Predicting Retail Loan Delinquency

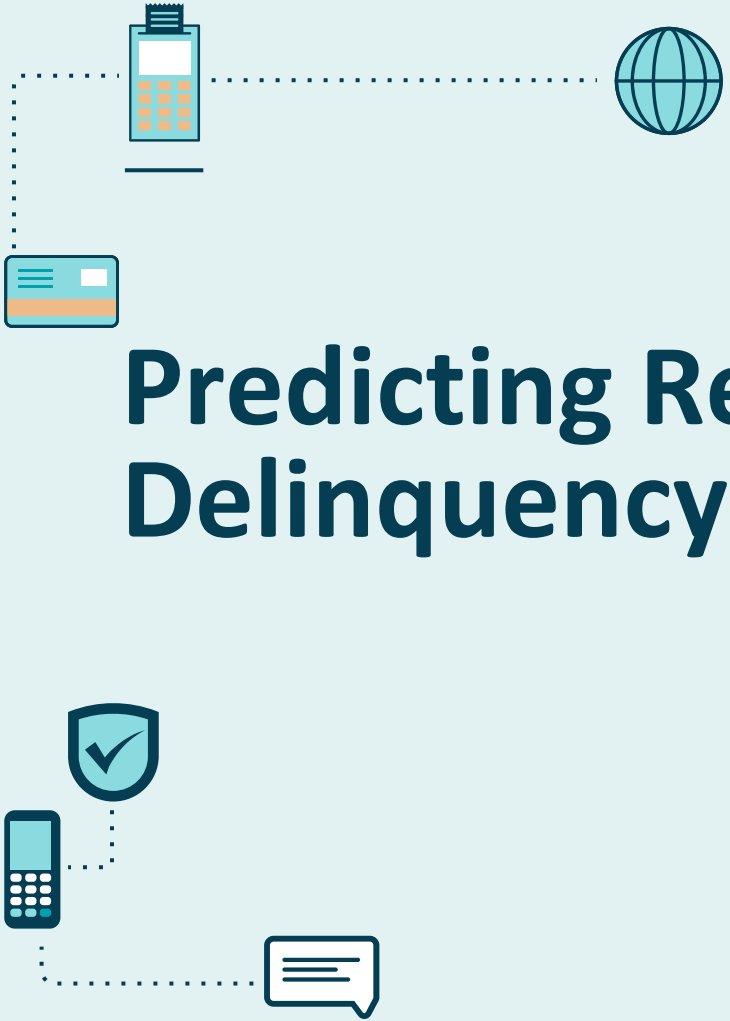


Table of Contents

01

Introduction

Indian Retail Lending Landscape

Industry |June 24

	Outstanding Balance (in Lakh Cr)	Y-o-Y Growth	Market share	90+ rates by Balance
Corporate*	78	1%	34%	15.2%
MSME*	29	10%	13%	2.0%
Retail	117	25%	51%	2.1%
MFI	4	27%	2%	1.3%
Total	**228**	**14%**		

Source: TransUnion Cibil Credit Landscape – Sept 24

Trends in India's Unsecured Lending

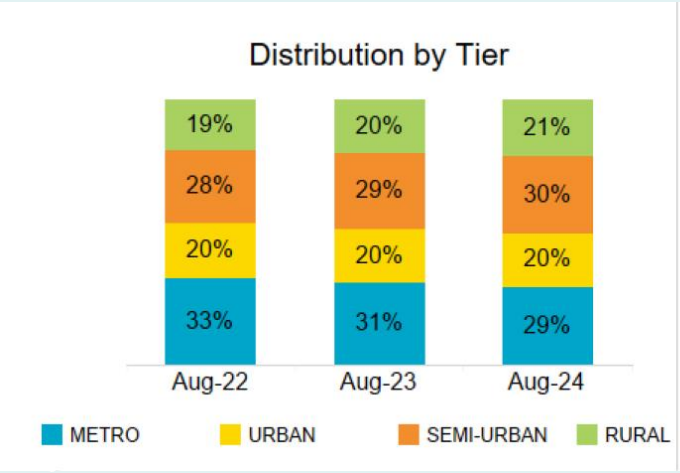
- The Indian retail lending market for unsecured loans has seen significant growth following the COVID-19 pandemic, enhanced by the rise of digital lending practices.
- Rise in Personal Loans and Credit Card Debt. There has been a substantial increase in personal loans and credit card debt, fueled by the convenience of digital platforms and shifts in consumer credit behavior.
- Data from the Reserve Bank of India (March 2023). 1

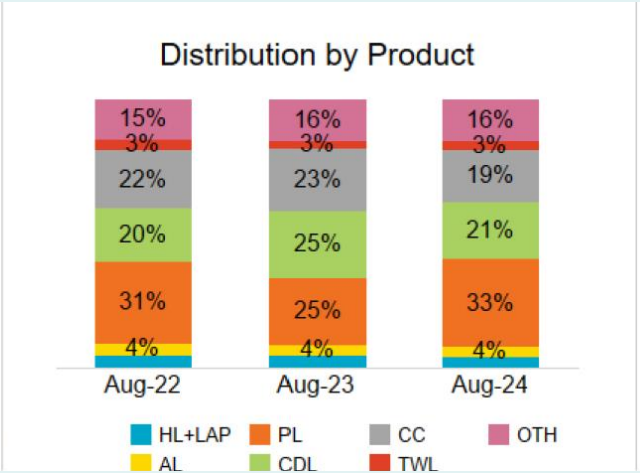
    - Credit card outstanding amounts rose by 28% to ₹2.10 lakh crore, up from ₹1.64 lakh crore in the previous year.
    - Personal loan credit increased by 22% to ₹11,00,404 crore.

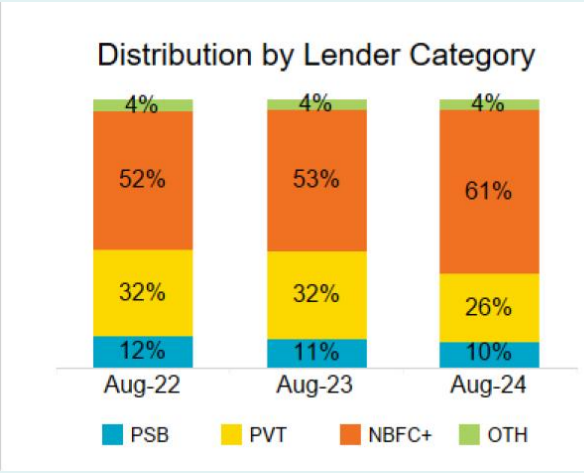
- Manageable Default Rates: Despite the rapid growth in unsecured loans, the rise in defaults has been proportional to loan disbursals, suggesting a manageable risk level.
- Continued Growth into 2024, Total credit card spending surged by 27% year-over-year to ₹18.26 trillion, influenced by the fiscal year-end and festive sales driving higher transaction volumes in March.
- The growth in unsecured loans has raised concerns regarding NPAs, given the higher risks associated with these collateral-free loans. Careful management is needed to prevent a potential increase in NPAs.

