Presented by Analytics Arc Group :

Bank Loan Default Prediction

Parisha desai

Introduction

- **Project objective:** Developing models for bank management to predict whether the loan will be defaulted or not based on the applicant details and loan specific details
- **Purpose:** Determine factors influencing loan default and create accurate models for assessment.
- **Dataset:** Dataset from Kaggle was utilized.
- **Audience:** Bank Personnel

Research Methodology

01 What are the important factors that help determine whether a customer will default?

02 Whether the person has ability to Repay the Loan or not?

Understanding and Preprocessing:

- Understand, tidy and transform data
- Addressed missing values.
- Initial dataset: 307,511 rows and 122 columns
- After Preprocessing: 307466 rows and 60 columns
- Outlier Analysis

Exploratory Data Analysis (EDA):

- Conducted univariate and bivariate analyses.
- Drawn correlations using graphs and statistics.

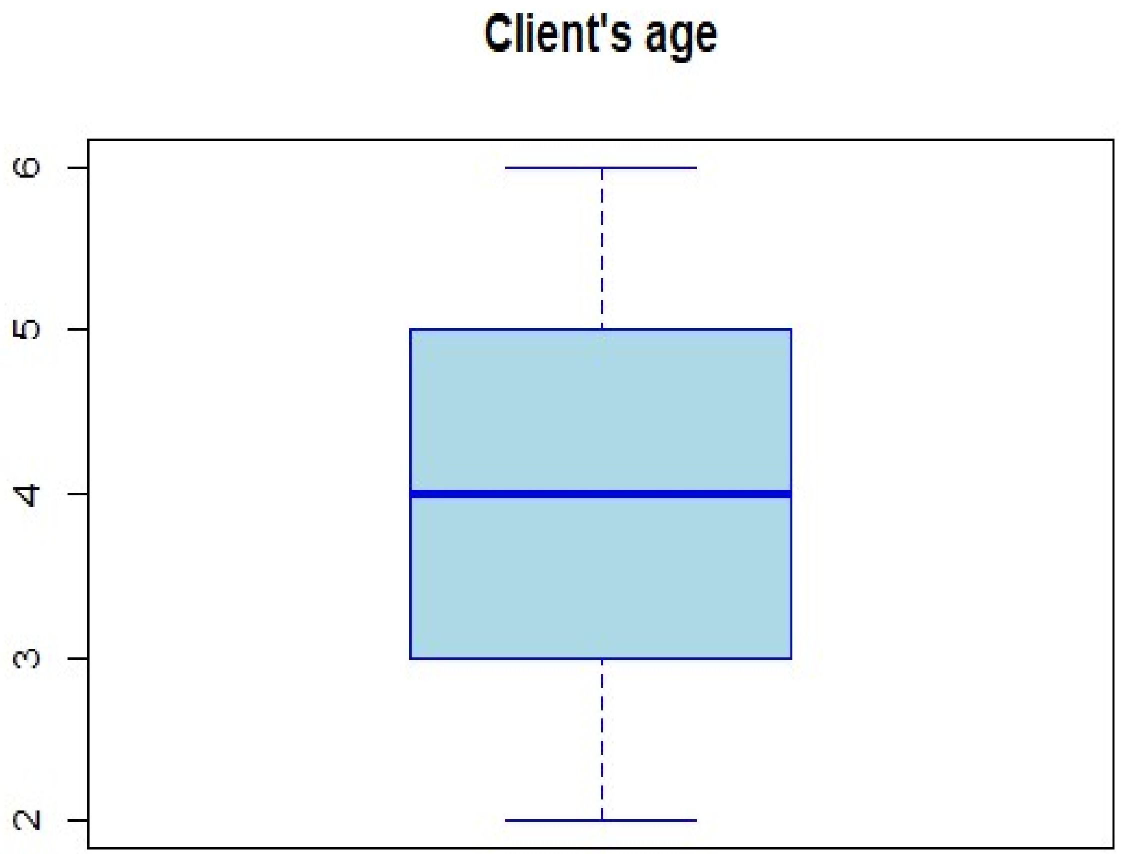
Prediction Models:

- Created Classification models based on identified parameters.
- Split the dataset into training and testing sets.

