Machine Learning

Credit Risk Score Prediction

Shijin Ramesh | Data Scientist & ML Engineer | October 30, 2025

Objective & Success Criteria

Objective: Build a machine learning model that can be useful to risk analytics team of any bank to predict the credit risk score of a customer.

Success Criteria:

- AUC, Gini > 86
- KS Statistic > 40
- Max KS Statistic should be achieved in first three Deciles
- Model Interpretation

Train & Test Data

- Source: Past two year's loan data

Feb 2022 to May 2024

Features - Customers

Variables	Description
cust_id	Customer ID
age	Age of the customer
gender	Gender
marital_status	Marital Status
employment_status	Employment Status
income	Income of the Customer
number_of_dependants	Number of dependents
residence_type	Residence Type
years_at_current_address	Years at Present Address
city	City
state	State
zipcode	Zipcode/Pincode

Features - Loans

Variables	Description
loan_id	Load ID
cust_id	Customer ID
loan_purpose	Loan Purpose
loan_type	Loan Type
sanction_amount	Sanction Amount
loan_amount	Loan Amount
net_disbursement	Amount Disbursed in Customer's Account
loan_tenure_months	Loan Tenure in Months
principal_outstanding	POS (Principal Outstanding)
bank_balance_at_application	Bank Balance at Application
disbursal_date	Disbursed Date
installment_start_dt	Installment Start Date
default	Default/No default

Features - Bureau Data

Variables	Description
cust_id	Customer ID
no_of_open_accounts	Total Number of Open Accounts Till Date
no_of_closed_accounts	Total Number of Closed Accounts Till Date
total_loan_months	Total Loan in Months
delinquent_months	Total Delinquent in Months
total_dpd	Total Due Passed Day
enquiry_count	Total Enquiry Count
credit_utilization_ratio	Credit Utilization Ratio

Significant Variables

Variables	IV	Inference
credit_utilization_ratio	2.35	Higher usage of available credit significantly increases default risk
delinquency_ratio	0.72	Higher delinquency rates are strongly linked to increased default risk
loan_to_income	0.48	Higher loan amounts relative to income increase the likelihood of default
avg_dpd_per_delinquency	0.4	Higher days past due per delinquency correlates with higher default risk
loan_purpose	0.37	Certain loan purposes are more likely to be associated with default
residence_type	0.25	Residence type has a moderate impact on default risk
loan_tenure_months	0.22	Longer loan tenures increases default risk
loan_type	0.16	Different loan types have a minor influence on default risk
age	0.09	Younger or older age has a minimal effect on default risk
number_of_open_accounts	0.08	More open accounts can lead to default risk

Modeling Steps

Dataset

Customers
Loans
Bureau Data

Target: Default (Binary value)

Data Preprocessing

- Loan purpose invalid value replaced with mode
- Feature selection using IV, VIF & domain knowledge
- Min max scaling for numeric features