1. Source: RBI Bulletin – Jan 24
2. Source: The retail lending burden: Behind the surge in unsecured loans – Oct 23
3. Source: Unsecured loans surge but no default risks, yet – Jul 23

Retail Lending



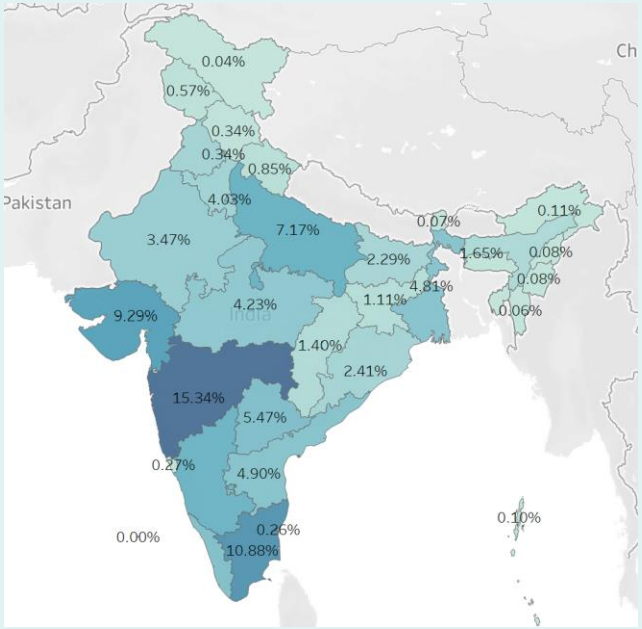




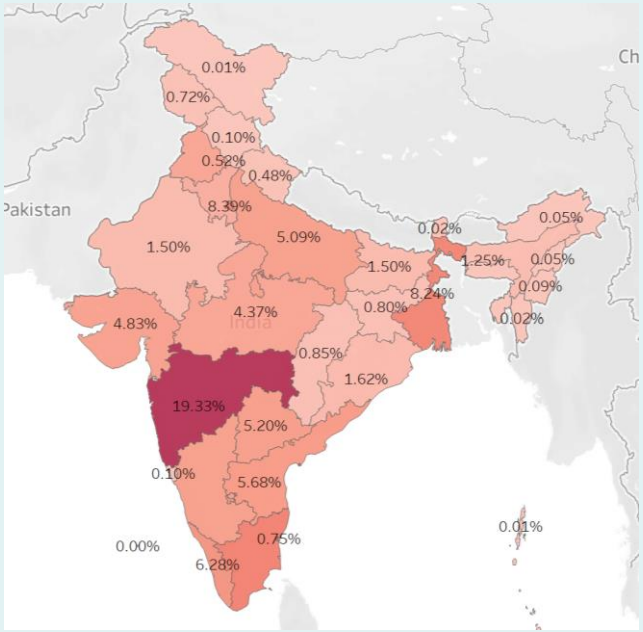
In Jun 2024, the total outstanding balance grew by 25% YoY, reaching ₹117 lakh crore. 90+ DPD by balance is 2% overall.

