CREDIT RISK EVALUATOR

Non-Banking Financial Company (NBFC) Credit Risk Modeling

Presented by Vaibhav Garg

PROBLEM STATEMENT

- Credit risk evaluation is complex due to the influence of diverse and interrelated factors like financial behavior, loan terms, and credit history.
- Traditional/manual assessment methods lack consistency and scalability, leading to delays and subjective decision-making.
- There is a high need for a standardized scoring system that can fairly and accurately quantify creditworthiness across varied applicant profiles.
- An automated, ML-based approach is essential to improve prediction accuracy, enable real-time decisions, and support high-volume processing.

PROJECT OBJECTIVES

- Develop an automated credit evaluation system that predicts the likelihood of loan default using machine learning.

- Generate a standardized credit score (e.g., 300–900) that reflects an applicant’s creditworthiness.

- Categorize applicants into clear risk tiers such as Poor, Average, Good, and Excellent for easy interpretation.

- Deliver real-time results via an interactive Streamlit app, allowing users to input applicant data and instantly receive:
  - Probability of default
  - Credit score
  - Risk category

BUSINESS REQUIREMENTS

- Recall (on Default Class) > 90%
  - The model must identify the majority of potential defaulters, even at the cost of some false positives. High recall ensures risky applicants are not missed.

- Precision > 50%
  - While not the top priority, maintaining a precision above 50% helps reduce false alarms and keeps the model practically useful.

- AUC > 85
  - A high AUC indicate strong model performance in distinguishing defaulters from non-defaulters.

- KS Statistic > 40 with Peak in First 3 Deciles
  - The Kolmogorov–Smirnov statistic should exceed 40% for production readiness. Ideally, the highest KS value should occur in the top 3 deciles to ensure early identification of high-risk applicants.

- Model Interpretability
  - Transparency is essential. The model must be easily understandable by business users to ensure trust, adoption, and compliance with regulatory standards.

DATA COLLECTION

Customers Table
Customer ID
Age
Gender
Marital Status
Employment Status
Income of the Customer
Number of dependents
Residence Type
Years at present address
City
State
Zipcode/Pincode

Loans Table
Loan ID
Customer ID
Loan Purpose
Loan Type
Sanction Amount
Loan Amount
Processing Fee
GST
Net Disbursement (Amount Disbursed in Customer's Account)
Loan Tenure in Months
POS (Principal Outstanding) (BookSize of Customer)
Bank Balance at application
Disbursed Date
Installment start date
Default **(Default / No Default)**

Target Variable ->

Bureau Table
Customer ID
Number Of Open Accounts (Total Number of open accounts till date)
Number Of Closed Accounts (Total Number of closed accounts till date)
Total Loan in months
Delinquent Months (Total delinquent in months)
Total DPD (Total Due Passed Day)
Total Enquiry count
Credit Utilization Ratio

DATA SPLITTING TO AVOID LEAKAGE

- Why it matters:
  - Prevents model from learning future information, avoiding unrealistic performance.

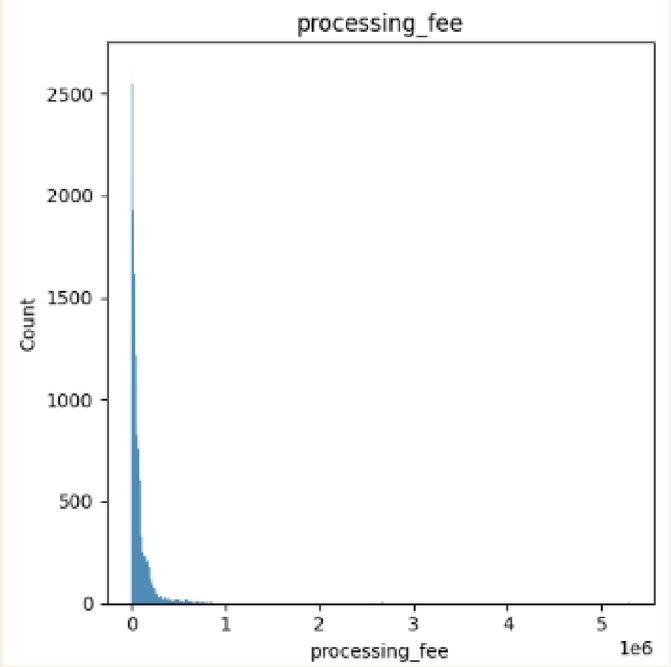
- Types of Leakage:
  - Target Leakage: Future outcome data accidentally used in training.
  - Train-Test Contamination: Preprocessing applied before splitting the data.

