CREDIT SCORING

GOWRI ABINESH

Background

- In today's dynamic financial environment, accurately assessing creditworthiness is more critical than ever. Traditional credit evaluation methods often fall short in predicting default risk effectively.

- This project aims to strengthen credit scoring practices by analyzing demographic, financial and behavioral data to uncover key risk indicators. By integrating machine learning with statistical modeling, it explores innovative ways to build data-driven scorecards that better identify default risks.

- Additionally, customer feedback from Net Promoter Score (NPS) surveys is analyzed to understand customer sentiments and identify areas for service improvement around the loan and credit process.

Project Objectives

ANALYZE DEFAULTERS AND ASSOCIATED FACTORS

DEVELOP A CREDIT SCORECARD USING BINARY LOGISTIC REGRESSION

COMPARE MODEL PERFORMANCE WITH ML METHODS

ANALYZE CUSTOMER FEEDBACK USING TEXT MINING

Data Source

This analysis is based on four datasets:

- Customer Credit Details
- Customer Info
- Customer Profile
- Customer feedback in text format

Data Snapshots

Customer Credit Details

```
##   custid debtinc creddebt othdebt preloan
## 1      1   12.34    13.26    5.88       2
## 2      2   18.65     2.12    5.13       1
## 3      3    7.22     3.31    3.65       1
## 4      4    6.15     2.95    2.34       1
## 5      5   20.64     2.67    4.07       2
## 6      6   12.44     3.06    2.57       2
## Dimensions: 6993 x 5
```

Customer Info

```
##   custid veh house selfemp account deposit emp address branch ref
## 1      1   2     1       2       1       2  18      12      1   2
## 2      2   2     1       1       2       1  11       7      1   2
## 3      3   1     2       1       1       1  16      15      1   1
## 4      4   1     2       1       1       2  15      14      2   1
## 5      5   2     1       1       1       1   2       1      2   2
## 6      6   2     2       1       1       2   5       5      1   2
## Dimensions: 6990 x 10
```

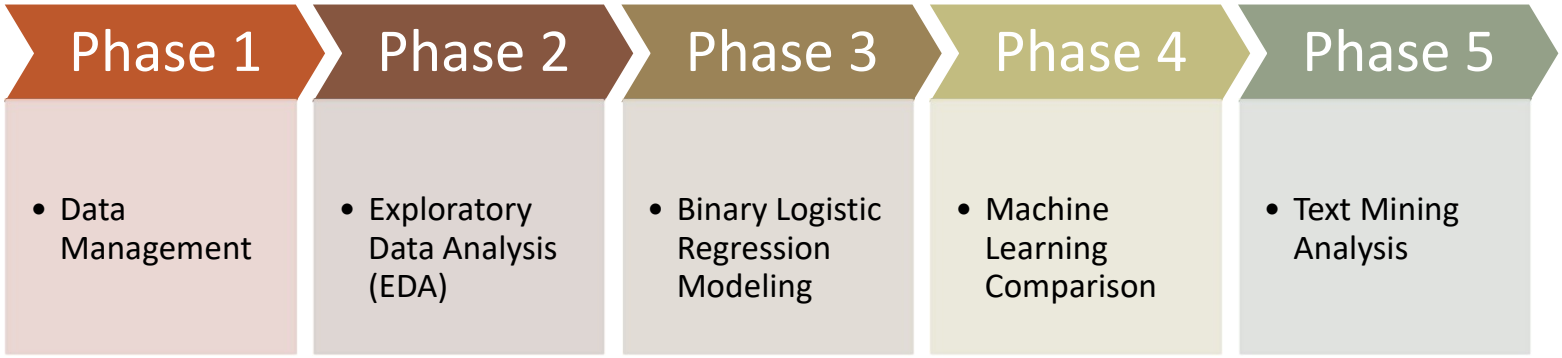
Customer Profile

```
##   custid age gender ms child zone bad
## 1      1   2      1  2     2    7   1
## 2      2   2      2  1     1    5   0
## 3      3   1      1  2     1   15   0
## 4      4   2      1  1     1    3   0
## 5      5   2      2  1     1   20   0
## 6      6   1      2  1     1    5   0
## Dimensions: 6990 x 7
```

Customer feedback in text format

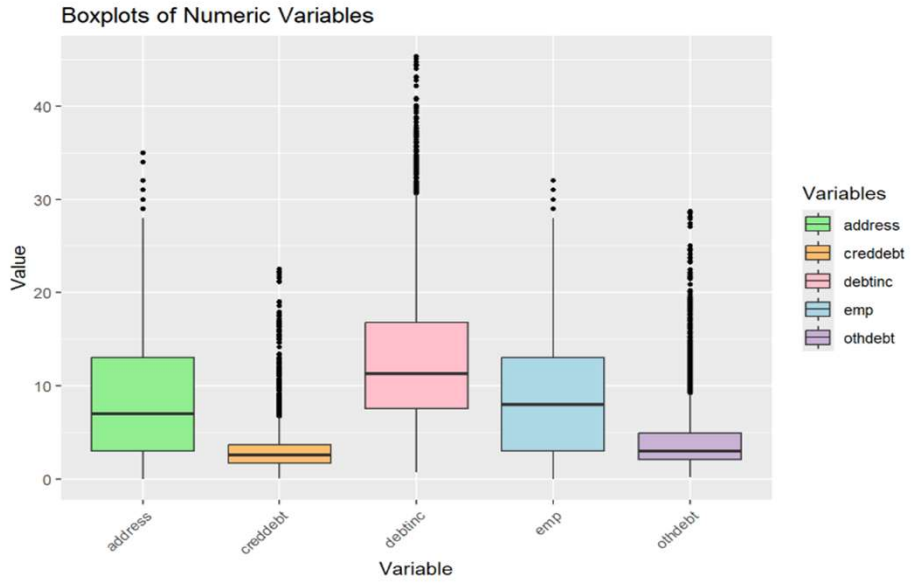
```
My experience of availing the personal loan was smooth.
I had a satisfactory experience with the Bank. The loan officer was helpful
rates were a bit higher than what I expected.
I had an excellent experience with the Bank while applying for a loan. The
few days.
I had to move on to another bank as it was taking a long time
I recently applied for a loan at the Bank, and the service was impeccable.
```