Source: TransUnion Cibil Credit Landscape – Sept 24

Unsecured Loans



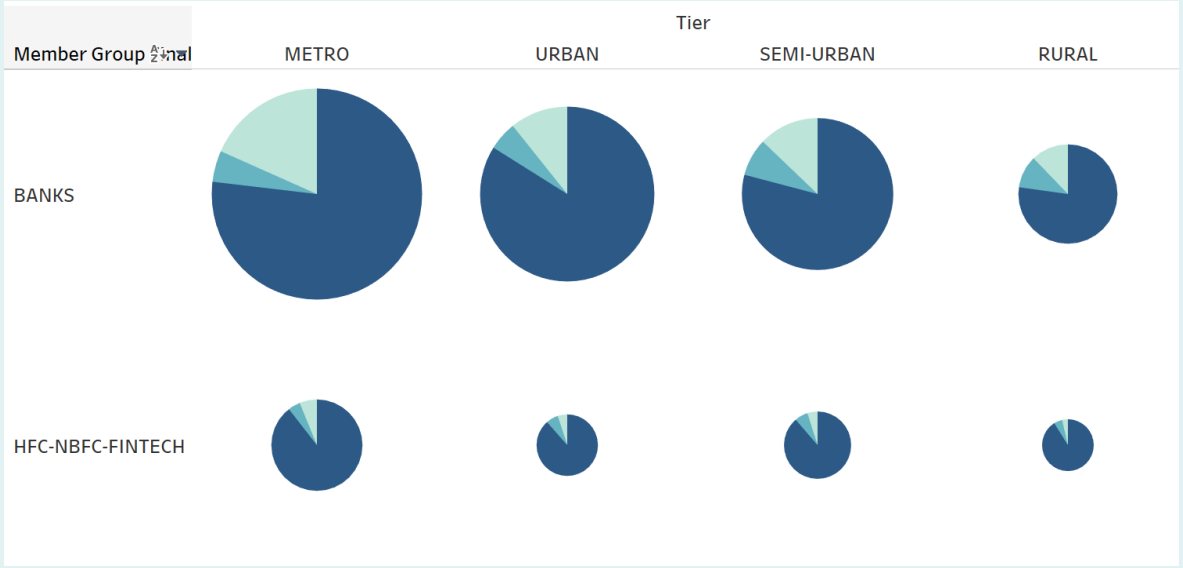
Current Balance Distribution



NPA Distribution

TU Market insight data depicted using Data Visualization

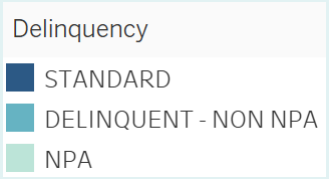
Unsecured Loans



Distribution across location tiers – Source Group wise

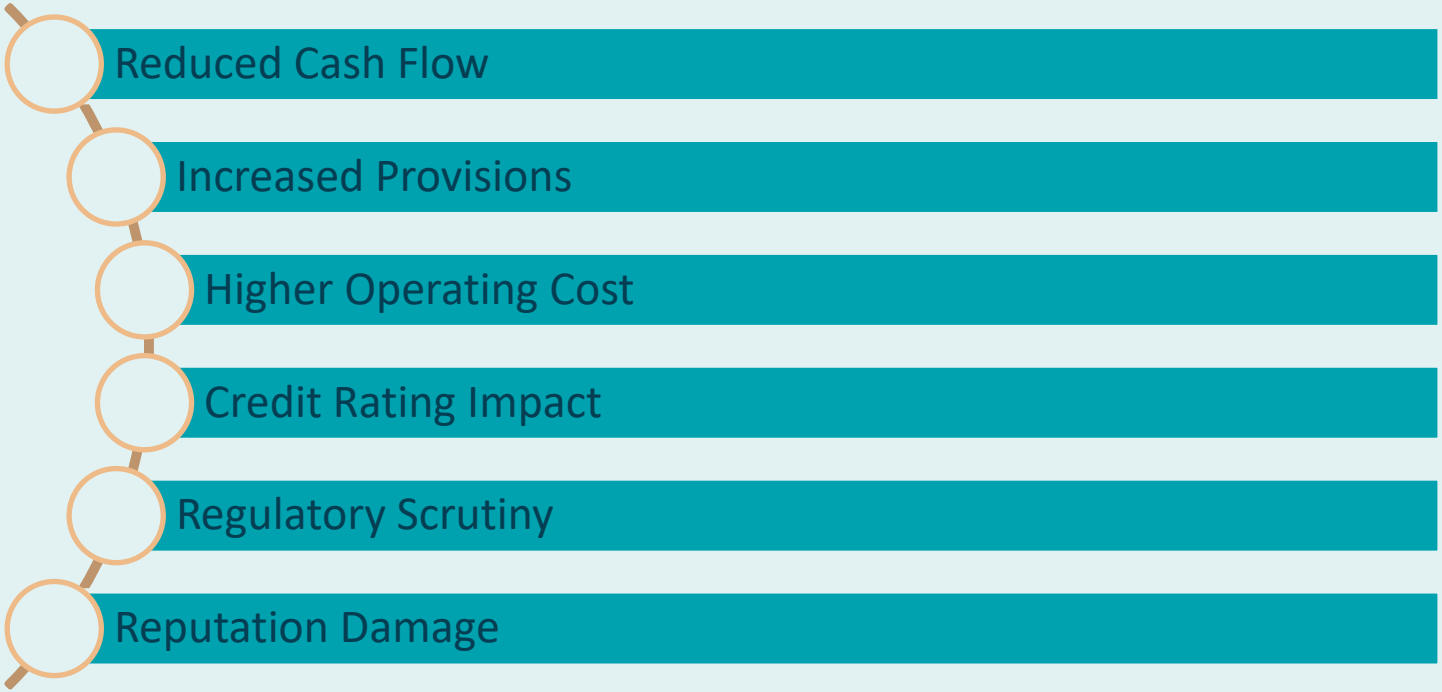
V3 Score Range	Delinquency: STANDARD	DELINQUENT - NON NPA	NPA
A.300-680	7.15%	3.16%	12.67%
B.681-700	6.65%	1.10%	0.71%
C.701-730	14.49%	1.12%	0.30%
D.731-750	17.60%	0.41%	0.03%
E.751-780	26.36%	0.25%	0.06%
F.781-800	6.93%	0.02%	0.00%
G.800+	0.98%	0.00%	0.00%

Cibil score band wise distribution of portfolio



TU Market insight data depicted using Data Visualization

Impact of Delinquency on Lenders



02

Objective

Develop a predictive model to identify unsecured loans, specifically **personal loans** and **credit cards**, at risk of delinquency using critical variables, tailored to the **Indian retail lending market**.

Problem Statement

- The consumer credit department of an Indian bank will implement guidelines aligned with local financial regulations to develop a statistically robust credit scoring model for approving unsecured loans, like personal loans and credit cards.
- The model will use data from existing loan underwriting processes, including successful applicants, to inform decision-making.
- It will leverage advanced predictive techniques, ensuring interpretability for clear decision justification, meeting regulatory standards, minimizing default risks, and promoting fairness and transparency in credit distribution.

- **Primary Objective:**
- **Perform Exploratory Data Analysis (EDA) and Feature Importance Evaluation:**
  - Conduct a comprehensive **exploratory data analysis** to identify and recommend the most influential features that the bank should consider when approving unsecured loans.
- **Develop a Predictive Classification Model:**
  - Construct a **classification model** to predict which clients are at a higher risk of defaulting on their unsecured loans. This model will utilize the identified key features from the EDA, applying advanced machine learning techniques to accurately classify potential defaulters

Dataset Overview

Sr. No	Variables	Description	Unique Values
1	ID	Unique ID of representative	
2	Loan Amount	Loan amount applied	
3	Funded Amount	Loan amount funded	
4	Funded Amount Investor	Loan amount approved by the investors	
5	Country	Country information is India	India
6	State	State information is Gujarat and Maharashtra	Maharashtra, Gujarat
7	City	Name of the city	Pune, Nagpur, Surat, Vadodara, Mumbai, Rajkot, Ahmedabad, Thane, Gandhinagar, Nashik
8	Pincode	Pincode of that city	
9	Tier	Tiers based on city population size	Urban, Metro, Semi-Urban
10	Age	Age of the customers	
11	Gender	Gender of the customers	Male, Female
12	Income	Income of the customers	
13	Term	Term of loan (in months)	
14	BatchEnrolled	Batch numbers to representatives	BAT2522922, BAT1586599, BAT2136391, BAT2428731, ... (41 unique values)
15	InterestRate	Interest rate (%) on loan	
16	Grade	Grade by the bank	B, C, F, A, G, E, D
17	SubGrade	Sub-grade by the bank	C4, D3, D4, C3, G5, C5, ... (35 unique values)
18	EmploymentDuration	Employment Duration	MORTGAGE, RENT, OWN
19	HomeOwnership_INR	Ownership of home	
20	VerificationStatus	Income verification by the bank	Not Verified, Source Verified, Verified
21	PaymentPlan	If any payment plan has started against loan	n
22	Loan Title	Loan title provided	Personal Loan, Credit Card, Education Loan, Medical Loan, Business Loan, Home Improvement Loan, Consumer Durable Loan, Debt Consolidation Loan, Festival Loan
23	Debit to Income	Ratio of representative's total monthly debt repayment divided by self-reported monthly income excluding mortgage	

Dataset period - **2007 to 2014**

Dataset Overview

Sr. No	Variables	Description	Unique Values
24	Delinquency – two years	Number of 30+ days delinquency in past 2 years	
25	Inquires six months	Total number of inquiries in last 6 months	
26	Open Account	Number of open credit line in representative's credit line	
27	Public Record	Number of derogatory public records	
28	Revolving Balance	Total credit revolving balance	
29	Revolving Utilities	Amount of credit a representative is using relative to revolving balance	
30	Total Accounts	Total number of credit lines available in representatives credit line	
31	Initial List Status	Unique listing status of the loan - W (Waiting), F (Forwarded)	w, f
32	Total Received Interest	Total interest received till date	
33	Total Received Late Fee	Total late fee received till date	
34	Recoveries INR	Post charge-off gross recovery	
35	Collection Recovery Fee	Post charge-off collection fee	
36	Collection 12 months Medical	Total collections in last 12 months excluding medical collections	
37	Application Type	Indicates whether the representative is an individual or joint	INDIVIDUAL, JOINT
38	Last week Pay	Indicates how long (in weeks) a representative has paid EMI after batch enrolled	
39	Accounts Delinquent	Number of accounts on which the representative is delinquent	
40	Total Collection Amount	Total collection amount ever owed	
41	Total Current Balance	Total current balance from all accounts	
42	Total Revolving Credit Limit	Total revolving credit limit	
43	Credit Score	Credit score of the borrower	
44	Loan Status	Loan Status: 1 = Defaulter, 0 = Non-Defaulter	1 = Defaulter, 0 = Non-Defaulter

Dataset Structure

- Programming Language Used: Python
- Data Source: Loan default data from Kaggle , Original source : URL: https://www.kaggle.com/datasets/kalaikumarr/loan-data
- Data set count: Dataset consist of **44 columns** and a total count of **67,463 records**