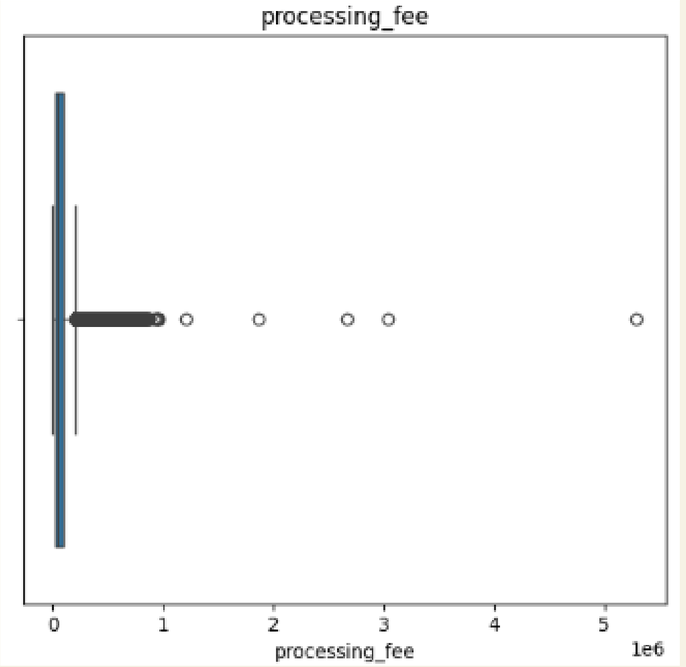
- Our Approach:
  - Used 75% Train – 25% Test split.
  - Split was performed before any preprocessing, scaling, or feature engineering.

DATA PREPROCESSING

- Data Cleaning
  - Imputed missing values in the residence_type column using the mode.
  - Dropped duplicate rows.

- Numerical Features Analysis:
  - Box Plots and Histograms were used to detect and visualize outliers.
  - Observed that the processing_fee column was highly right-skewed and compressed, indicating potential outliers.





- Removed the records where the processing_fee exceeded 3% of the loan_amount, as per business rules.

- Categorical Feature Analysis
  - Checked unique values in each categorical column.

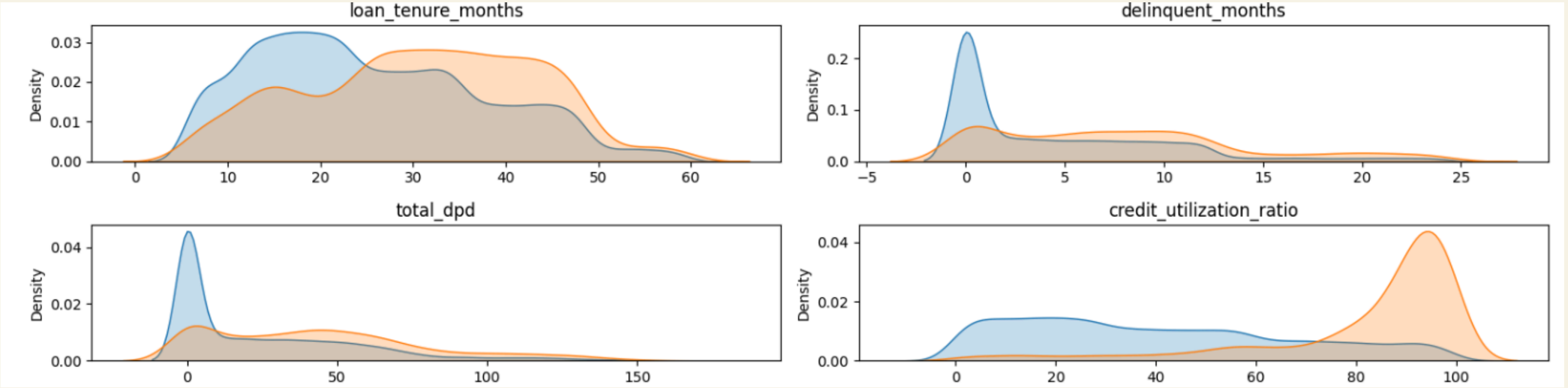
```
gender --> ['M' 'F']
marital_status --> ['Married' 'Single']
employment_status --> ['Self-Employed' 'Salaried']
residence_type --> ['Owned' 'Mortgage' 'Rented']
city --> ['Hyderabad' 'Mumbai' 'Chennai' 'Bangalore' 'Pune' 'Kolkata' 'Ahmedabad'
 'Delhi' 'Lucknow' 'Jaipur']
state --> ['Telangana' 'Maharashtra' 'Tamil Nadu' 'Karnataka' 'West Bengal'
 'Gujarat' 'Delhi' 'Uttar Pradesh' 'Rajasthan']
zipcode --> [500001 400001 600001 560001 411001 700001 380001 110001 226001 302001]
loan_purpose --> ['Home' 'Education' 'Personal' 'Auto' 'Personaal']
loan_type --> ['Secured' 'Unsecured']
default --> [0 1]
```