Project Phases Overview



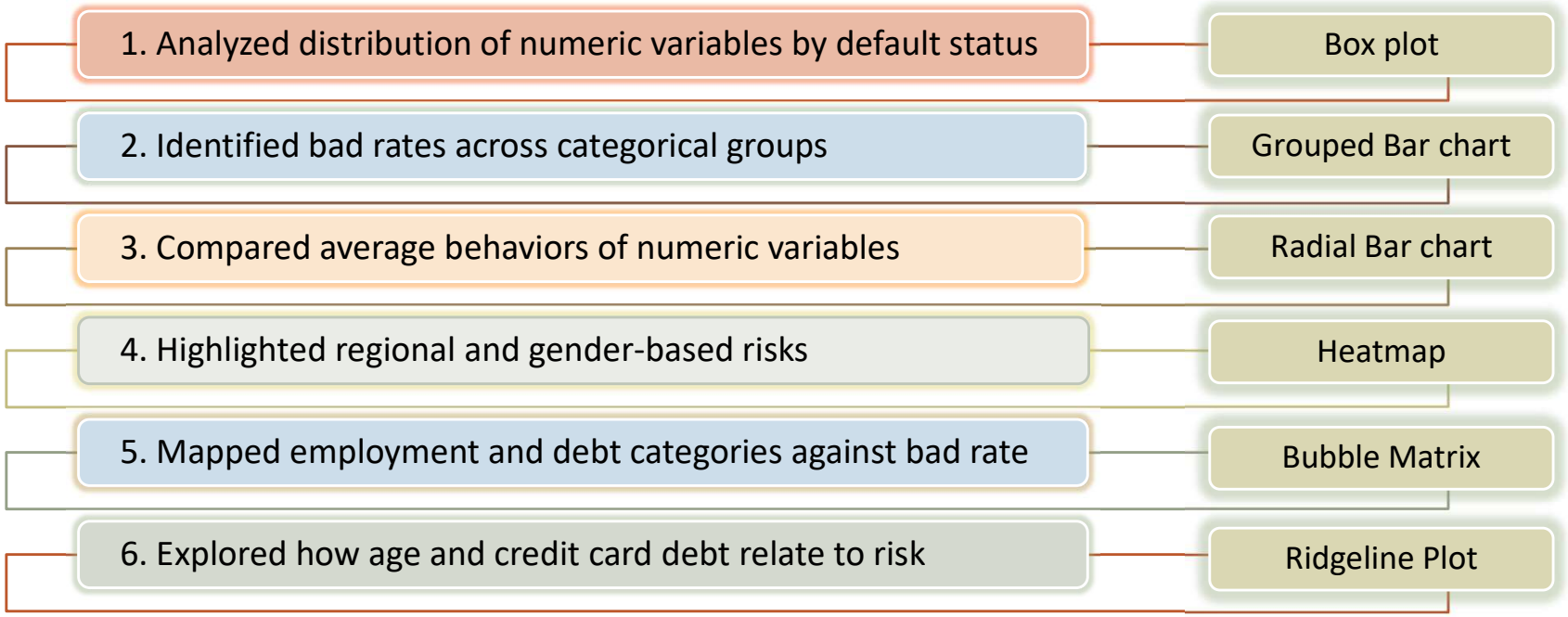
Phase 1 - Data Management

1. Imported and merged structured datasets into master data
2. Cleaned missing values and duplicates
3. Recoded numerical variables into categories
4. Calculated overall Bad Rate



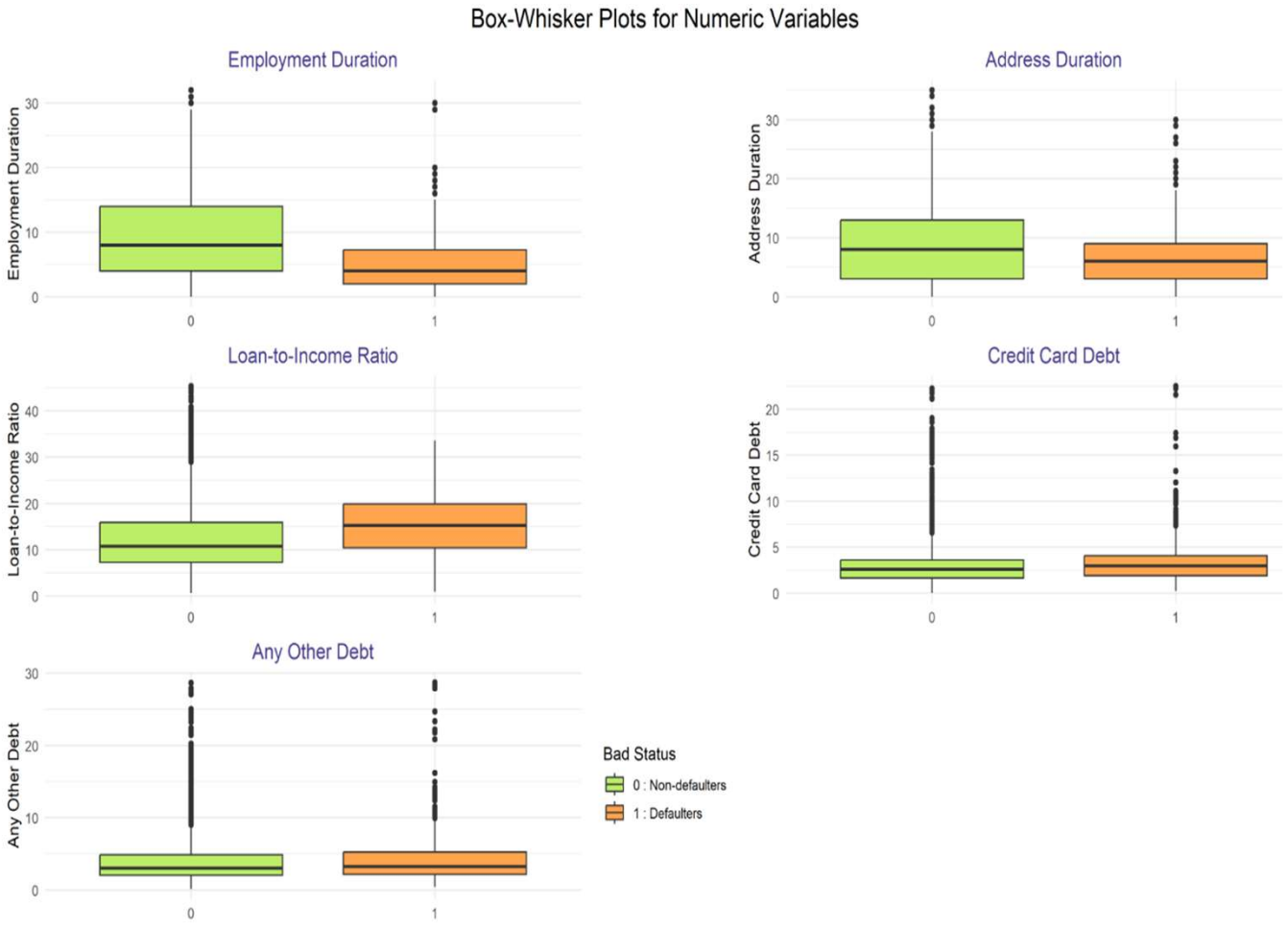
The cleaned dataset revealed notable outliers in debt-related variables and an overall defaulter rate of 13.2%, providing the preliminary understanding of default risk.