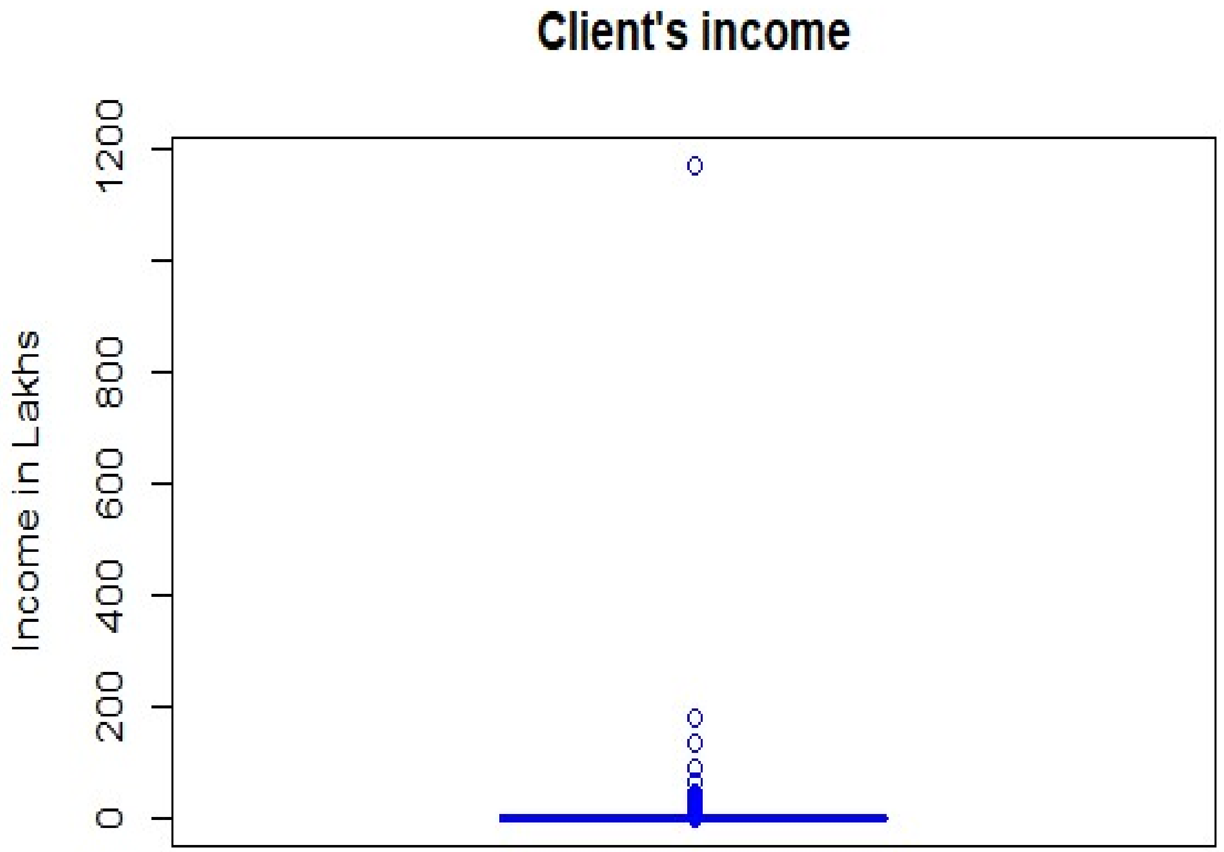
Model Evaluation and Selection:

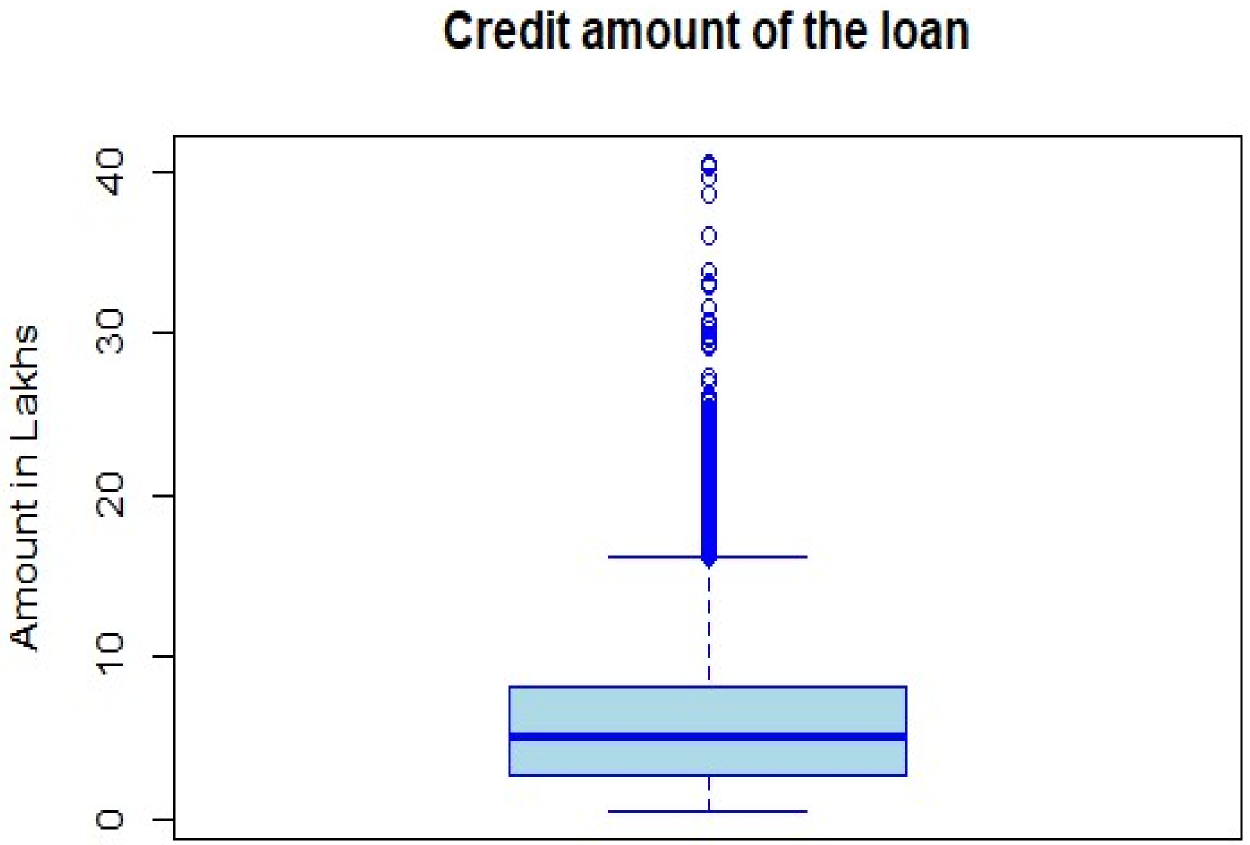
- Computed accuracy and precision.
- Analyzed performance metrics to choose the best model.

Outlier Analysis



Age 1:0-20 2:20-30 3:30:40 4:40-50 5:50-60 6:60-70





Client's age seems to have no outliers at all. No imputation or treatment required.

AMT_INCOME_TOTAL(Income of the client) shows that some of the applicants have very high income as compared to others.

AMT_CREDIT has outliers, expected due to varying loan amounts based on eligibility. Many applications fall in the lower range, below 5 lakhs.

Exploratory Data Analysis

UNIVARIATE ANAYSIS

Best customers: Females aged between 30-40, married, with a house and secondary education.