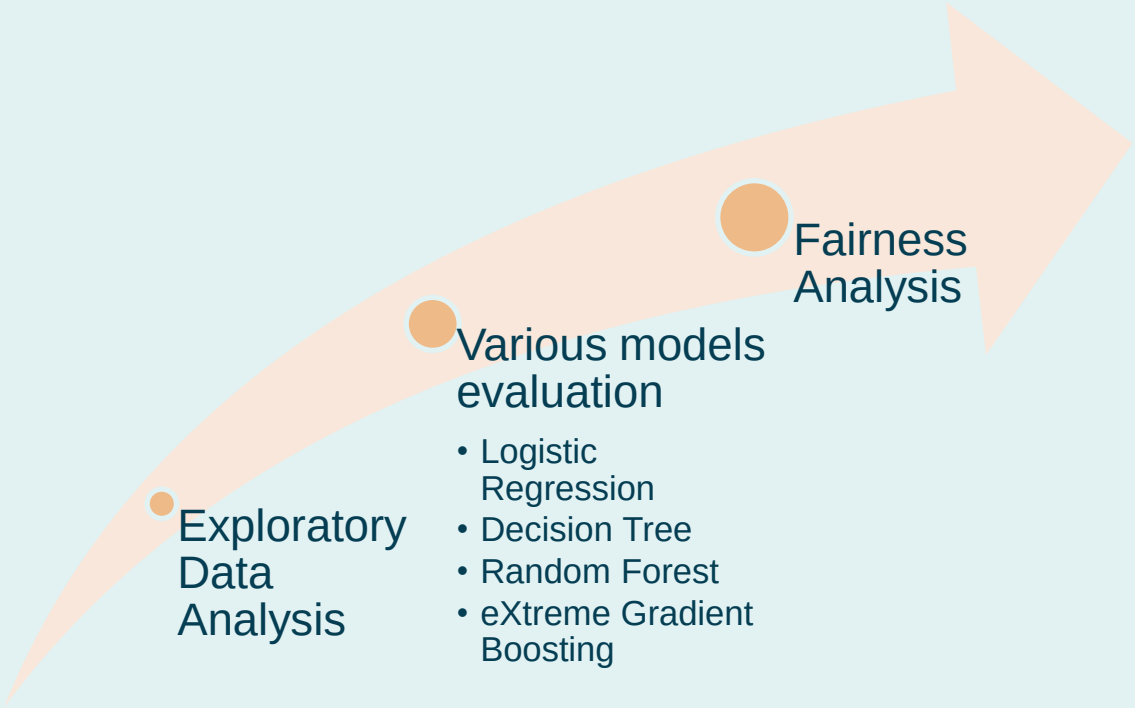
Integer	Float	Categoric
ID	Loan Amount	Country
Pincode	Funded Amount	State
Age	Funded Amount Investor	City
Income	Interest Rate	Tier
Term	Home Ownership	Gender
Delinquency - two years	Debit to Income	Batch Enrolled
Inquires - six months	Revolving Balance	Grade
Open Account	Revolving Utilities	Sub Grade
Public Record	Total Received Interest	Employment Duration
Total Accounts	Total Received Late Fee	Verification Status
Collection 12 months Medical	Recoveries	Payment Plan
Last week Pay	Collection Recovery Fee	Loan Title
Accounts Delinquent	Total Collection Amount	Initial List Status
Credit Score	Total Current Balance	Application Type
Loan Status	Total Revolving Credit Limit	

- Loan Status: 1 = Defaulter; 0 = Non-Defaulter

03

Methodology

Mode of Operation



04

Exploratory Data Analysis

Exploratory Data Analysis

Collection

Import of various Libraries & upload of data set

Cleaning

Removing errors and irrelevant data

Preparation

Reduction - Variables with very low distinct value excluded

Statistical

Distribution analysis & Correlation analysis done

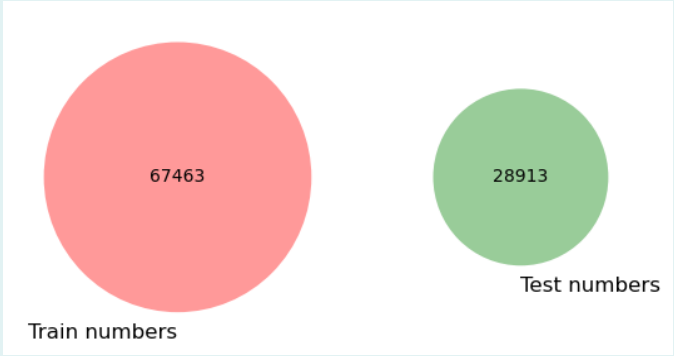
Visualization

Correlation analysis & Multicollanearity Check done

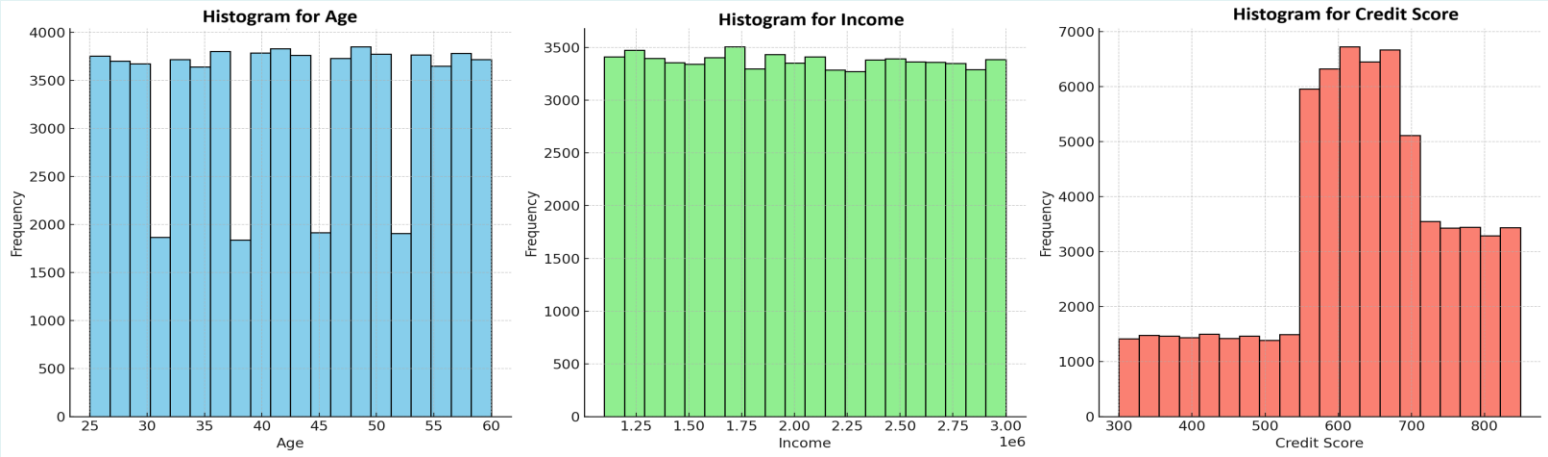
Hybrid Sampling & Spilt

Hybrid sampling, Transformation, Train data spilt into training & testing (70:30)