Phase 2 - Exploratory Data Analysis



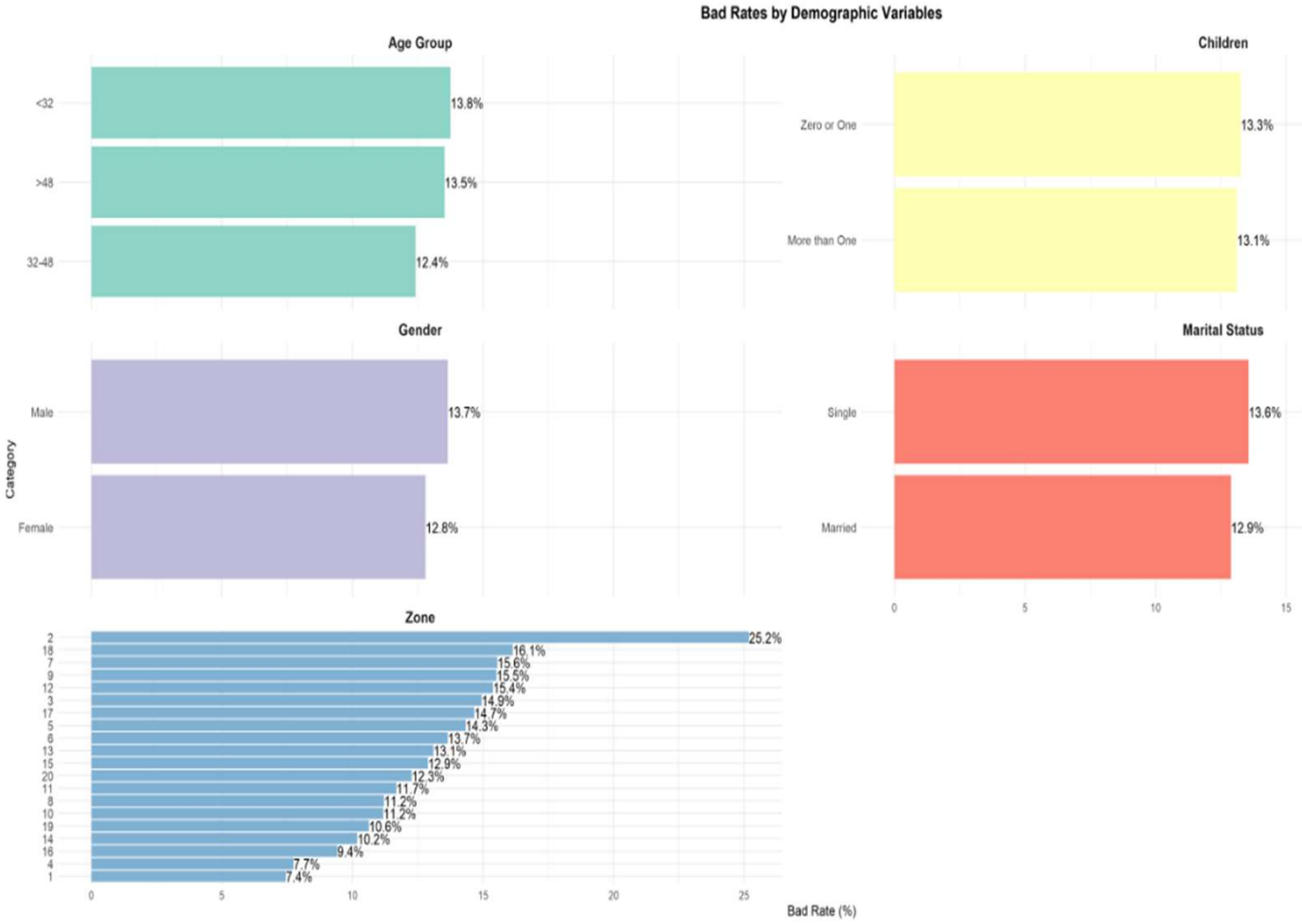
1. Distribution of Numeric Risk Indicators

Defaulters tend to have shorter employment and address histories and carried slightly higher levels of debt and loan-to-income ratios, compared to non-defaulters.



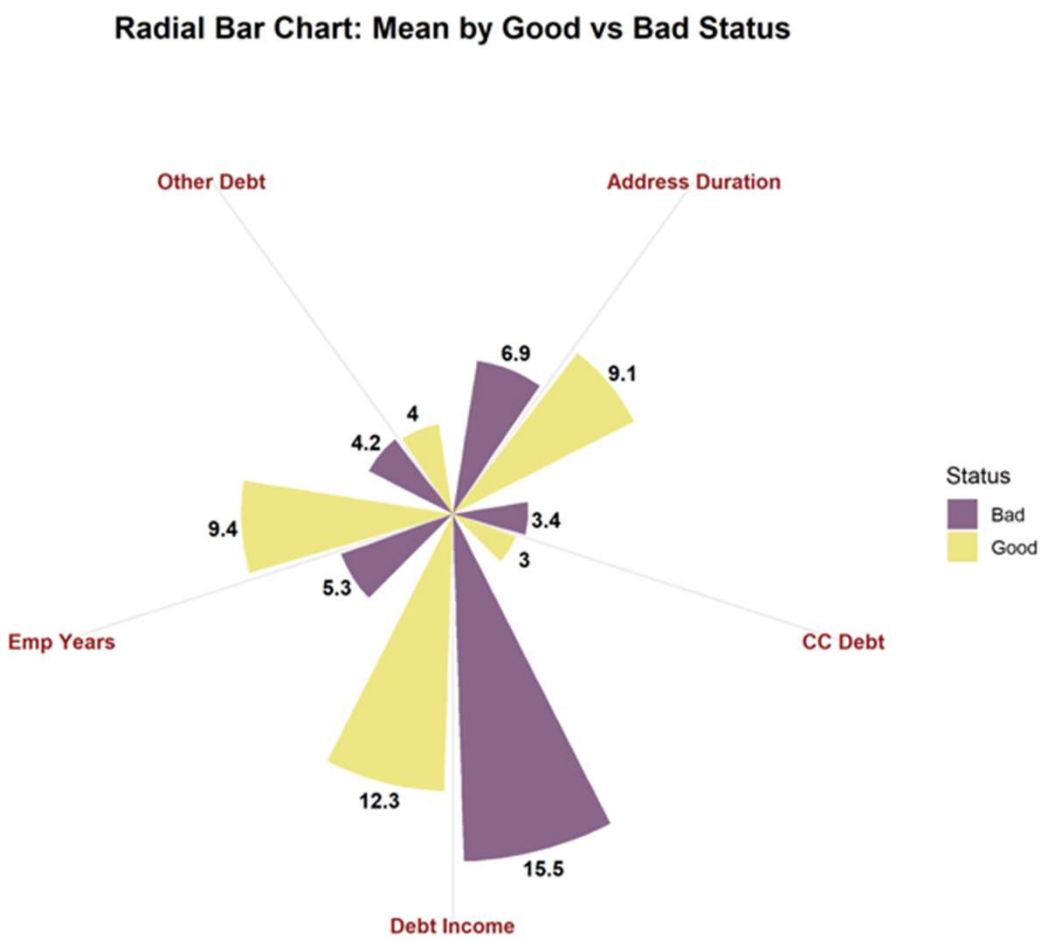
2. Bad Rates Across Demographic Segments

Slightly higher default rates were observed among younger individuals, males, singles and certain zones, suggesting subtle demographic influences on credit risk.



