmini Thamodya **s16340**


Maheesha Sewmini **s16349**


Poornima Tharangani **s16368**


Kavina Dias **s16200**