  - Fixed inconsistent entries in the loan_purpose column by replacing 'Personaal' with 'Personal'.

EXPLORATORY DATA ANALYSIS (EDA)

- Plotted KDE graphs for all numerical features to visualize their distribution against the target variable.

- Observed that higher values in the following features correlated with a greater likelihood of default:
  - loan_tenure_months
  - delinquent_months
  - total_dpd
  - credit_utilization

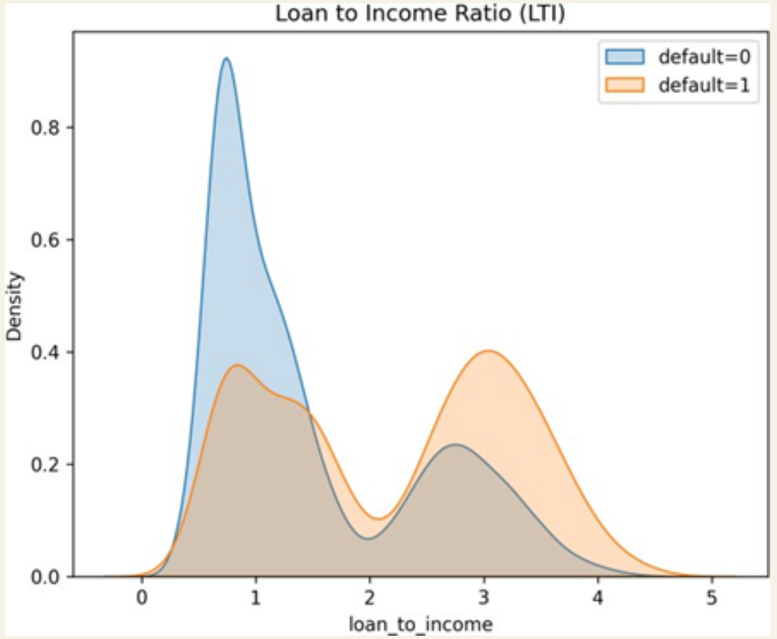


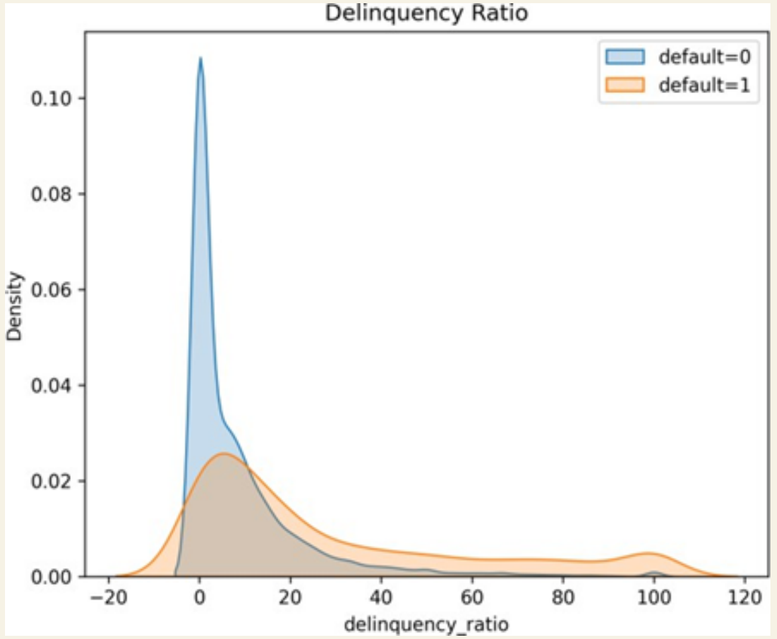
- These features were identified as strong predictors of default.