Train, Test, Split

- 75% - Training
- 25% - Test

Model Training

- Logistic Regression
- XGBoost
- Random Forest

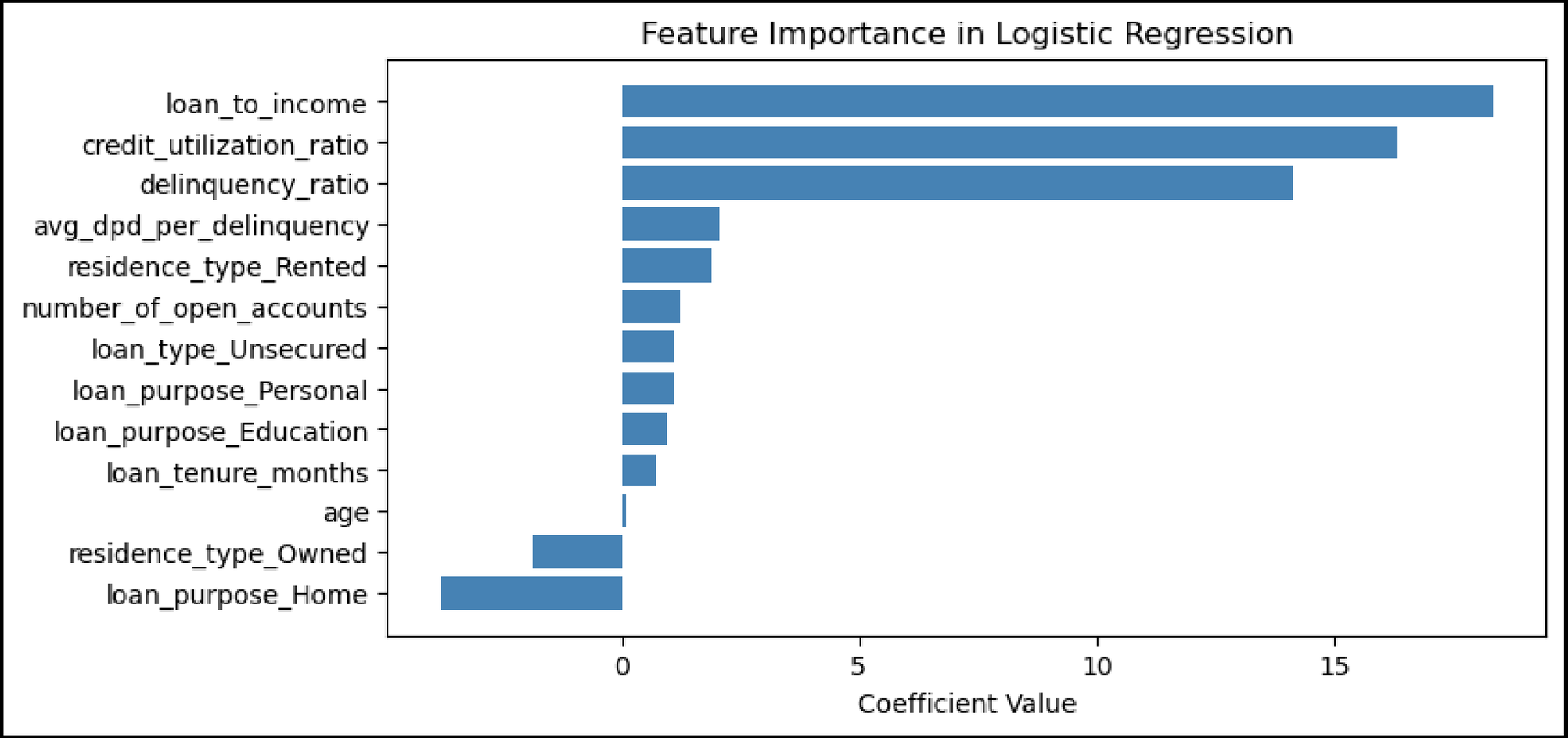
Fine Tuning

- RandomizedSearchCV
- Optuna

Model Evaluation

- AUC, KS Statistic, Gini Coefficient
- Classification Report

Feature Importance



Trials & Performance

Models	AUC	Gini Coefficient
Logistic Regression	0.98	0.96
XGBoost	0.98	0.97

Model Evaluation

	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	0.82	1.00	898.00	352.00	71.84	28.16	898.00	352.00	83.61	3.08	80.53
1	8	0.21	0.82	162.00	1088.00	12.96	87.04	1060.00	1440.00	98.70	12.61	86.09
2	7	0.03	0.21	9.00	1240.00	0.72	99.28	1069.00	2680.00	99.53	23.46	76.07
3	6	0.00	0.03	5.00	1245.00	0.40	99.60	1074.00	3925.00	100.00	34.36	65.64
4	5	0.00	0.00	0.00	1249.00	0.00	100.00	1074.00	5174.00	100.00	45.29	54.71
5	4	0.00	0.00	0.00	1250.00	0.00	100.00	1074.00	6424.00	100.00	56.24	43.76
6	3	0.00	0.00	0.00	1250.00	0.00	100.00	1074.00	7674.00	100.00	67.18	32.82
7	2	0.00	0.00	0.00	1249.00	0.00	100.00	1074.00	8923.00	100.00	78.11	21.89
8	1	0.00	0.00	0.00	1250.00	0.00	100.00	1074.00	10173.00	100.00	89.06	10.94
9	0	0.00	0.00	0.00	1250.00	0.00	100.00	1074.00	11423.00	100.00	100.00	0.00

AUC	Gini Coefficient	Top 3 Decile Capture Rate
0.98	0.96	99.53%

Appendix

- *Code Link: https://github.com/shijin/CustomerCreditRiskAnalysis*