Exploratory Data Analysis



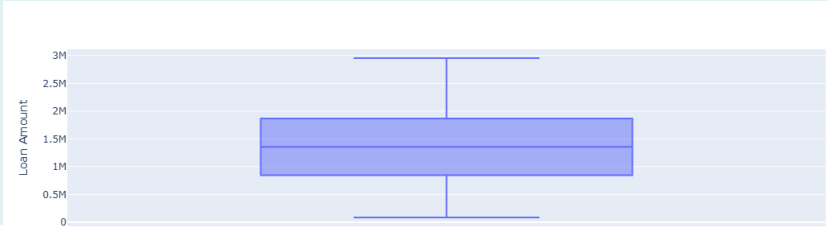
The ratio between the training and testing datasets is 70:30

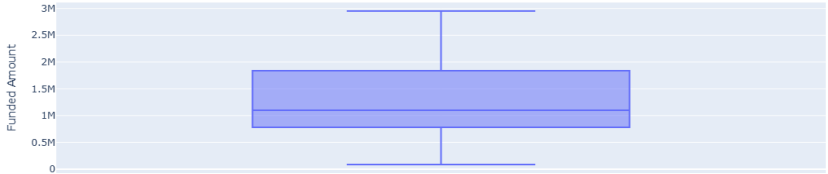


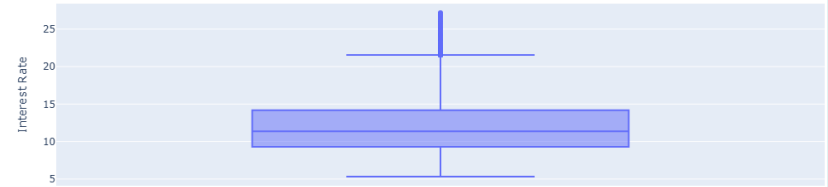
Variables with low distinct values

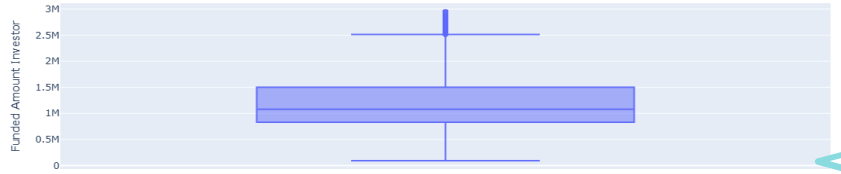
Variable	Distinct Values	Observation	Recommendation
Accounts Delinquent	0: 28,885, 1: 28	Highly skewed, with almost all records having a value of 0.	Exclude from the model due to lack of variability and predictive power.
Payment Plan	n: 28,913	Only one unique value across all records.	Exclude from the model as it provides no discriminative information.
Application Type	INDIVIDUAL: 28,883, JOINT: 30	Highly skewed, with "INDIVIDUAL" dominating the records.	Retain for now combining into one label unless "JOINT" is critical.
Collection 12 Months Medical	0: 28,347, 1: 566	Skewed distribution, but the minority class (1) could provide predictive insights.	Retain for now; analyze further to confirm its predictive value, as the minority class might correlate with defaults.

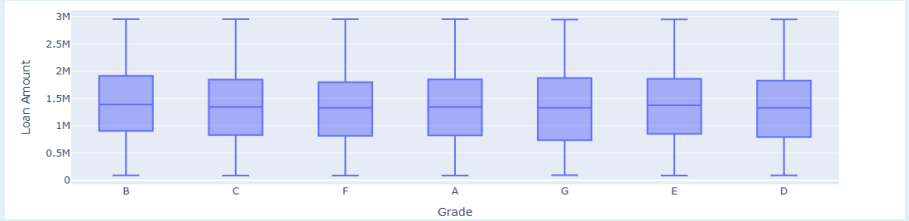
Visually exploring the distribution and variability