BIVARIATE ANALYSIS

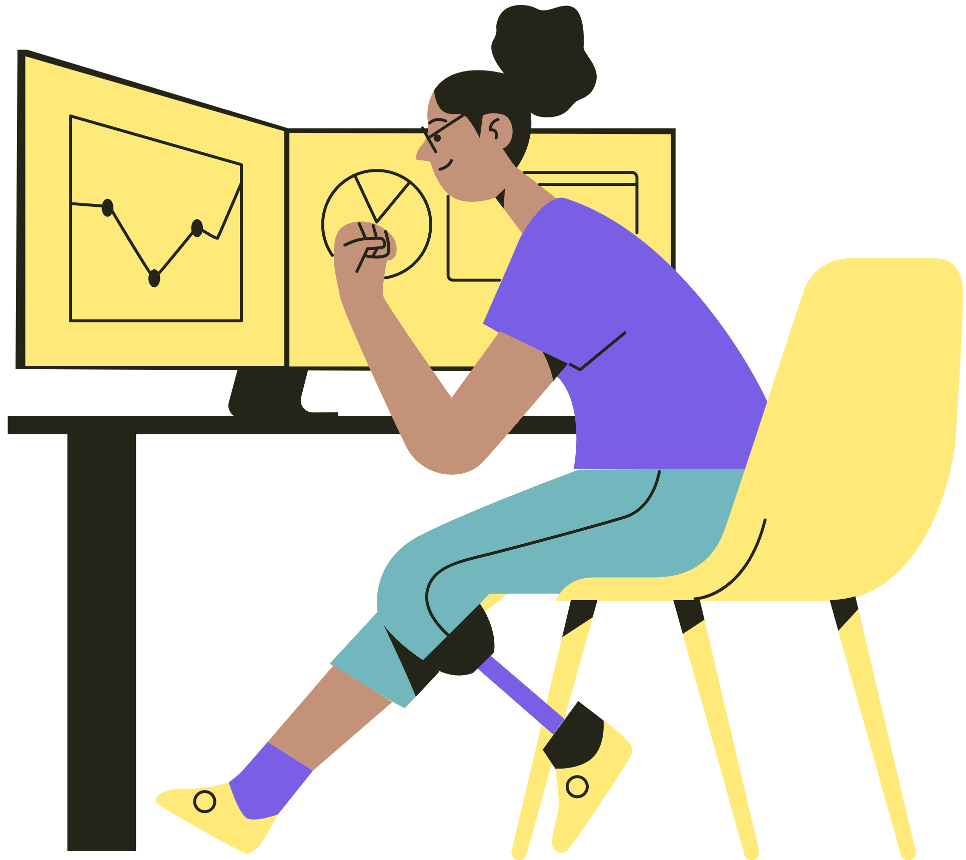
- Majority on-time payments: Females and age 30-40.
- Good payers: Lower credit, low income, working group.
- Strong Correlation: Marital status, education, and housing impact payment likelihood.

CORRELATION MATRIX INSIGHTS

- On-time payers get higher credit and better rates.
- People with Higher education had larger credit and made timely payments; secondary education faced challenges.

Age, Gender, Income Type, Education, Housing, Income Amount, Credit Amount, Marital Status And Credit Rate

Classification Models



- Logistic Regression (LR)
- Decision Tree (DT) Classification
- Naive Bayes (NB)

Logistic Regression

Analysis:

- Performs well in predicting non-default cases (class 0) with high precision (96%) and reasonable recall (69%).
- Struggles in predicting default cases (class 1) similar to Naive Bayes, with low precision (16%) and moderate recall (67%).
- High F1 score for class 0, indicating a good balance between precision and recall.
- Lower F1 score for class 1 highlights challenges in achieving a balanced trade-off between precision and recall for default cases.