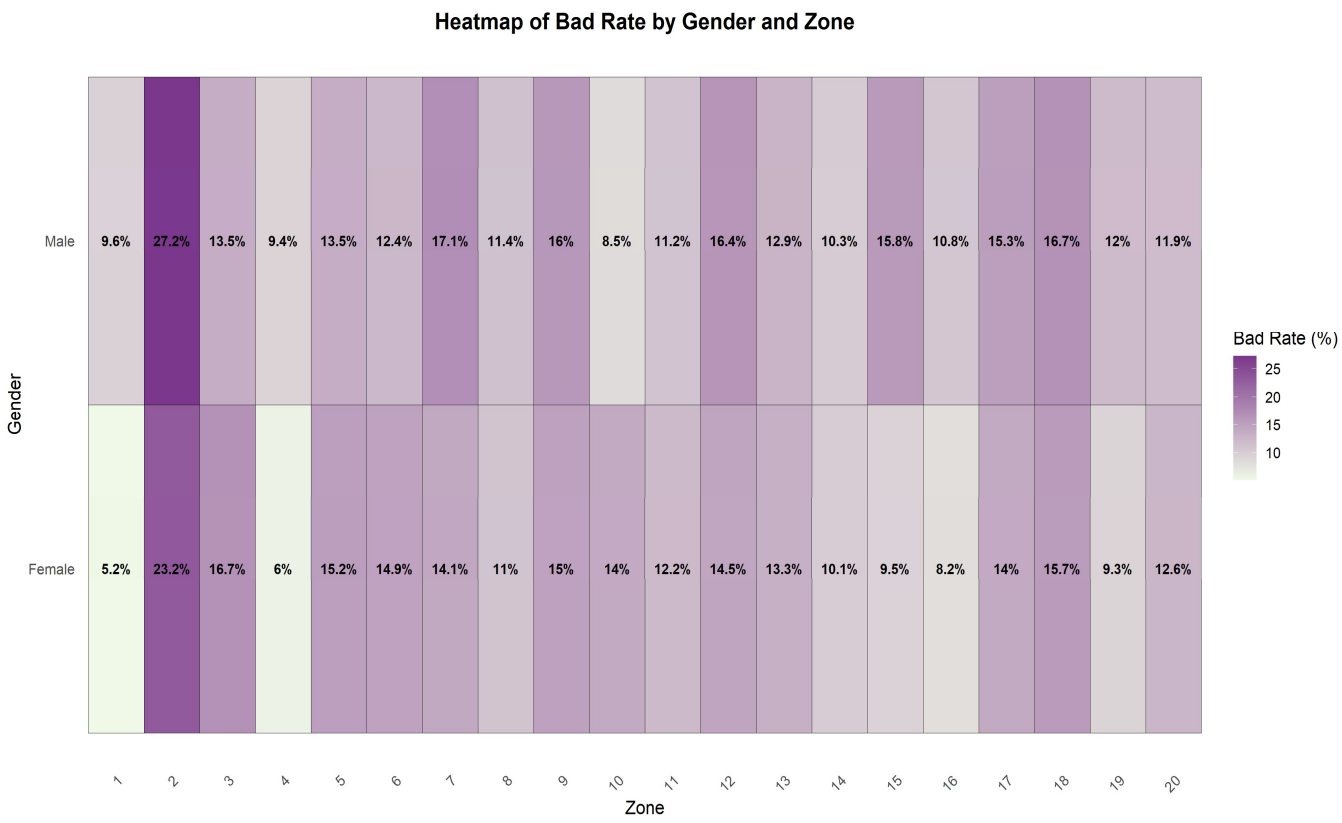
3. Mean Comparison: Good vs Bad Customers

Radial bar chart shows that the defaulters have noticeably higher average values in debt related variables and shorter durations in employment and address history.



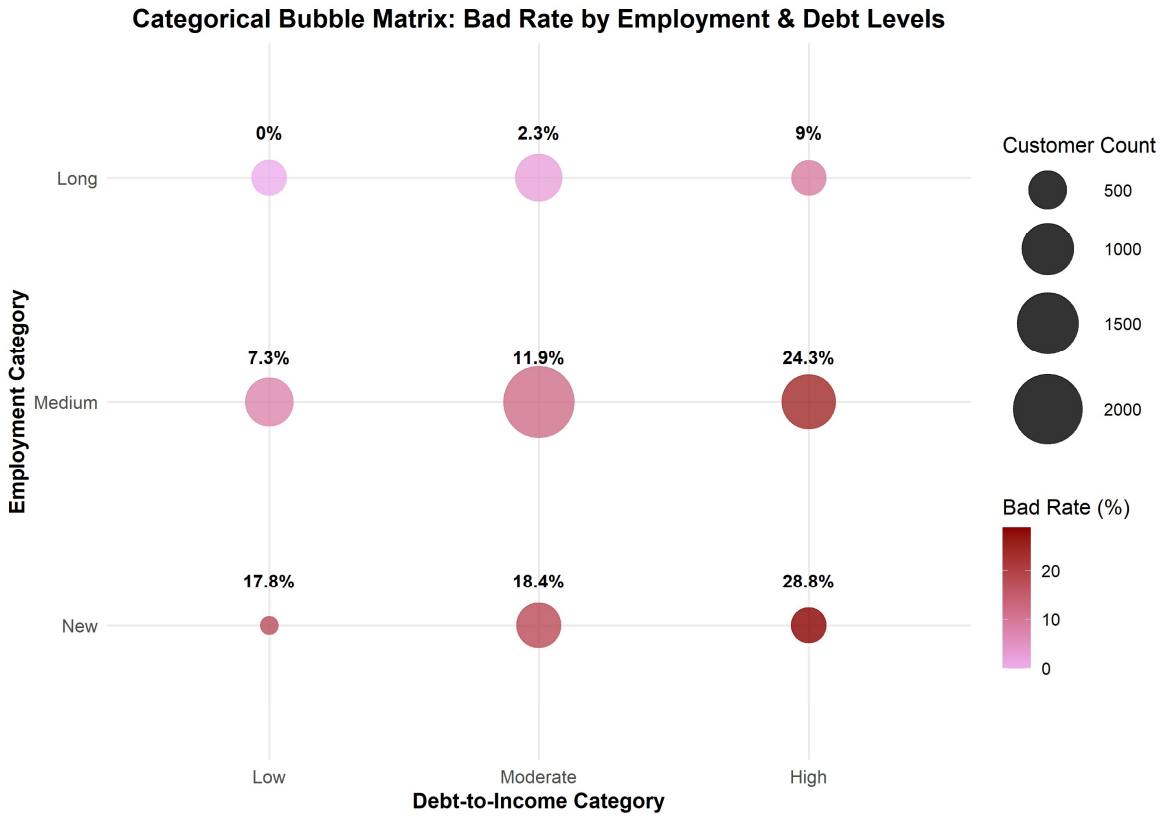
4. Risk Variation by Gender and Zone

Gender-Zone heatmap revealed concentrated default risk among males in certain zones, particularly in Zone 2 and Zone 7.