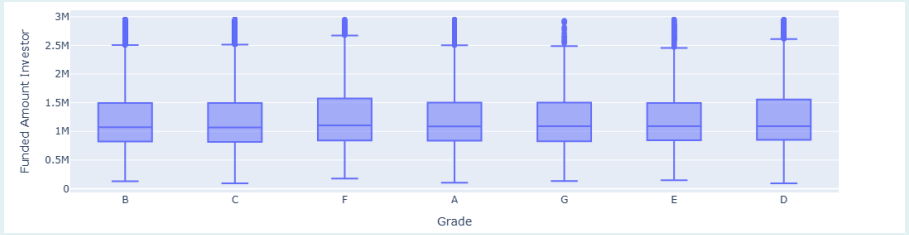


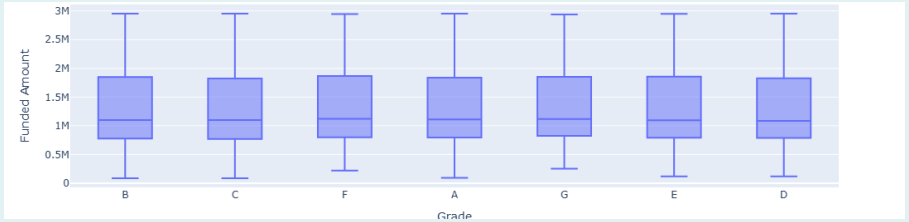


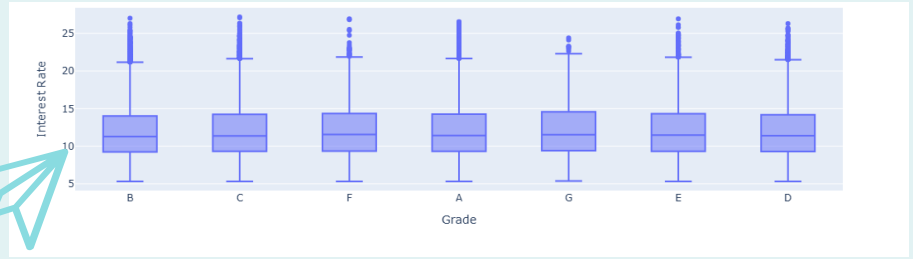




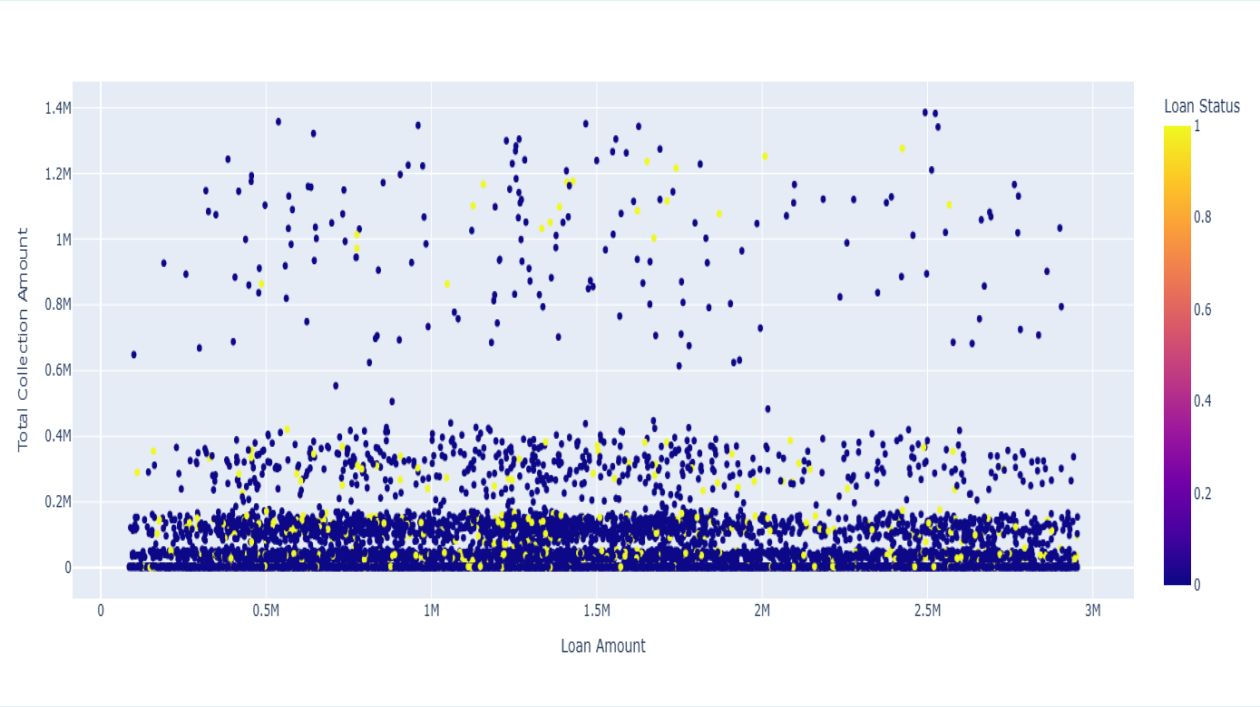








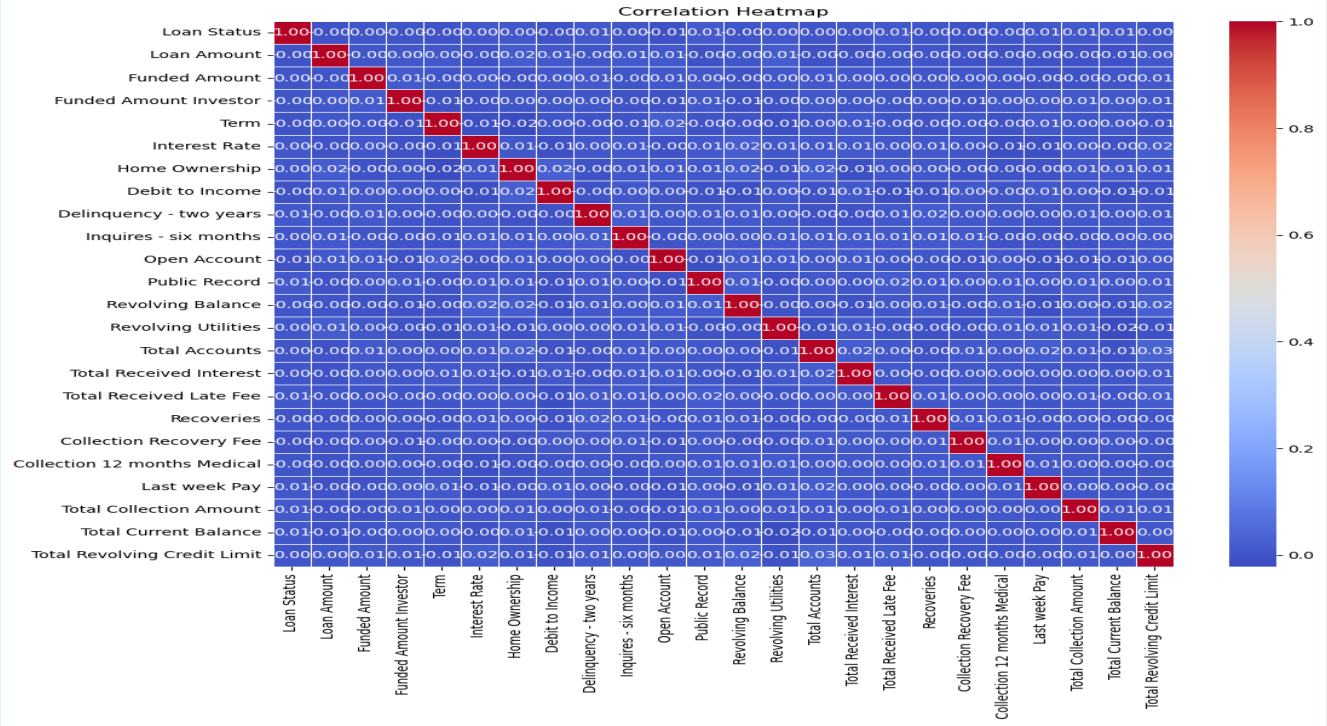
Scatter Plot



- Most data points cluster near the lower end of the Total Collection Amount axis, suggesting that many loans have low collection amounts regardless of the loan size.

- As loan amounts increase, the variability in **Total Collection Amount** grows. This indicates that datapoints that are in yellow region are more defaulter in compared to blue region.

Correlation Heatmap



- After performing correlation and doing careful analysis we have retained only the highly correlated variables for modelling based on Multicollinearity.

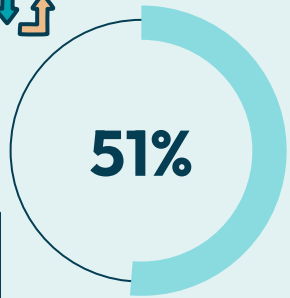
Insights from EDA

- "**Accounts Delinquent**," "**Payment Plan**," due to low variability need not be taken for modelling
- **Batch BAT1135695** has high default rate
- **Subgrade G3** has high default rate
- Loan title "**Debt Consolidation Loan"** and **"Consumer Durable Loan“** have high default rate
- **Delinquency two years** and **Inquires six months** as potential predictors as they show variability that might influence loan defaults.
- Borrowers with 4 **Inquiries in six months** exhibit the highest default rate.
- People with 4 **Public Records** have higher default rate
- People with higher **total collect amount** have higher default rate

05

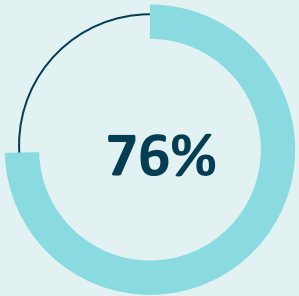
Model Evaluation

Model Comparision



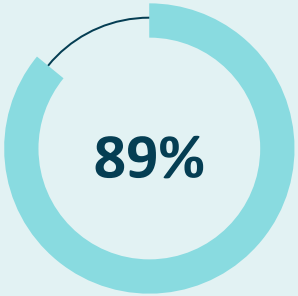
Logistic Regression

Accuracy: 0.5093
Precision: 0.5103
Recall: 0.4555
Specificity: 0.5631
Sensitivity: 0.4555
Confusion Matrix:
[[1406 1091]
[1359 1137]]