Confusion Matrix

-	Predicted	
Actual	0	1
0	58338	26455
1	2434	5013

Performance by output class

	Accuracy	Precision	Recall	F1
0	0.69	0.96	0.69	0.80
1	0.69	0.16	0.67	0.26

Model Evaluation

Accuracy	0.69
Macro-precision	0.56
Macro-Recall	0.68
Macro-F1	0.53
ROC area	0.68

Decision Tree Classification

Non-Default Prediction:

- High precision (94%) and recall (80%).

Default Prediction:

- Low precision (16%) suggests many false positives.
- Relatively low recall (43%) indicates missing actual default cases.

Overall Model Evaluation:

- High F1 score for class 0 (non-default), indicating a good balance.
- Lower F1 score for class 1 (default), highlighting challenges in precision-recall trade-off.

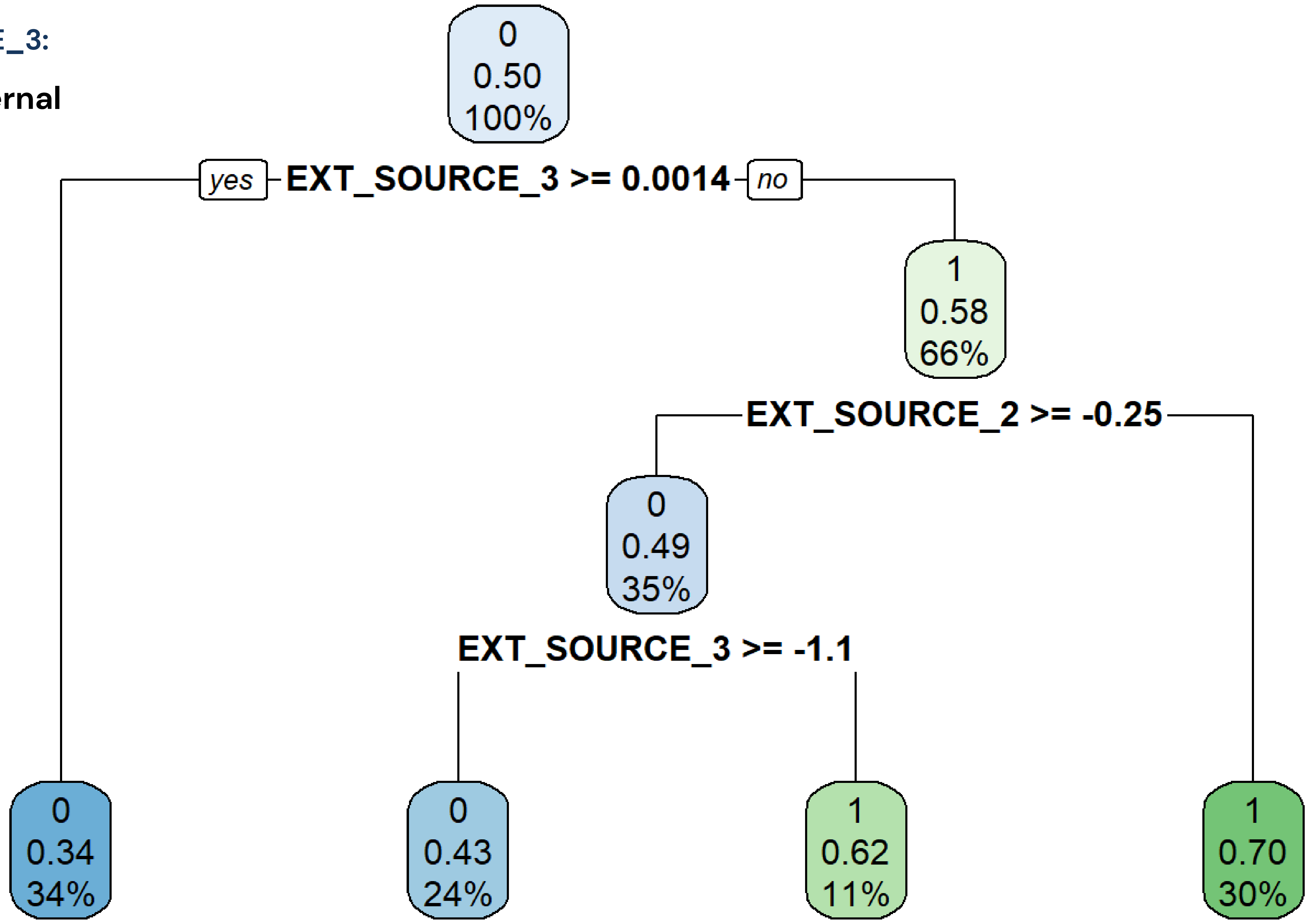
Confusion Matrix		
-	Predicted	
Actual	0	1
0	67441	17352
1	4225	3222