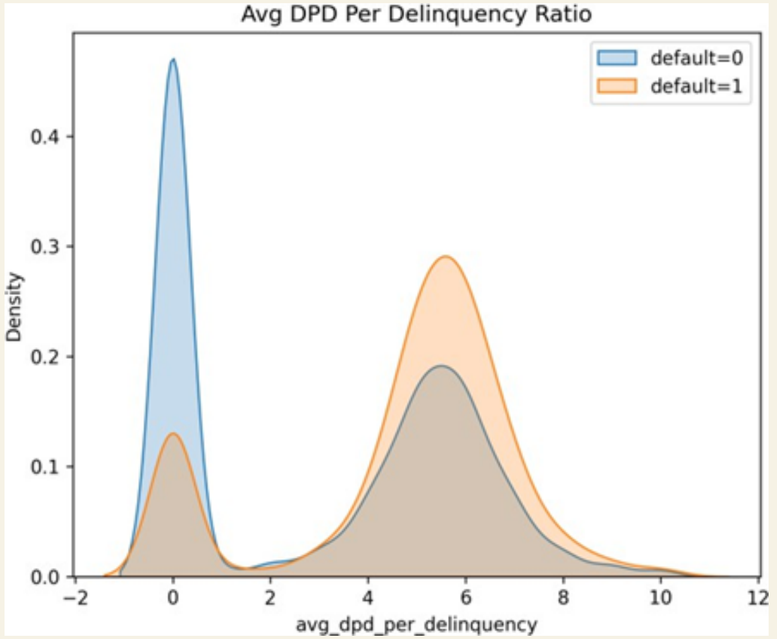
- Most other features did not show significant separation between default and non-default distributions.

FEATURE ENGINEERING

- Created new features based on business recommendations to improve model interpretability and predictive strength.
  - Loan-to-Income (LTI) Ratio – Higher values indicate increased default risk.
  - Delinquent-to-Tenure Ratio – Normalizes delinquency by loan duration; higher ratios show higher risk.
  - Average DPD per Delinquent Month – Captures delay severity; strongly correlates with defaults.







- Based on Technical & Business Knowledge
  - Removed cust_id and loan_id as they are identifiers with no predictive value.
  - Dropped features like disbursal_date, installment_start_dt, loan_amount, income, total_loan_months, delinquent_months, and total_dpd based on business input to avoid redundancy and leakage risk.

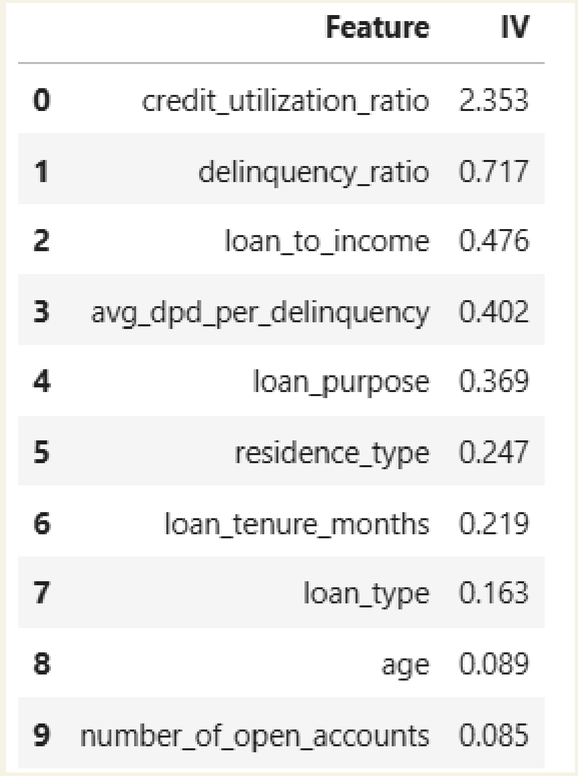
- Checked Multicollinearity using Variance Inflation Factor (VIF)
  - Calculated VIF scores for all numerical features after scaling them using MinMaxScaler.
  - Dropped features with VIF > 10, such as sanction_amount, processing_fee, gst, net_disbursement, principal_outstanding and recalculated VIF.

