Credit Risk Prediction - Technical Report

GenAI Capstone Project - NST Sonipat - February 2026

---

1. Problem Statement

Access to credit is a fundamental enabler of economic participation, yet lending institutions face the persistent challenge of distinguishing creditworthy borrowers from those likely to default. Inaccurate risk assessments lead to significant financial losses through non-performing assets or, conversely, the denial of credit to deserving applicants.

This project addresses the binary classification problem of predicting loan default risk. Given a set of applicant and loan attributes - such as income, employment history, loan grade, and prior default records - the goal is to build a supervised machine learning model that can predict whether a borrower will default on a loan (loan_status = 1) or repay successfully (loan_status = 0).

The trained model is deployed as an interactive Streamlit web application that enables real-time credit risk evaluation for new applicants.

2. Data Description

The dataset is a publicly available credit risk dataset hosted on GitHub, containing 32,581 records of historical loan applications.

2.2 Features

Feature	Type	Description
person_age	Numerical	Age of the applicant
person_income	Numerical	Annual income of the applicant
person_emp_length	Numerical	Employment length in years
person_home_ownership	Categorical	Home ownership status (RENT, OWN, ...)
loan_intent	Categorical	Purpose of loan (PERSONAL, EDUCATION, ...)
loan_grade	Categorical	Credit grade assigned (A-G)
loan_amnt	Numerical	Loan amount requested
loan_int_rate	Numerical	Interest rate on the loan
loan_percent_income	Numerical	Loan amount as % of annual income
cb_person_default	Binary	Previous default on record (Y/N)
cb_person_cred_hist	Numerical	Length of credit history
loan_status	Binary	Target (0 = Non-Default, 1 = Default)

Class Distribution: Non-Default (0): ~78% | Default (1): ~22%

3. Exploratory Data Analysis (EDA)

- Target Distribution: Confirmed class imbalance (addressed via balanced weighting).

- Income vs. Default: Defaulters tend to have lower incomes.
- Loan Amount vs. Default: Higher loan amounts associate with increased risk.
- Loan-to-Income Ratio: Strongest separator; high ratios indicate higher risk.
- Interest Rate vs. Default: Riskier borrowers are assigned higher rates.
- Correlation: Loan percent income has the strongest positive correlation with default.

4. Methodology

4.1 Preprocessing: Median imputation for missing values in employment length and interest rate; Ordinal encoding for grades; One-hot encoding for categorical features; StandardScaler normalization.

4.2 Train-Test Split: 75% Training, 25% Testing (random_state=42).

4.3 Model: Decision Tree Classifier (max_depth=10, min_samples_split=20, min_samples_leaf=10, class_weight='balanced').

Rationale: Decision Tree was selected over Logistic Regression for its higher overall accuracy (90.8%) and superior recall for the default class (0.77).

5. Evaluation

5.1 Model Comparison

Metric	Logistic Regression	Decision Tree
Accuracy	84.52%	90.81%
ROC-AUC	0.8636	0.6817

5.2 Classification Report (Decision Tree)

Class	Precision	Recall	F1-Score	Support
Non-Default (0)	0.93	0.95	0.94	6331
Default (1)	0.81	0.77	0.79	1815
Weighted Avg	0.91	0.91	0.91	8146

5.3 Confusion Matrix

Actual / Predicted	Non-Default	Default
Non-Default	6,006 (TN)	325 (FP)
Default	424 (FN)	1,391 (TP)

6. Optimization & Limitations

- Class Imbalance: used balanced weighting.
- Pruning: max_depth=10 to prevent overfitting.
- Limitations: Binary default history misses partial defaults; limited probability calibration (ROC-AUC 0.68).

7. System Architecture

Data Flow: [Borrower Data/CSV] -> [Data Preprocessing] -> [Scaling & Encoding] -> [ML Model: Decision Tree] -> [Risk Score / Prediction] -> [Interactive UI: Streamlit]

8. Input-Output Specification

Input Specification:

Format: CSV or Manual Form Entry. Key features: Annual Income, Age, Loan Amount, and Historical Default records.

Output Specification:

Result: Probability Score (0.0 to 1.0) and Classification (Approved / Needs Review / Rejected). Includes Default Probability % and repayment likelihood.

9. Key Risk Drivers

- Loan Percent Income: Most significant indicator; higher ratios correlate with higher risk.
- Loan Grade: Strong monotonic trend from Grade A (lowest risk) to G (highest risk).
- Interest Rate: Higher rates are major predictive drivers of default likelihood.
- Income: Lower income levels show a statistically higher frequency of default.

10. Deployment

Streamlit App: https://genaicapstone-a7eipdbqudn2niewt9s2mp.streamlit.app/

11. Team Contribution

Team Member	Contribution
Palak	Data preprocessing, EDA, model training & evaluation
Samarth	App development, UI/UX design, deployment, bug fixes