Performance by output class				
	Accuracy	Precision	Recall	F1
0	0.77	0.94	0.94	0.80
1	0.77	0.16	0.94	0.43

Model Evaluation	
Accuracy	0.77
Macro-precision	0.55
Macro-Recall	0.61
Macro-F1	0.55
ROC area	0.61

EXT_SOURCE_2 , EXT_SOURCE_3:

Normalized score from external data source



The Decision tree model suggests using the normalized score from external data sources (EXT_SOURCE_2 & EXT_SOURCE_3) to determine if a person has the ability to service the loan.

Naive Bayes

Naive Bayes Performance:

- Good at predicting non-default cases (precision: 95%, recall: 64%).
- Struggles with default cases: low precision (12%), moderate recall (58%).

F1 Score Insights:

- High F1 score for class 0 (non-default), indicating a good balance.
- Lower F1 score for class 1 (default), showing challenges in precision-recall trade-off.

Confusion Matrix		
-	Predicted	
Actual	0	1
0	54056	17352
1	3113	4334

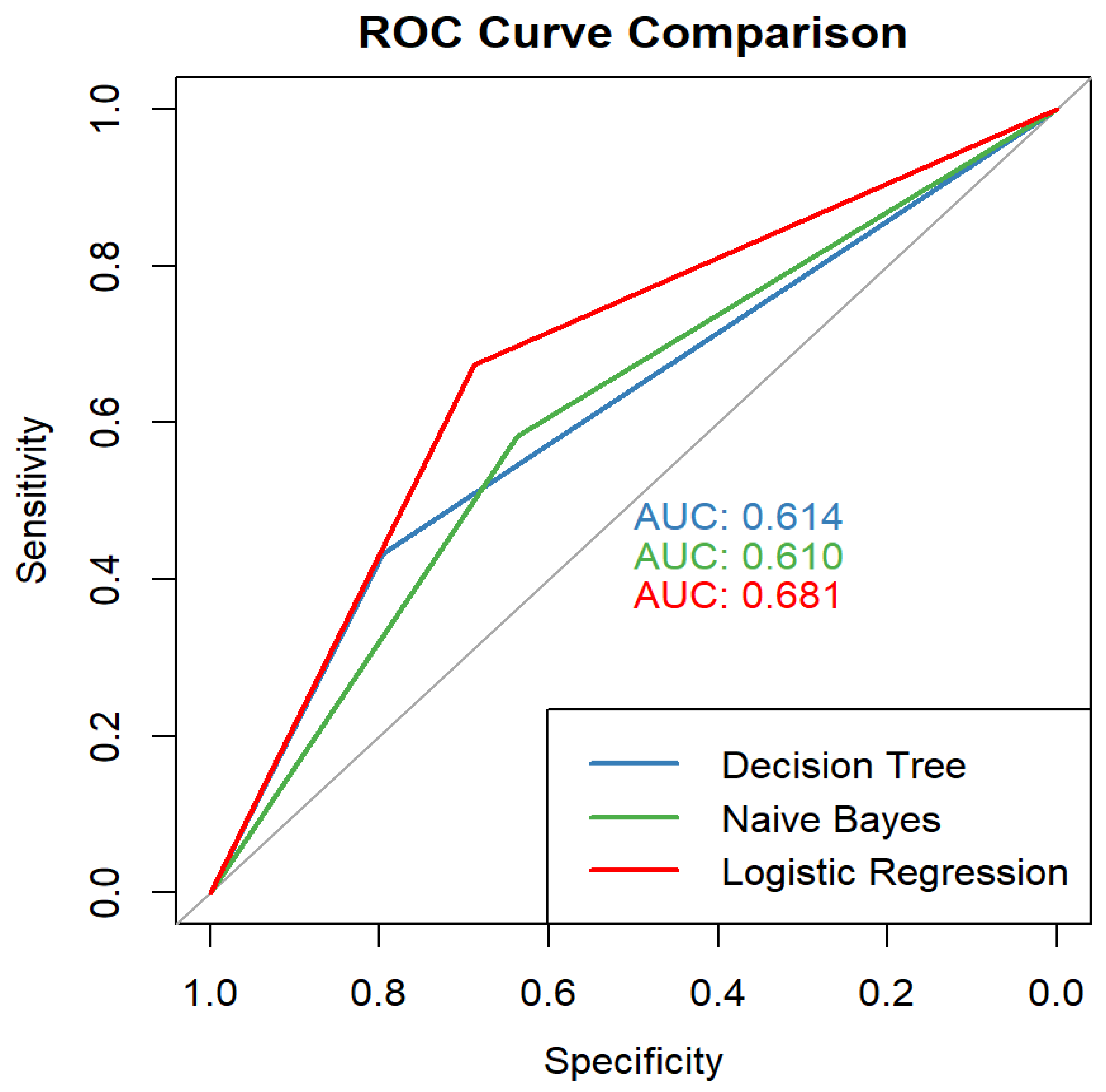
Performance by output class				
	Accuracy	Precision	Recall	F1
0	0.63	0.95	0.64	0.76
1	0.63	0.12	0.58	0.20