	Column	VIF
0	age	5.55
1	number_of_dependants	2.72
2	years_at_current_address	3.36
3	sanction_amount	101.08
4	processing_fee	inf
5	gst	inf
6	net_disbursement	inf
7	loan_tenure_months	6.17
8	principal_outstanding	16.32
9	bank_balance_at_application	9.33
10	number_of_open_accounts	4.38
11	number_of_closed_accounts	2.36
12	enquiry_count	6.33
13	credit_utilization_ratio	2.90
14	loan_to_income	6.89
15	delinquency_ratio	1.93
16	avg_dpd_per_delinquency	2.90

	Column	VIF
0	age	5.27
1	number_of_dependants	2.72
2	years_at_current_address	3.34
3	loan_tenure_months	6.01
4	bank_balance_at_application	1.80
5	number_of_open_accounts	4.35
6	number_of_closed_accounts	2.35
7	enquiry_count	6.30
8	credit_utilization_ratio	2.88
9	loan_to_income	4.54
10	delinquency_ratio	1.93
11	avg_dpd_per_delinquency	2.90

- Information Value (IV) Filtering
  - Applied binning where necessary to prepare features for IV computation.
  - Calculated IV for both numerical and categorical features.
  - Retained only features with IV > 0.02 to ensure strong predictive power, transparency, and compliance with credit scoring standards.

	Feature	IV
0	credit_utilization_ratio	2.353
1	delinquency_ratio	0.717
2	loan_to_income	0.476
3	avg_dpd_per_delinquency	0.402
4	loan_purpose	0.369
5	residence_type	0.247
6	loan_tenure_months	0.219
7	loan_type	0.163
8	age	0.089
9	number_of_open_accounts	0.085
10	enquiry_count	0.008
11	bank_balance_at_application	0.006
12	employment_status	0.004
13	years_at_current_address	0.002
14	number_of_dependants	0.002
15	city	0.002
16	zipcode	0.002
17	state	0.002
18	number_of_closed_accounts	0.001
19	marital_status	0.001
20	gender	0.000



	Feature	IV
0	credit_utilization_ratio	2.353
1	delinquency_ratio	0.717
2	loan_to_income	0.476
3	avg_dpd_per_delinquency	0.402
4	loan_purpose	0.369
5	residence_type	0.247
6	loan_tenure_months	0.219
7	loan_type	0.163
8	age	0.089
9	number_of_open_accounts	0.085

- Applied One-Hot Encoding to Nominal Features
  - Converted non-ordinal categorical features into binary columns using one-hot encoding.

MODEL TRAINING

- Trained three baseline models on original imbalanced data:
  - Logistic Regression – Chosen for interpretability and regulatory alignment.
  - Random Forest – Used for its ensemble strength and ability to capture non-linearity.
  - XGBoost – Included for its robustness and performance on tabular data.

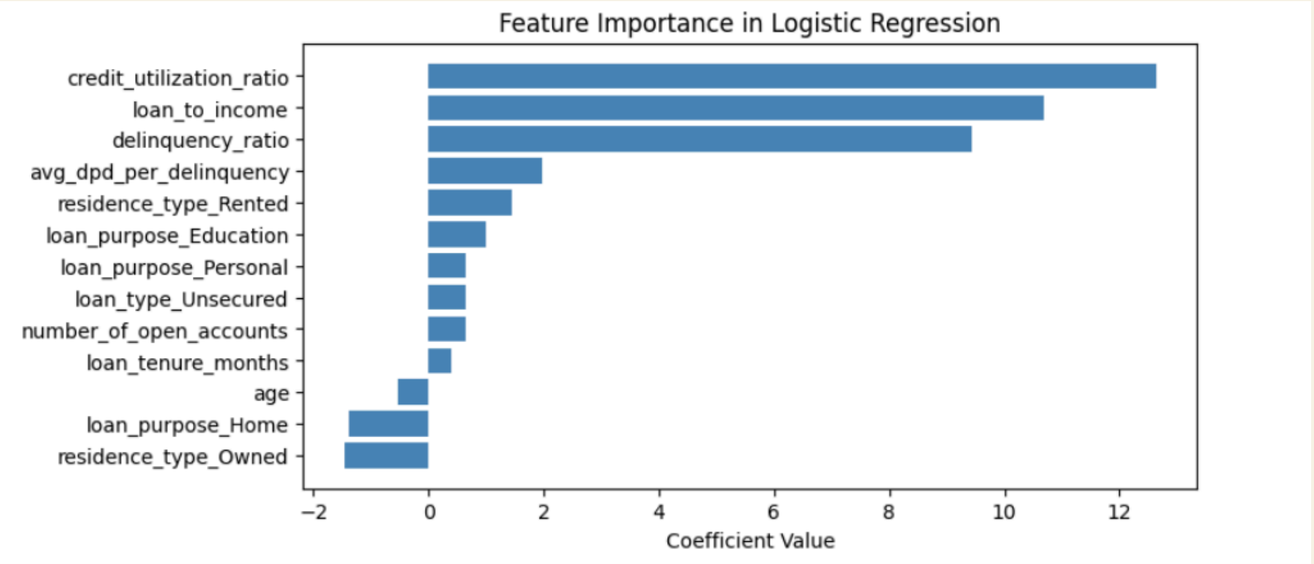
- All baseline models performed poorly on the default class, with low recall values, indicating failure to detect defaulters.