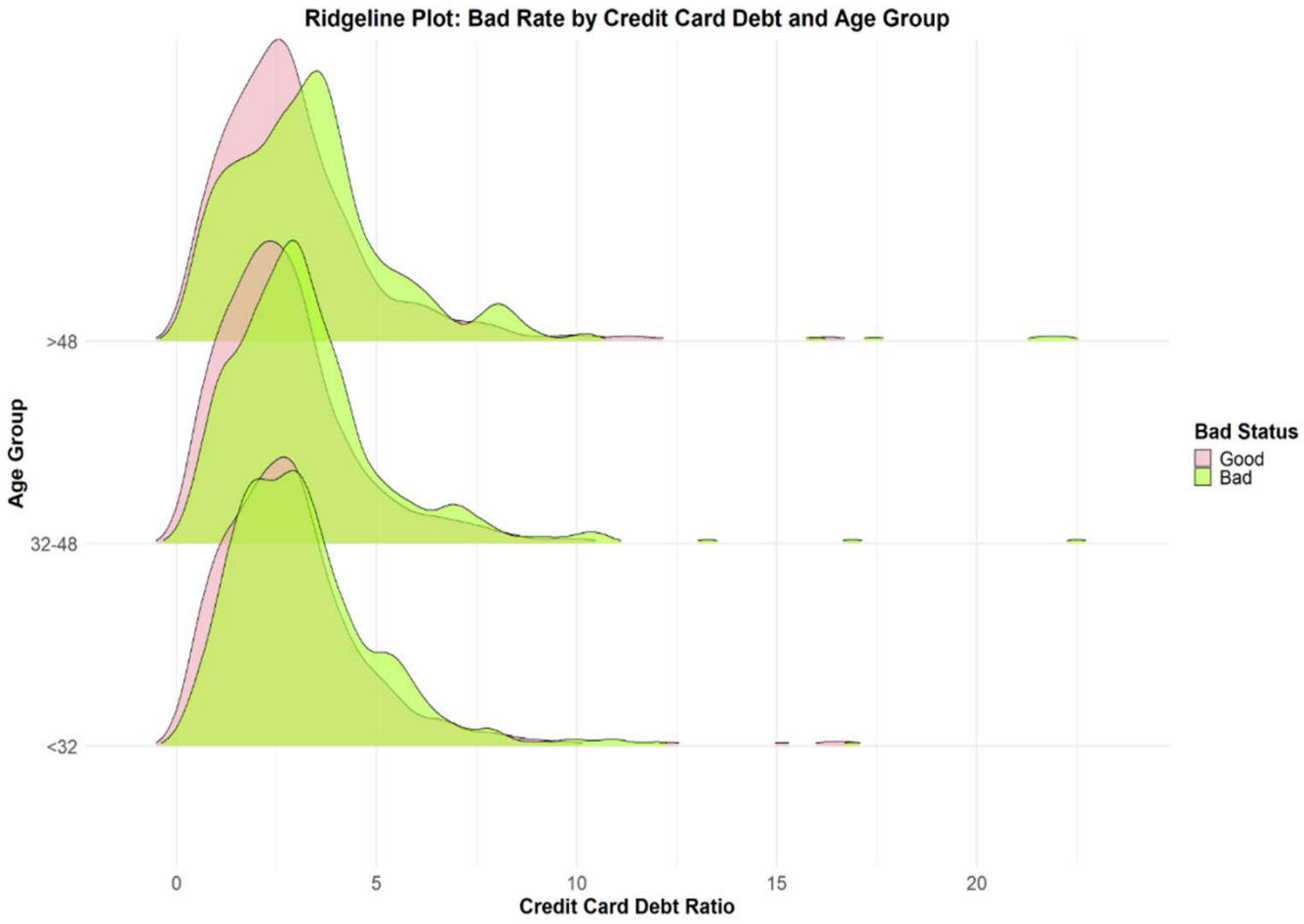
5. Risk Profile by Employment Category & Debt Level

From the Categorical Bubble Matrix, default rates are highest among newly employed individuals with high debt-to-income ratios.



6. Age wise Credit Card Debt Risk Patterns

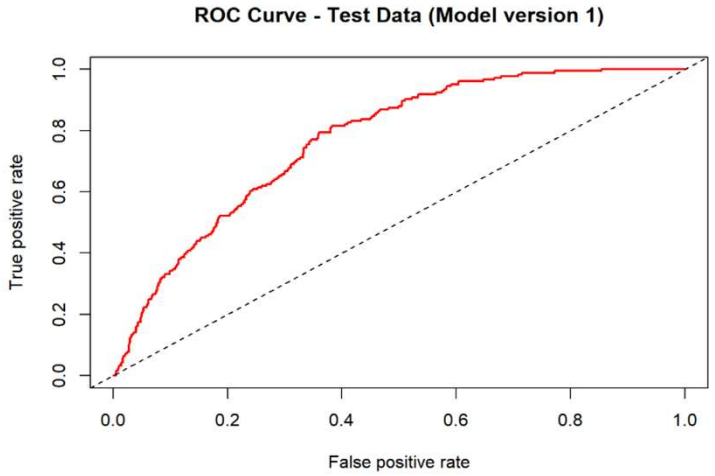
The Ridgeline Plot reveals that younger age group has a wider spread of credit card debt among defaulters.



Phase 3 - Binary Logistic Regression Model

Model 1 Performance - Numeric Variables

- Used original numeric variables to build the binary logistic regression model.
- The ROC curve of model 1 test data indicated reliable separation between defaulters and non-defaulters.
- Model Version 1 achieved a test AUC of 0.77, showing strong discriminatory power.
- At the optimal threshold of 0.147, the confusion matrix showed strong balance between true positive and true negative rates.