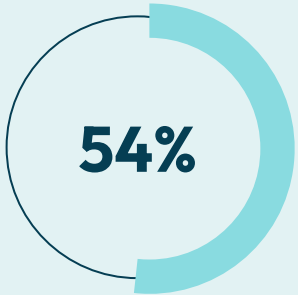
Decision Tree

Accuracy: 0.7659
Precision: 0.7196
Recall: 0.8710
Specificity: 0.6608
Sensitivity: 0.8710
Confusion Matrix:
[[1650 847]
[322 2174]]



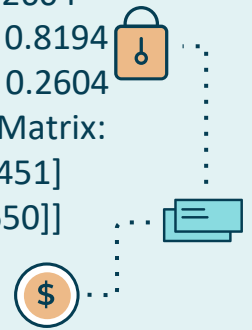
Random Forest

Accuracy: 0.8890
Precision: 0.9746
Recall: 0.7989
Specificity: 0.9792
Sensitivity: 0.7989
Confusion Matrix:
[[2445 52]
[502 1994]]

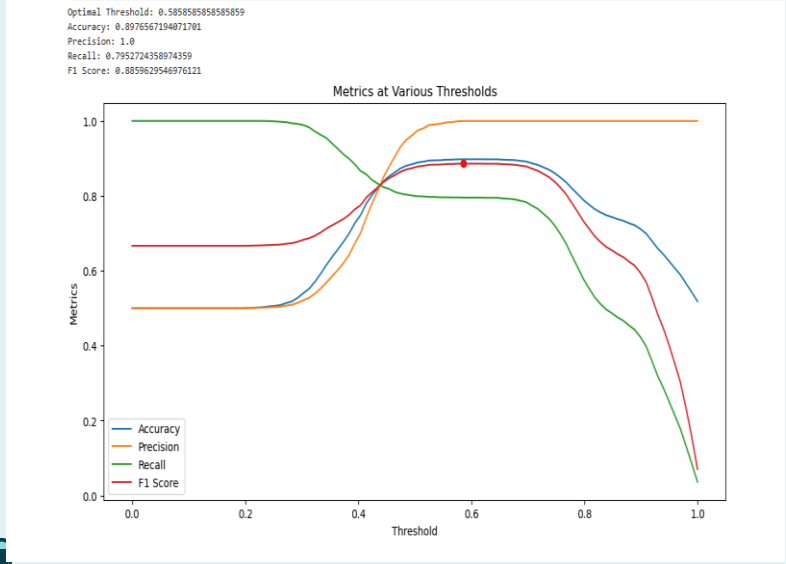


XGBoost

Accuracy: 0.5400
Precision: 0.5904
Recall: 0.2604
Specificity: 0.8194
Sensitivity: 0.2604
Confusion Matrix:
[[2046, 451]
[1846, 650]]



Random Forest Confusion Matrix



1. Optimal Threshold (0.5858)
2. Deploy the model with its current strong metrics of 89% accuracy, 88% F1-Score.
3. Random Forest Metrics:

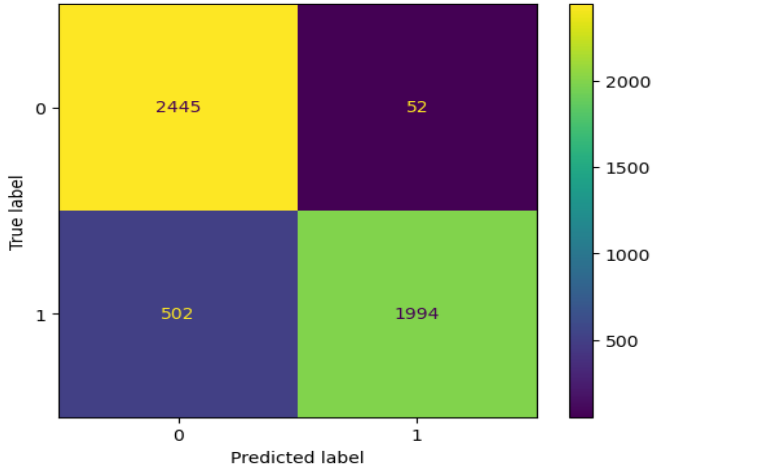
Accuracy: 0.8890 **Recall:** 0.7989 **Precision:** 0.9746
Specificity: 0.9792 **Sensitivity:** 0.7989

Random Forest Test Metrics:

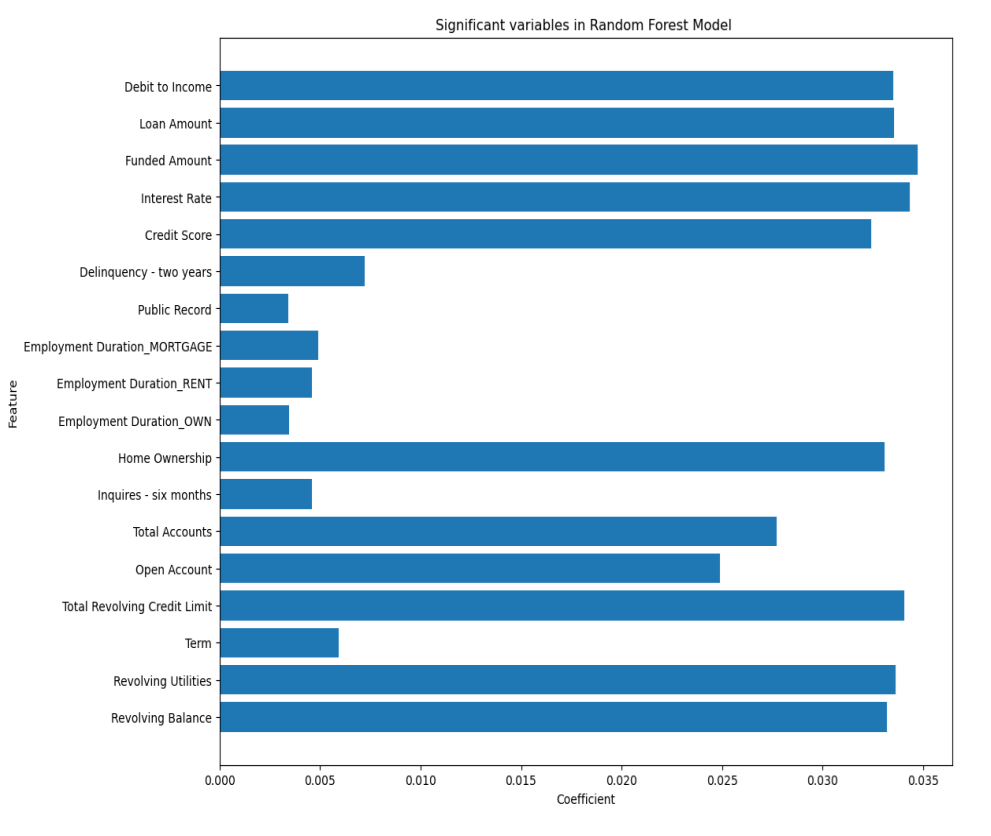
	precision	recall	f1-score	support
0	0.83	0.98	0.90	2497
1	0.97	0.80	0.88	2496
accuracy			0.89	4993
macro avg	0.90	0.89	0.89	4993
weighted avg	0.90	0.89	0.89	4993

