Credit Score Calculator

Credit scoring plays a crucial role in the financial ecosystem, helping institutions assess the creditworthiness of individuals. In this project, we build a machine learning model to predict the likelihood of a loan default using customer demographic and financial information.

CrediSure

table of content

- Dataset Description
- Data Preprocessing
- Model Building
- Evaluation Metrics
- Results and Analysis
- Conclusion

Dataset Description

- The dataset used in this project is named credit_data.csv.
- It contains a variety of financial and demographic features related to individual customers.
- Key variables include:
  - Year of birth (used to calculate age)
  - Annual income and declared income
  - Number of children and dependents
  - Employment status
  - Outstanding loan amounts
  - Housing-related details such as home value and mortgage due

- The primary target variable is loan_status, which represents the creditworthiness of a borrower:
  - 0 → Indicates a "good" loan (the borrower is likely to repay)
  - 1 → Indicates a "bad" loan (the borrower has defaulted or is at high risk of default)
- The objective is to build a model that accurately predicts loan_status based on the provided features.

Data Preprocessing

- Handled missing and infinite values, and renamed columns for better clarity and consistency.
- Engineered new features like debt-to-income ratio, EMI-to-income ratio, and home equity to capture deeper financial insights.
- Treated outliers and normalized skewed distributions using transformations to improve data quality.
- Addressed class imbalance in the target variable (loan_status) using SMOTE to ensure fair model learning.
- Standardized feature values and split the dataset into stratified training and testing sets.

Data Cleaning & Formatting

Feature Engineering

Outlier & Distribution Handling

Class Imbalance Treatment

Data Scaling & Splitting

Model Building

- Logistic Regression
- Support Vector Classifier (SVC)
- Random Forest Classifier
- K-Nearest Neighbors
- Voting Classifier

Handling Class Imbalance

The dataset showed imbalance in target classes. To address this, SMOTE (Synthetic Minority Oversampling Technique) was used to oversample the minority class during training.

Pipeline

1. Train-test split
2. Feature scaling using StandardScaler
3. Model fitting and prediction

Evaluation Metrics

To assess the effectiveness of our classification models in predicting loan default risk, we used a combination of performance metrics, each capturing different aspects of model behavior:

Accuracy Score

Measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of predictions. While useful, it may be misleading in imbalanced datasets.

ROC AUC Score

(Receiver Operating Characteristic - Area Under Curve)

Evaluates the model's ability to distinguish between the two classes (good vs. bad loans). A higher AUC indicates better discrimination. This metric is especially important in credit scoring, where class imbalance is common.

Classification Report

Provides a detailed breakdown of:

- Precision: The percentage of positive predictions that were actually correct (important to avoid falsely labeling a good loan as bad).
- Recall (Sensitivity): The proportion of actual positives correctly identified (crucial for catching risky loans).
- F1-Score: The harmonic mean of precision and recall, balancing both metrics to give a single performance score.

Feature Importance & Explainability

To understand which inputs drive the model's decisions—and ensure our results align with financial intuition—we employed a combination of feature-analysis techniques

Tree-Based Importance

- For ensemble methods like Random Forest, we extracted built-in impurity-based importance scores.
- Highest scores were observed for Debt-to-Income Ratio (dti_ratio), Declared Income, Home Equity, and Outstanding Loans.

Model Coefficients (Logistic Regression)