- To improve minority class detection, applied SMOTE Tomek:
  - SMOTE generated synthetic examples of defaulters.
  - Tomek Links removed overlapping/noisy majority class instances.
  - Significantly improved recall while retaining class distribution balance.

- Performed hyperparameter tuning using Optuna on Logistic Regression:
  - Tuned key parameters such as C (regularization strength) and solver.
  - Achieved better recall while preserving model interpretability and simplicity.
  - Optuna enabled efficient, automated optimization over a wide parameter space.

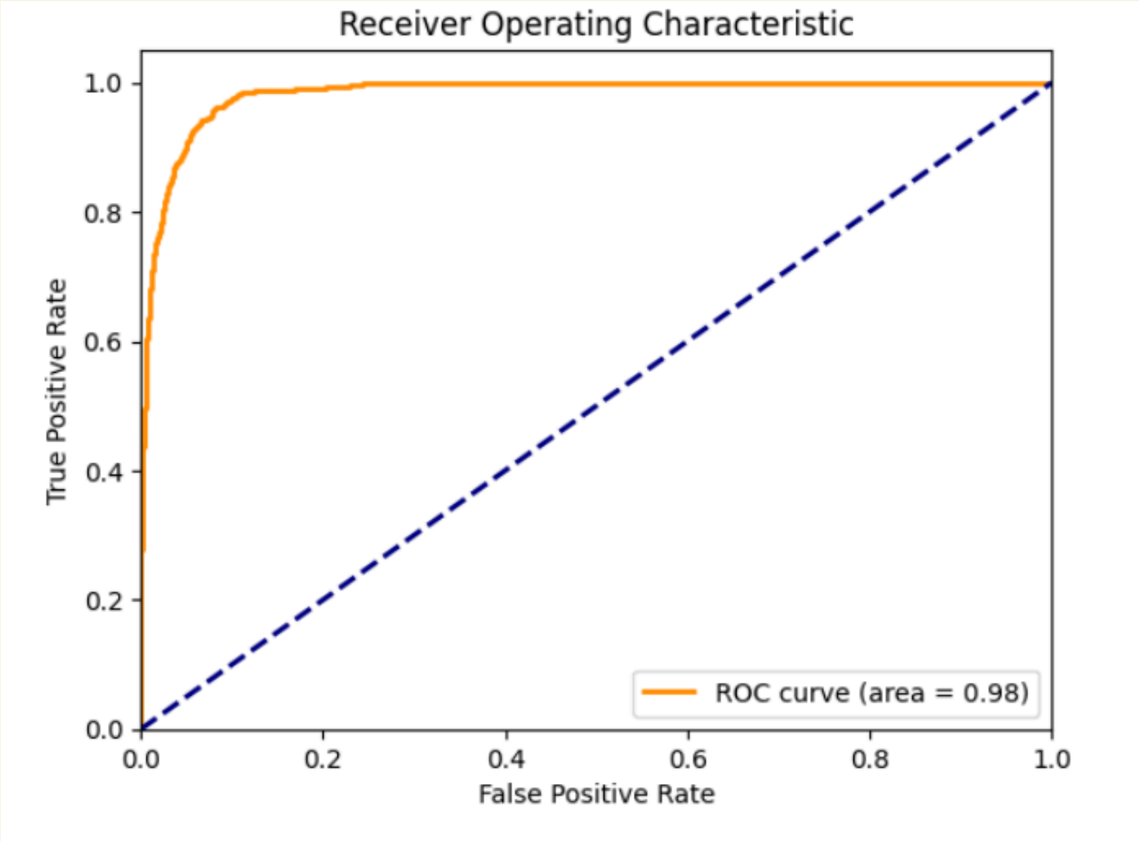
- Selected Logistic Regression with SMOTE Tomek and Optuna as the final model based on:
  - High recall on default class.
  - Strong balance between interpretability and performance.

```
Best trial:
  F1-score: 0.9459121201808944
  Params:
    C: 3.704319021882243
    solver: lbfgs
    tol: 0.00020521917884401469
    class_weight: balanced
              precision    recall  f1-score   support

           0       0.99      0.93      0.96     11423
           1       0.56      0.94      0.70      1074

    accuracy                           0.93     12497
   macro avg       0.78      0.94      0.83     12497
weighted avg       0.96      0.93      0.94     12497
```



