

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 the denial of credit to deserving applicants.

This project addresses the binary classification problem of predicting loan default risk. Given applicant attributes - income, employment history, loan grade, prior default records - the goal is to predict whether a borrower will default (status=1) or repay (status=0).

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

2. Data Description

The dataset contains 32,581 records of historical loan applications.

Features

| Feature | Type | Description |
|-----------------------|-------------|------------------------------------|
| person_age | Numerical | Age of the applicant |
| person_income | Numerical | Annual income |
| person_emp_length | Numerical | Employment length (years) |
| person_home_ownership | Categorical | RENT, OWN, MORTGAGE, OTHER |
| loan_intent | Categorical | PERSONAL, EDUCATION, MEDICAL, etc. |
| loan_grade | Ordinal | Credit grade (A-G) |
| loan_amnt | Numerical | Loan amount requested |
| loan_int_rate | Numerical | Interest rate (%) |
| loan_percent_income | Numerical | Loan amount as % of income |
| cb_person_default | Binary | Previous default on record (Y/N) |
| cb_person_cred_hist | Numerical | Credit history length (years) |
| loan_status (Target) | Binary | 0 = Non-Default, 1 = Default |

Class Distribution: Non-Default ~78% | Default ~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-to-Income Ratio: Strongest class separator; high ratios = higher risk.
- Interest Rate: Defaulters concentrated in higher rate brackets.
- Loan Grade: Default rates increase progressively from A to G.
- Correlation: loan_percent_income has strongest positive correlation with default.

4. Methodology

4.1 Preprocessing

1. Missing values: Median imputation (emp_length, int_rate)

2. Ordinal encoding: Loan grade A-G mapped to 1-7
3. Binary encoding: Default history Y/N to 1/0
4. One-hot encoding: Home ownership, loan intent (drop_first)
5. Feature scaling: StandardScaler normalization

4.2 Train-Test Split

75% training / 25% testing (random_state=42)

4.3 Models Trained

Model 1: Logistic Regression (baseline, max_iter=1000)

Model 2: Decision Tree Classifier (selected) - max_depth=10, min_samples_split=20, min_samples_leaf=10, class_weight='balanced'

Rationale: Decision Tree selected for higher accuracy (90.8% vs 84.5%) and better recall on the default class (0.77), which is critical in credit risk assessment.

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 class_weight='balanced' to upweight minority class.
- Tree Pruning: max_depth=10 prevents overfitting while maintaining generalization.
- Limitation: Binary default history; cannot capture partial default rates.
- Limitation: ROC-AUC of 0.68 indicates room for probability calibration improvement.

7. Deployment

Live App: <https://genaicapstone-a7eipdbqudn2niewt9s2mp.streamlit.app/>

- Real-time prediction via interactive form.
- Three-tier decision: Approved (<20%), Needs Review (20-50%), Rejected (>50%).
- Displays default probability, repayment likelihood, and loan grade.

8. Team Contribution

| Member | Contribution |
|---------|--|
| Palak | Data preprocessing, EDA, model training & evaluation |
| Samarth | Streamlit app, UI/UX design, deployment, bug fixes |

9. Tech Stack

| Component | Technology |
|-----------------|-----------------------------|
| Language | Python 3.x |
| ML Libraries | scikit-learn, pandas, NumPy |
| Visualization | Matplotlib, Seaborn |
| Web Framework | Streamlit |
| Deployment | Streamlit Cloud |
| Version Control | Git & GitHub |

10. System Architecture

The system is structured into three functional layers:

- Layer 1 - User Interface (Streamlit): Handles applicant forms and input validation.
- Layer 2 - Intelligence (Python/Scikit-Learn): Manages preprocessing, scaling, and Decision Tree model logic.
- Layer 3 - Output Display: Visualizes risk metrics, tier status, and probability scores.

The Data Journey:

1. Input: User provides details (Income, Age, etc.) via the web form.
2. Transformation: Data is cleaned and scaled for the model.
3. Prediction: Decision Tree calculates default probability.
4. Action: Result is categorized and displayed on the dashboard.

11. Input-Output Specification

System Inputs (Borrower Profile)

| Category | Key Features | Purpose |
|--------------|---------------------|---------------------------------|
| Demographics | Age, Emp Length | Assess stability and life-stage |
| Financials | Income, Loan Amount | Calculate Loan-to-Income ratio |
| History | Default/Credit Hist | Factor in past behavior |
| Context | Home/Loan Purpose | Collateral type and context |

System Outputs (Credit Decision)

| Result Type | Output Detail | Description |
|-------------|----------------|-----------------------------------|
| Score | Default Prob % | Statistical risk of non-repayment |
| Decision | Status Tier | Approved / Review / Rejected |
| Metrics | Repayment % | Confidence in repayment |

12. Key Risk Drivers (Ranked by Impact)

The following features are the primary drivers of credit risk, ranked by impact:

- 1. Loan-to-Income Ratio: Single biggest predictor; high debt is a primary rejection trigger.
- 2. Interest Rate: Strong signal of pre-existing risk profiles assigned by lenders.

- 3. Loan Grade: Summary quality metric; Grades E-G carry heavy penalties.
- 4. Annual Income: Baseline affordability check; low income keeps risk high.

13. References

1. scikit-learn Documentation - <https://scikit-learn.org/stable/>
2. Streamlit Documentation - <https://docs.streamlit.io/>
3. Credit Risk Dataset - Hosted on GitHub repository
4. Decision Tree Classifier - Breiman, L. (1984). Classification and Regression Trees