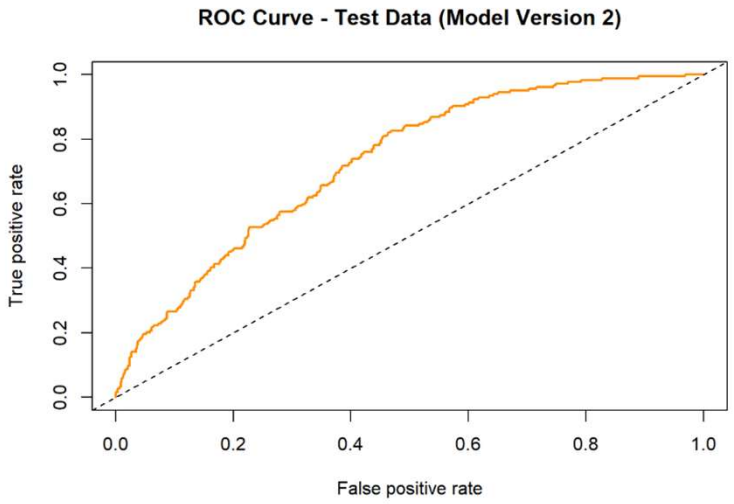


Confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 837  59
##          1 375 125
##
##                Accuracy : 0.6891
##                  95% CI : (0.6641, 0.7133)
##     No Information Rate : 0.8682
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.214
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.67935
##             Specificity : 0.69059
##          Pos Pred Value : 0.25000
##          Neg Pred Value : 0.93415
##              Prevalence : 0.13181
##          Detection Rate : 0.08954
##    Detection Prevalence : 0.35817
##       Balanced Accuracy : 0.68497
##
##        'Positive' Class : 1
##
```

Model 2 Performance - Categorical Variables

- Used recoded categorical versions of the same variables for comparison.
- The ROC curve of model 2 test data showed moderate class separation.
- Model Version 2 recorded a test AUC of 0.7281, indicating slightly lower performance compared to model 1.
- Using the optimal threshold of 0.139, the test confusion matrix showed reasonable classification balance.



Confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 817  71
##          1 395 113
##
##                Accuracy : 0.6662
##                  95% CI : (0.6408, 0.6909)
##     No Information Rate : 0.8682
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.165
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.61413
##             Specificity : 0.67409
##          Pos Pred Value : 0.22244
##          Neg Pred Value : 0.92005
##              Prevalence : 0.13181
##          Detection Rate : 0.08095
##    Detection Prevalence : 0.36390
##       Balanced Accuracy : 0.64411
##
##        'Positive' Class : 1
##
```

Model Summary

```
##
## Call:
## glm(formula = bad ~ emp + address + branch + debtinc + creddebt +
##     zone, family = binomial, data = train_data)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.478822   0.280404  -8.840  < 2e-16 ***
## emp                  -0.144226   0.009522 -15.146  < 2e-16 ***
## address              -0.027002   0.007205  -3.748 0.000179 ***
## branchNearest Branch -0.268789   0.084155  -3.194 0.001403 **
## debtinc               0.036856   0.006273   5.876 4.21e-09 ***
## creddebt              0.185775   0.021724   8.551  < 2e-16 ***
## zone2                 1.626332   0.294599   5.520 3.38e-08 ***
## zone3                 1.051297   0.307287   3.421 0.000623 ***
## zone4                 0.282251   0.346026   0.816 0.414676
## zone5                 0.834347   0.318740   2.618 0.008854 **
## zone6                 1.008445   0.313547   3.216 0.001299 **
## zone7                 1.180097   0.309520   3.813 0.000137 ***
## zone8                 0.653392   0.322679   2.025 0.042878 *
## zone9                 1.215759   0.313945   3.873 0.000108 ***
## zone10                0.492811   0.335573   1.469 0.141951
## zone11                0.752211   0.319612   2.354 0.018597 *
## zone12                0.999847   0.313988   3.184 0.001451 **
## zone13                0.771756   0.316150   2.441 0.014642 *
## zone14                0.465999   0.328265   1.420 0.155730
## zone15                0.846351   0.327176   2.587 0.009686 **
## zone16                0.584903   0.329317   1.776 0.075715 .
## zone17                0.894285   0.314436   2.844 0.004454 **
## zone18                1.160176   0.306417   3.786 0.000153 ***
## zone19                0.635501   0.328839   1.933 0.053290 .
## zone20                0.768799   0.324763   2.367 0.017920 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4370.1  on 5589  degrees of freedom
## Residual deviance: 3821.6  on 5565  degrees of freedom
## AIC: 3871.6
##
## Number of Fisher Scoring iterations: 6
```

Metric	Model 1	Model 2
Test AUC	0.7700	0.7281
Accuracy	0.6891	0.6662
Sensitivity	0.6794	0.6141
Specificity	0.6906	0.6741

Comparison with EDA Findings

- High bad rate zones and high debt ratios observed in EDA were confirmed as significant predictors in modeling.
- Variables like age, gender and house ownership had minor differences in EDA but were not statistically significant in the model.

Insights from Logistic Regression Modeling

- Two logistic regression models were built and evaluated using test data, with Model 1 (numeric predictors) outperforming Model 2 (categorical predictors) in terms of AUC and classification accuracy.
- ROC curves and confusion matrices confirmed that Model version 1 provides better separation and balance at the optimal threshold, indicating its potential as a baseline for comparison with machine learning models.

Phase 4 - Machine Learning Models

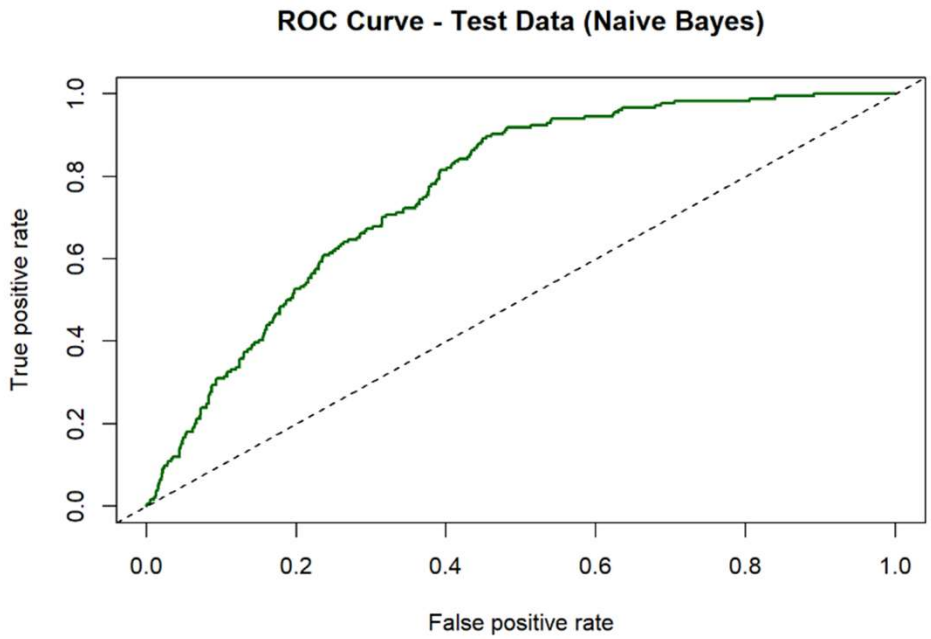
- Built four machine learning models to improve prediction accuracy of loan defaults.
- Models Applied:

  - Naive Bayes

  - Support Vector Machine (SVM)

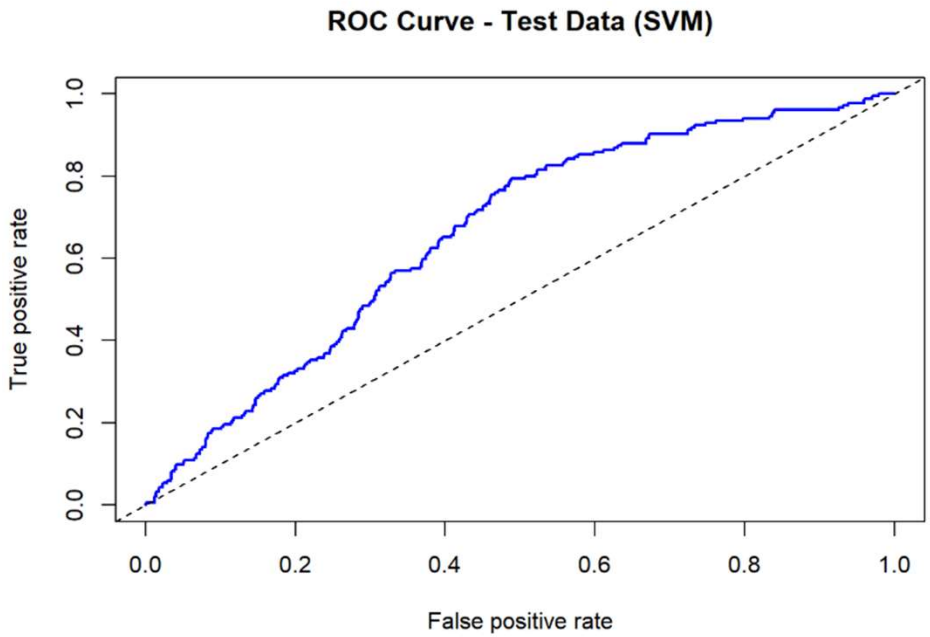
  - Decision Tree

  - Random Forest
- Goal: Compare ML models with logistic regression from Phase 3.
- Evaluated using ROC curve, AUC values and Confusion Matrices on train and test data.



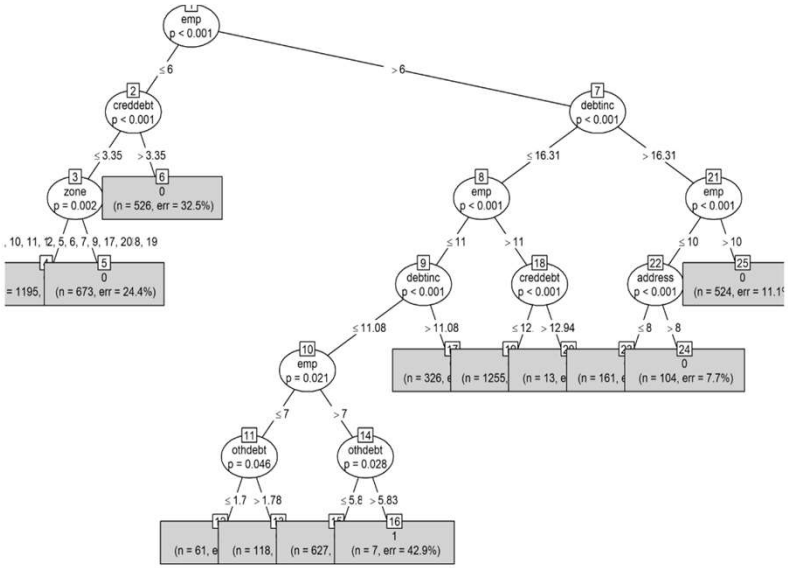
1. Naive Bayes

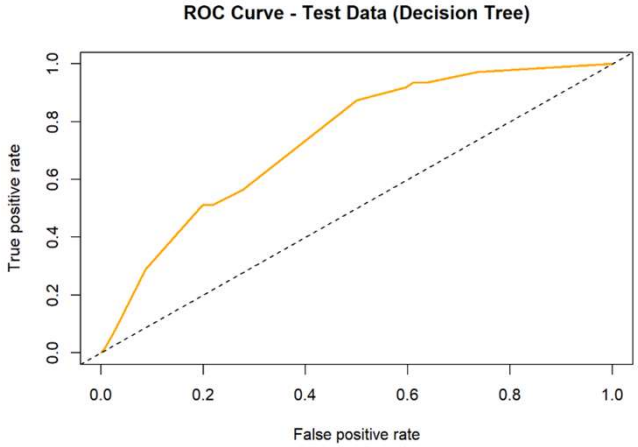
- A simple model that uses probabilities assuming features are unrelated.
- **Test AUC: *0.7665***
- Accuracy: *0.8596*
- High specificity 98.02%, but low sensitivity 6.52%



2. Support Vector Machine (SVM)

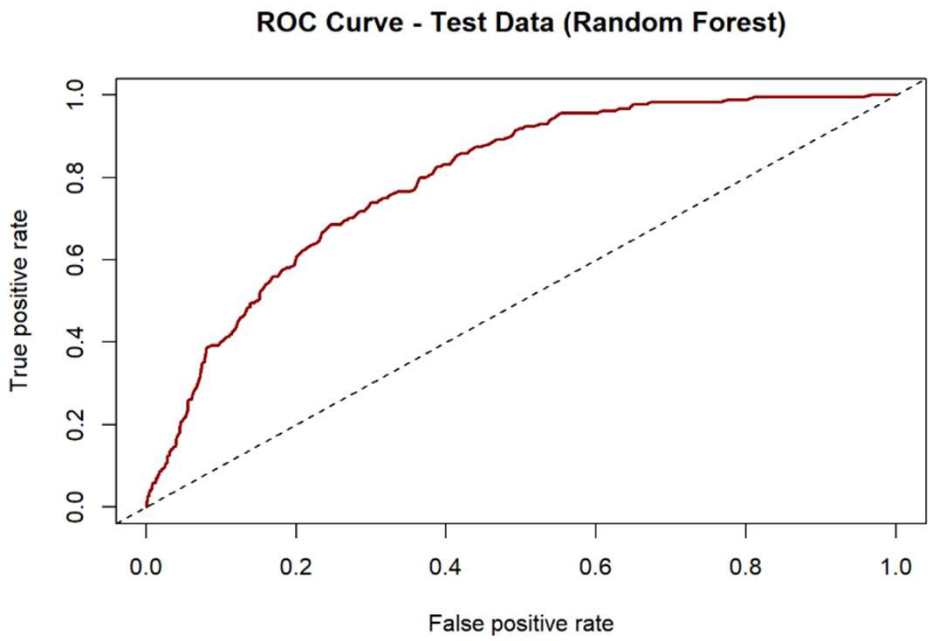
- A linear SVM model was applied to separate defaulters and non-defaulters.
- **Test AUC: *0.662***
- Accuracy: *0.8682*
- Perfect specificity (100%) but no ability to detect actual defaulters (0% sensitivity)