MODEL EVALUATION

- ROC–AUC Curve
  - Evaluated the model's ability to separate default vs. non-default classes.
  - Achieved a high AUC score of 0.98, indicating excellent class discrimination.
  - ROC curve showed a strong balance between true positive and false positive rates.



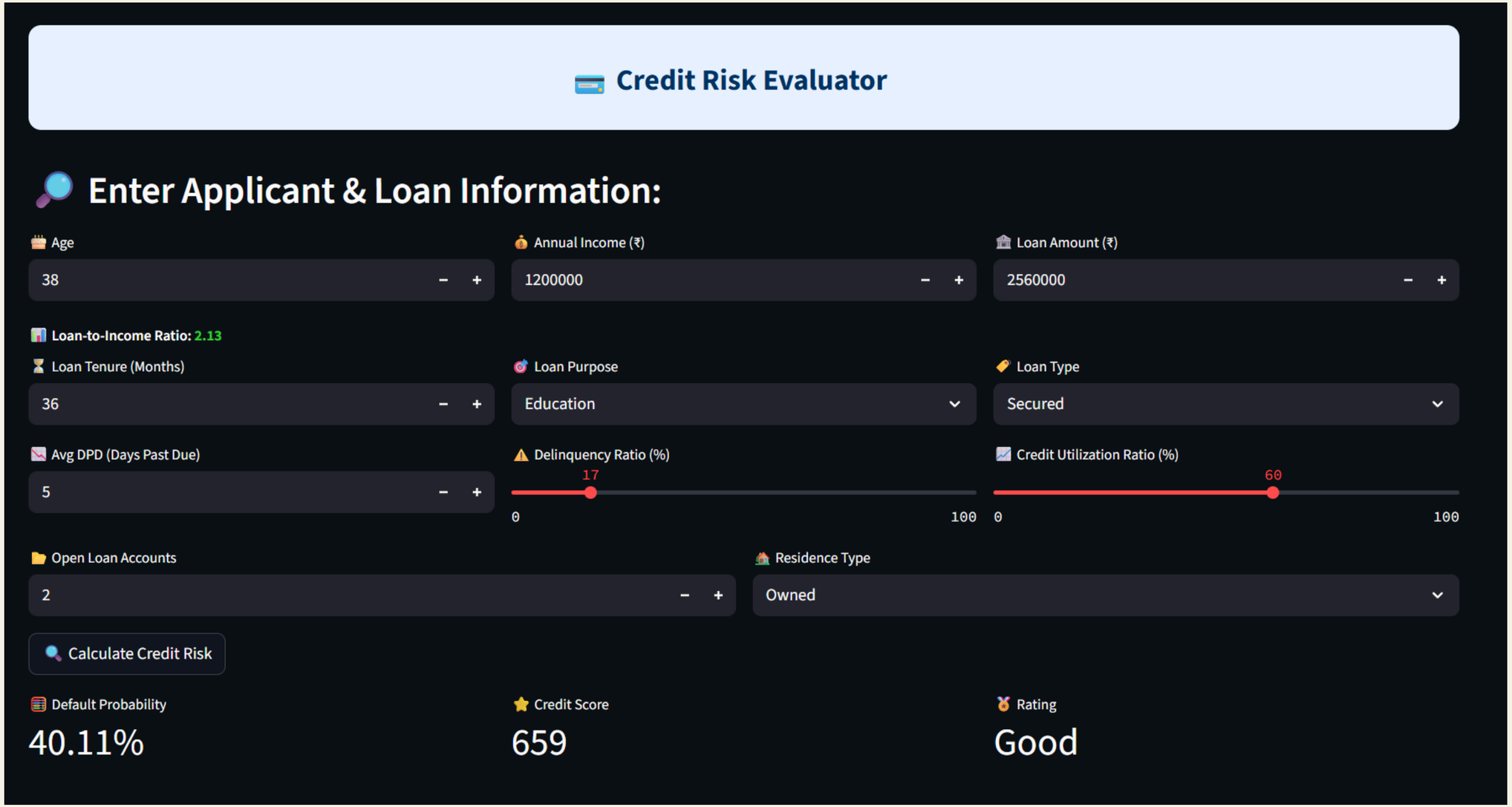
- KS (Kolmogorov–Smirnov) Statistic
  - Used to measure the maximum separation between the cumulative distributions of defaulters and non-defaulters.
  - Achieved a KS score above 86%, significantly exceeding the industry threshold of 40%.
  - Indicates that the model is highly effective at distinguishing risky applicants.
  - Most of the KS separation was observed within the first 3 deciles, aligning with business expectations for early-risk identification.
  - Validates the model's readiness for deployment in credit risk evaluation systems.