Model Evaluation	
Accuracy	0.63
Macro-precision	0.53
Macro-Recall	0.61
Macro-F1	0.48
ROC area	0.61

Comparison between DT, NB and LR models

Method	Accuracy	Precision	Recall	FScore	ROC
Decision Tree	0.77	0.55	0.61	0.55	0.61
Naive Bayes	0.63	0.53	0.61	0.48	0.61
Logistic Regression	0.69	0.56	0.68	0.53	0.68

01	**Decision Tree:** Highest overall accuracy but lower true positive rate for defaulters.
02	**Naive Bayes:** Similar ROC scores to Decision Tree, indicating comparable performance.
03	**Logistic Regression:** Highest ROC score, balancing precision and recall effectively.



By considering ROC and other metrics, we find logistic regression is the best model for predicting if a person has the ability to repay the loan ().

Conclusion

Key Variables for Default Prediction:

- Identified crucial factors: age, gender, income type, education, housing, income amount, credit amount, marital status and credit rate. Bank Personnel are suggested to study past trends based on these parameters.

Best Classification Model to predict Default Likelihood:

- Logistic Regression stands out as the top model for predicting loan default probability, considering accuracy, precision, F1 score, and ROC.

Dataset Limitations:

- Unable to use historical data for trend analysis.
- The dataset lacks location specificity, limiting the universal applicability of observations.

Future Scope:

- Explore regression models to determine the safe loan amount for individuals and aid bank personnel in better decision-making during loan sanctioning.

References

- Ali Abdullatif Ali Albastaki, ” Loan Default Prediction System” RIT Scholar Works, Feb 2022. https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12544&context=theses
- Utkarsh Lal,” Mastering Loan Default Prediction: Tackling Imbalanced Datasets for Effective Risk Assessment”, Medium Blog, April 2023 https://medium.com/geekculture/mastering-loan-default-prediction-tackling-imbalanced-datasets-for-effective-risk-assessment-8e8dfb2084d0
- Utkarsh,” Loan default Prediction “, Scaler Topics, May 2023 https://www.scaler.com/topics/data-science/loan-default-prediction/

Q&A

Thank You