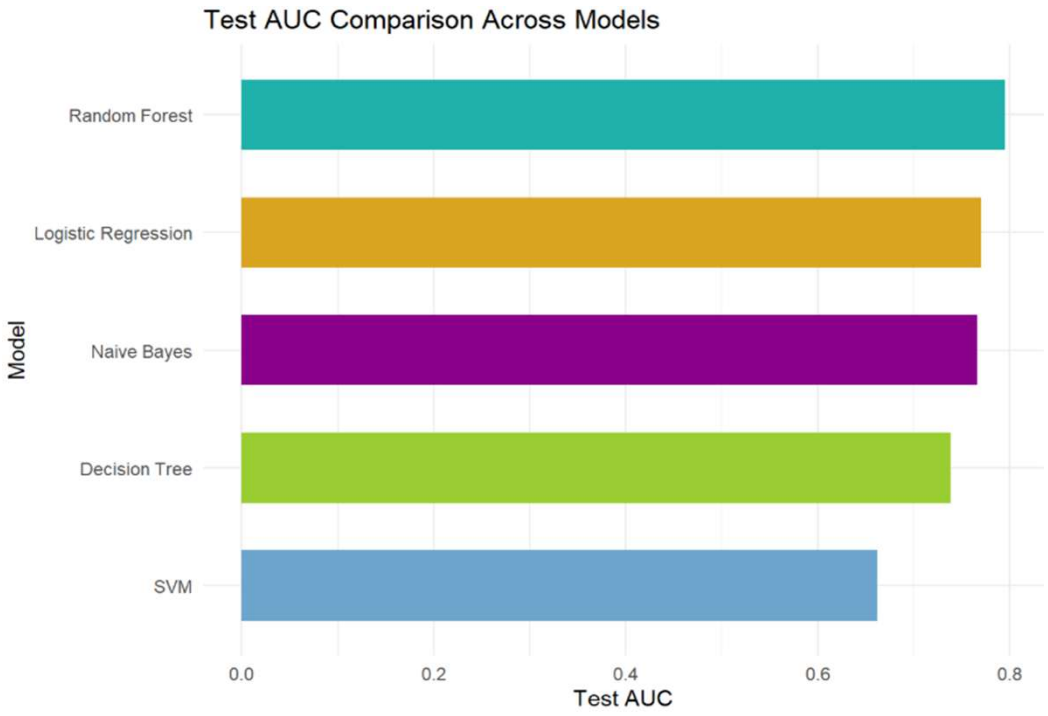
3. Decision Tree

- Easy to interpret with clear decision rules.
- Captured important patterns like debt ratio and zones.
- **Test AUC: *0.7381***
- Accuracy: *0.8646*
- High specificity (99.42%) but extremely low sensitivity (1.09%)



4. Random Forest

- Combines multiple decision trees to improve accuracy and reduce overfitting.
- **Test AUC: *0.7953***
- Accuracy: *0.8689*
- High specificity (99.59%), but low sensitivity (3.26%)

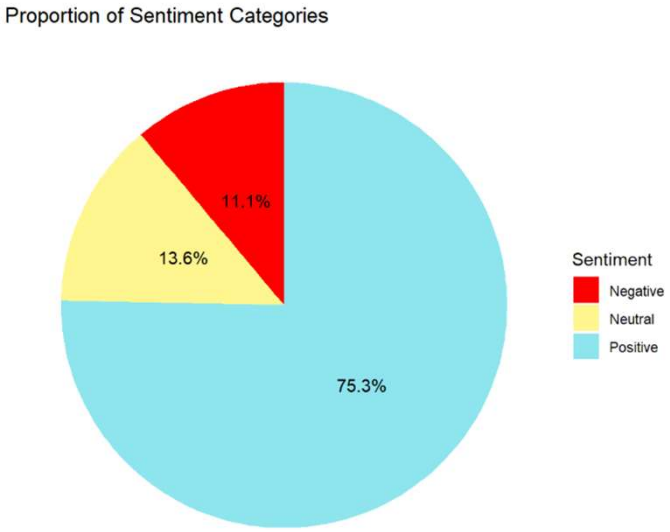


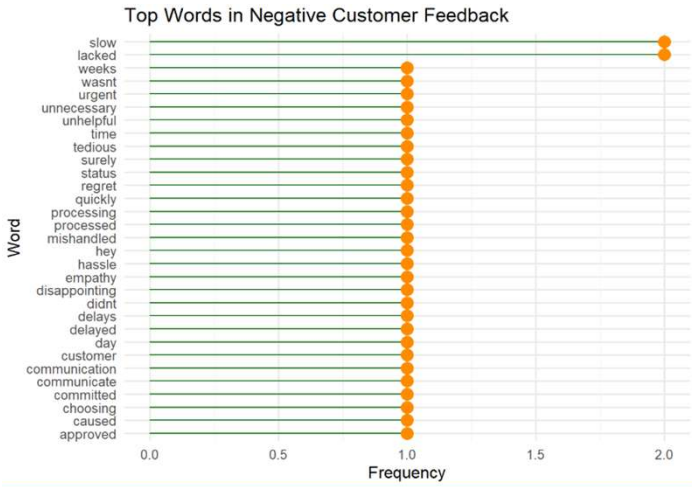
Model Comparison & Decision

- Random Forest outperformed all other models in overall predictive performance.
- Achieved the highest test AUC of **0.7953** among all models.
- Final model selected: **Random Forest**, based on highest AUC score on test data.

Phase 5 - Customer Feedback Analysis

- Analyzed customer feedback data from NPS survey responses.
- Text was cleaned by removing punctuation, numbers, stop words and irrelevant terms.
- Created a word cloud to visualize frequently mentioned words.
- Common themes included service quality, speed, staff support and loan process clarity.





Sentiment Analysis & Insights

- Sentiment scores were calculated using sentence-level analysis and averaged per feedback entry.
- Feedback classified into Positive, Neutral and Negative categories.
- Majority of the responses were positive, showing overall satisfaction.
- Negative sentiments pointed to delays and lack of proactive support.
- These insights help highlight areas for improving customer experience.

Summary

- Explored credit risk patterns using customer demographic, financial and behavioral data to understand critical factors of loan default.
- Performed in depth EDA using bad rate tables, boxplots and heat maps to identify high-risk customer segments.
- Built predictive models using binary logistic regression and four machine learning techniques. Random Forest achieved the best test AUC.
- Validated findings by comparing model outcomes with EDA insights, ensuring consistency and meaningful interpretations.
- Analyzed customer feedback using sentiment analysis, revealing mostly positive perceptions, with key improvement areas identified.

Thank you!