	Decile	Minimum Probability	Maximum Probability	Events	Non-events	Event Rate	Non-event Rate	Cum Events	Cum Non-events	Cum Event Rate	Cum Non-event Rate	KS
0	9	0.818	1.000	900.000	350.000	72.000	28.000	900.000	350.000	83.799	3.064	80.735
1	8	0.215	0.818	160.000	1090.000	12.800	87.200	1060.000	1440.000	98.696	12.606	86.090
2	7	0.029	0.214	9.000	1240.000	0.721	99.279	1069.000	2680.000	99.534	23.461	76.073
3	6	0.004	0.029	5.000	1245.000	0.400	99.600	1074.000	3925.000	100.000	34.361	65.639
4	5	0.001	0.004	0.000	1249.000	0.000	100.000	1074.000	5174.000	100.000	45.295	54.705
5	4	0.000	0.001	0.000	1250.000	0.000	100.000	1074.000	6424.000	100.000	56.237	43.763
6	3	0.000	0.000	0.000	1250.000	0.000	100.000	1074.000	7674.000	100.000	67.180	32.820
7	2	0.000	0.000	0.000	1249.000	0.000	100.000	1074.000	8923.000	100.000	78.114	21.886
8	1	0.000	0.000	0.000	1250.000	0.000	100.000	1074.000	10173.000	100.000	89.057	10.943
9	0	0.000	0.000	0.000	1250.000	0.000	100.000	1074.000	11423.000	100.000	100.000	0.000

STREAMLIT APP INTEGRATION

- Developed an interactive web application using Streamlit for real-time credit risk evaluation.
- Integrated the trained Logistic Regression model with preprocessing and feature engineering pipeline for seamless predictions.
- Enabled users to input details such as age, income, loan purpose, loan tenure, and delinquency-related information.
- On clicking "Calculate Credit Risk", the app provides:
  - Probability of default
  - Credit score (scaled 300–900)
  - Risk category (Poor / Average / Good / Excellent)
- All preprocessing steps are handled within the app to ensure consistent and accurate predictions.
- Deployed the app on Streamlit Cloud (streamlit.io) for public access.
- Designed for business and non-technical users, allowing quick decision-making through an intuitive interface.

USER INTERACTION PREVIEW



PROJECT SUMMARY

- Built a machine learning system to evaluate credit risk and generate a credit score (300–900) using applicant demographic, financial, and bureau data.
- Cleaned and preprocessed the dataset, handled outliers in processing fees, and fixed categorical inconsistencies.
- Engineered key features like Loan-to-Income Ratio, Delinquent-to-Tenure Ratio, and Average DPD per Delinquent Month.
- Performed feature selection using domain knowledge, VIF analysis, and Information Value (IV) filtering.
- Trained Logistic Regression, Random Forest, and XGBoost; improved recall using SMOTE Tomek and Optuna.
- Final model (Logistic Regression) achieved Recall > 90%, AUC > 98% and KS > 86%.
- Deployed the solution as an interactive Streamlit web app with real-time credit scoring and risk categorization.
- Live App: https://vaibhav-project-credit-risk-evaluator.streamlit.app/
- GitHub Repository: https://github.com/vaibhavgarg2004/Credit-Risk-Evaluator

THANK YOU