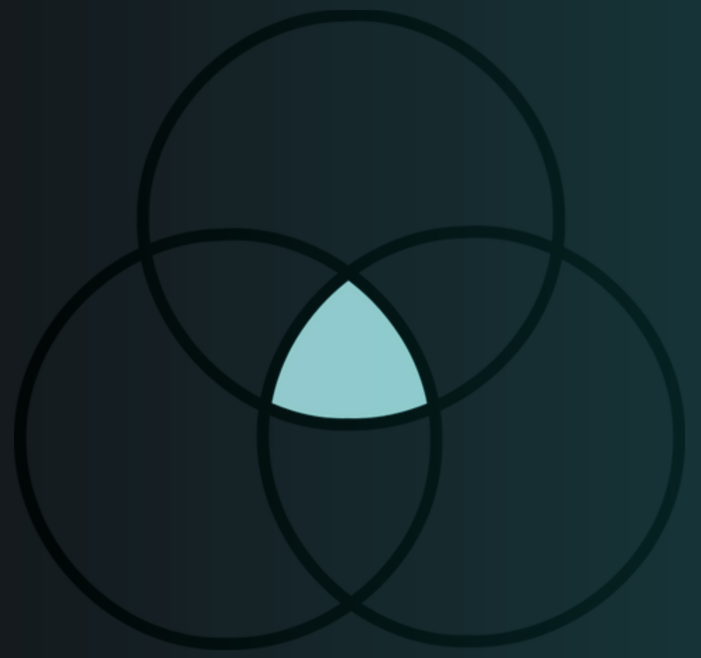
- We examined standardized regression coefficients to gauge the direction and strength of each feature's impact.
- Positive coefficients (e.g., higher dti_ratio) increased default risk, while negative coefficients (e.g., higher declared income) reduced it.

Permutation Importance

- By randomly shuffling each feature and measuring the resulting drop in performance, we obtained a model-agnostic ranking of feature relevance.
- Consistently, permuting dti_ratio or home_equity led to the largest degradation in AUC and accuracy.

Partial Dependence Plots (PDPs)

- PDPs illustrated how changes in a single feature—holding others constant—affect predicted default probability.
- For example, default risk climbed sharply once dti_ratio exceeded 0.4, and home equity below a certain threshold sharply increased risk.



These analyses collectively confirmed that:

- Debt-to-Income Ratio (dti_ratio) is the strongest predictor of default.
- Declared Income serves as a robust indicator of repayment capacity.
- Home Equity and Outstanding Loans capture borrowers' net worth and debt burden, respectively, making them critical risk factors.

User Persona

Applicant Overview

- Name: Vandana Tripathi
- Age: 34
- Employment: Full-Time Software Engineer
- Marital Status: Married, 1 child
- Phone Access: Yes

💰 Financial Profile

- Annual Income: ₹12,00,000
- Outstanding Loans: ₹50,000
- Monthly EMI: ₹4,000
- Home Value: ₹90,00,000
- Mortgage Balance: ₹15,00,000

🤖 ML Model Output

- Loan Status Prediction: No Default
- Probability of Default: 0.11
- Risk Level: 🟢 Low Risk
- Recommendation: Eligible for approval with minimal conditions.

📊 Risk Classification

- 🟢 Low Risk: Probability < 0.45
- 🟡 Moderate Risk: 0.45 – 0.55
- 🔴 High Risk: Probability > 0.55

Hurrah! 🎉 Vandana’s loan has been officially approved through CrediSure — the intelligent credit risk prediction system. With her strong financial profile and low risk score, she’s all set to chase her dreams, backed by trust, data, and CrediSure’s smart evaluation!

CrediSure

Conclusion

- Machine learning models were effectively used to assess credit risk based on financial and demographic data.
- Engineered features such as debt-to-income ratio, EMI-to-income ratio, home equity, and outstanding loans significantly improved the model's predictive ability.
- Rigorous data preprocessing—including handling of missing values, feature scaling, and class imbalance correction—ensured clean, consistent input for modeling.
- Ensemble methods like Random Forest and Gradient Boosting captured non-linear patterns and outperformed simpler models in both accuracy and interpretability.
- A customized risk stratification framework was introduced by mapping predicted probabilities to intuitive risk categories:
- 🟢 Low Risk (< 0.45): Likely to repay loan
- 🟡 Moderate Risk (0.45–0.55): Borderline, requires caution
- 🔴 High Risk (> 0.55): Elevated chance of default
- This probability-based framework makes model outputs more actionable, bridging the gap between data science insights and real-world loan decision-making.
- Overall, the project showcases how machine learning can make credit assessments more accurate, explainable, and aligned with domain logic.

CrediSure

-Created by Het Patel

Thank You

IIT Guwahati

☎ +91 9173320321

✉ p.vishnubhai@iitg.ac.in