Confusion Matrix on Test Set:
[[2445 52]
[502 1994]]

Out[87]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x247ea95f8c0>

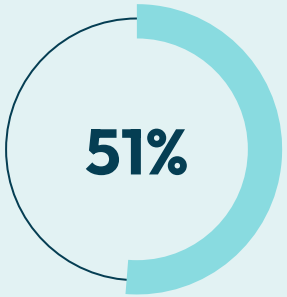


Insights from Random Forest

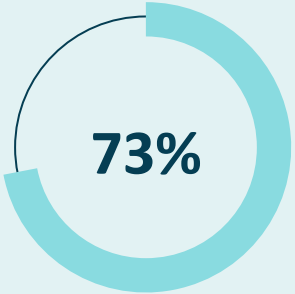


Sr. No	Variable	Significance and Insights
1	**Debit to Income**	Highly Significant: Reflects financial stress; ensure meaningful variation across the dataset.
2	**Loan Amount**	Highly Significance: Include realistic distributions correlating moderately with default likelihood.
3	**Funded Amount**	Highly Significance: Maintain consistency with Loan Amount while varying slightly.
4	**Interest Rate**	Highly Significant: Focus on realistic interest rate ranges and their impact on repayment.
5	**Credit Score**	Highly Significant: Use a range accurately reflecting default risks.
6	**Home Ownership**	Moderate Significance: Ensure the data reflects ownership status variations.
7	**Total Accounts**	Moderate Significant: Maintain basic data for modeling completeness.
8	**Open Account**	Moderate Significant: Include values that reflect real-world scenarios without overemphasis.
9	**Total Revolving Credit Limit**	Moderate Significant: Include a broad range with realistic limits.
10	**Revolving Utilities**	Moderate Significant: Use realistic values but with minimal focus.
11	**Revolving Balance**	Moderate Significant: Maintain data integrity without detailed emphasis.
12	**Delinquency two years**	Moderate Significance: Reflect a balanced range of delinquencies.
13	**Term**	Moderate Significance: Reflect loan terms with variability.
14	**Inquiries six months**	Moderate Significant: Model realistic inquiries indicative of financial behavior.
15	**Employment Duration MORTGAGE**	Less Significance: Capture variations across different employment lengths.
16	**Employment Duration RENT**	Less Significance: Capture variations across different employment lengths.
17	**Employment Duration OWN**	Less Significance: Capture variations across different employment lengths.
18	**Public Record**	Less Significance: Ensure realistic values to understand external financial records' influence.

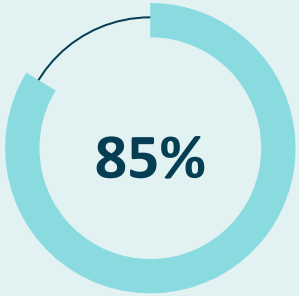
Cross Validation



Logistic Regression



Decision Tree



Random Forest

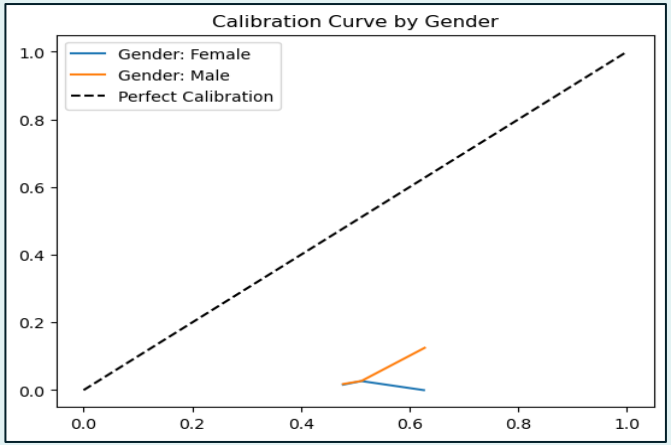
06

Fairness Analysis

Fairness and Explainability in Predictive Modeling

Why It Matters??

- **Fairness**
  - **Equitable Lending Practices:** Ensures gender- and age-neutral loan approvals.
  - **Regulatory Compliance:** Adheres to RBI guidelines, avoiding bias.
  - **Trust & Reputation:** Fair outcomes build customer and stakeholder trust.
  - **Example:** Predicted positive rates:
  - Females: **12.43%**, Males: **12.33%** (Demographic Parity).
- **Explainability**
  - **Transparent Decisions:** Models must justify predictions.
  - **Stakeholder Confidence:** Clear insights for executives.
  - **Customer Education:** Borrowers understand and improve eligibility.
  - **Example:** LIME identifies **Total Collection Amount** and **Credit Score** as key factors.



Fairness and Explainability in Predictive Modeling

Key Principles Based on Regulatory Guidelines

- **Fairness**
  - **Demographic Parity:** Equal positive outcomes across groups.
  - **Equalized Odds:** TPR & FPR parity (Females TPR: **18.97%**, Males TPR: **17.69%**).
  - **AIR Compliance:** Achieved at **1.0080**.
- **Explainability**
  - **Global Interpretability:** Features like Debt-to-Income Ratio drive insights.
  - **Local Interpretability:** LIME highlights individual prediction factors.
  - **Actionable Insights:** Refine policies for high-risk borrowers.

Methods to Ensure Fairness & Explainability

- **Fairness Analysis:**
  - Balanced gender predictions; subgroup analysis confirms accuracy (Females 31–44: **87.16%**).
- **Explainability Analysis:**
  - **Feature Importance:** Debt-to-Income Ratio, Credit Score, Funded Amount.
  - **Tools Used:** LIME and Partial Dependence Plots (PDPs) for non-linear feature effects.

07

Recommendation

Recommendation

- **Debt-to-Income Ratio, Credit Score, Interest Rate, Loan Amount:** These are the most significant indicators of default risk. Focus on these for better risk stratification.
- Offer **financial counseling** to borrowers with high Debt-to-Income Ratios, guiding them for regular repayment of the loan.
- Borrowers with **Debt-to-Income Ratio, Credit Score, Delinquency two years, Public Record, Loan Amount, Funded Amount, Inquiries six months** consistently show a high default likelihood. Implement strict thresholds for these metrics as potential "hard stops" for loan approvals.
- Review exceptional cases where these thresholds are exceeded to understand the rationale behind approvals and refine decision-making policies.

Recommendation

- Use strict cutoffs for **delinquencies** to avoid high-risk approvals.
- Use insights from explainability tools to tailor loan terms, interest rates, and eligibility criteria to individual borrower profiles.
- Ensure **equal opportunities for all demographic groups**, avoiding systematic exclusions while addressing business growth opportunities.
- Model showing Accuracy: 88.9% ensures reliable loan default predictions.
- Model showing Precision: 97.46% minimizes false approvals, protecting against high-risk loans.
- The model demonstrates fairness across demographics. Use this as a foundation to maintain trust and compliance with regulatory standards.
- Leverage explainability tools like feature importance (global) and LIME (local) to provide clear, transparent decision justifications to borrowers.

Thanks!

Do you have any questions?