**CREDIT RISK ANALYSIS REPORT**

abdulbasit adefugabi

## Date: 03/24/2025

**FINAL CREDIT RISK ANALYSIS REPORT**

**1. Introduction**

**1.1 Project Overview**

Credit risk assessment is a crucial aspect of the banking and financial industry, as it helps institutions minimize losses due to loan defaults. This project aims to develop a robust machine learning model capable of predicting the likelihood of loan default using a dataset containing financial, personal, and credit history attributes of applicants.

**1.2 Business Objective**

The primary objectives of this study include:

* Identifying key factors that influence loan default rates.
* Developing a reliable credit risk prediction model using machine learning techniques.
* Providing actionable insights to financial institutions to optimize their lending strategies and minimize risks.

**2. Data Overview**

**2.1 Dataset Description**

The dataset comprises **45,000 loan applications** with **14 key features**, including demographic details, loan information, and credit history. The main features are:

* **Demographic Information:**
  + person\_age: Age of the applicant.
  + person\_gender: Gender (Male/Female).
  + person\_education: Educational background.
  + person\_income: Annual income of the applicant.
  + person\_emp\_exp: Years of employment experience.
  + person\_home\_ownership: Type of home ownership (Rent/Own/Mortgage).
* **Loan-Specific Details:**
  + loan\_amnt: Requested loan amount (USD).
  + loan\_intent: Purpose of the loan (Personal, Education, Business, etc.).
  + loan\_int\_rate: Interest rate applied to the loan.
  + loan\_percent\_income: Loan amount as a percentage of income.
* **Credit History:**
  + cb\_person\_cred\_hist\_length: Length of applicant’s credit history (years).
  + credit\_score: Numeric representation of creditworthiness.
  + previous\_loan\_defaults\_on\_file: Indicator of past loan defaults.
  + loan\_status: **Target variable** (1 = Default, 0 = Non-Default).

**2.2 Data Cleaning and Preprocessing**

A structured data preprocessing approach was implemented:

* **Handling Missing Values:**
  + Records with missing values were dropped to maintain dataset integrity.
* **Encoding Categorical Variables:**
  + Categorical variables (person\_gender, person\_education, etc.) were label-encoded.
* **Feature Scaling:**
  + Continuous numerical features were standardized using StandardScaler to optimize model performance.

**3. Exploratory Data Analysis (EDA)**

**3.1 Key Insights**

* **Income Distribution:**
  + The income variable exhibits a **right-skewed distribution**, indicating a wide variance in earnings.
* **Credit Score vs Default Rate:**
  + **Higher credit scores correspond to lower default rates**, confirming its strong predictive power.
* **Loan Amount and Default Rate:**
  + Applicants with **larger loan amounts** have a higher probability of defaulting.
* **Loan Purpose and Default Risk:**
  + **Personal loans and business loans** have the highest risk of default.
* **Correlation Analysis:**
  + **Credit Score and Credit History Length** negatively correlate with loan defaults.

**4. Machine Learning Model Development**

**4.1 Model Selection & Training**

Multiple machine learning models were explored:

1. **Random Forest Classifier:**
   * Baseline model using ensemble learning with 100 decision trees.
   * Moderate accuracy but lacked interpretability.
2. **XGBoost Classifier:**
   * Fine-tuned model with 200 estimators, learning rate = 0.05, max depth = 6.
   * Outperformed Random Forest in both accuracy and AUC-ROC.

**4.2 Model Performance Evaluation**

**Random Forest Classifier:**

* **Accuracy:** 85%
* **AUC-ROC Score:** 0.78
* **Classification Report:** Balanced precision, recall, and F1-score.

**XGBoost Classifier:**

* **Accuracy:** 89%
* **AUC-ROC Score:** 0.84
* **Confusion Matrix Insights:** Lower false positives, better classification of high-risk applicants.

**4.3 Feature Importance Analysis**

The top five predictors of loan default were:

1. **Credit Score** - The most influential factor in predicting default risk.
2. **Loan Amount** - Larger loan sizes increased default probability.
3. **Income-to-Loan Ratio** - Lower ratios corresponded with higher default risks.
4. **Credit History Length** - Longer histories were associated with lower risks.
5. **Previous Defaults** - Applicants with past defaults had a higher probability of defaulting again.

**5. Business Recommendations**

Based on the findings, the following strategic recommendations are proposed:

* **Stricter Credit Score Requirements:**
  + Implement a **higher minimum credit score threshold** for loan approvals.
* **Loan-to-Income Ratio Limits:**
  + Avoid approving loans where requested amounts exceed **a safe percentage of annual income**.
* **Enhanced Risk Monitoring:**
  + Increase scrutiny on applicants with **short credit histories and prior defaults**.
* **Differentiated Interest Rates:**
  + Implement **risk-adjusted interest rates** for high-risk borrowers to mitigate financial exposure.

**6. Conclusion**

This study successfully developed a **credit risk prediction model** that enables banks and financial institutions to:

* Minimize **loan default risks** through predictive analytics.
* Optimize **loan approval strategies** based on reliable indicators.
* Enhance financial stability by **mitigating potential losses**.

This project highlights the potential of machine learning in **financial risk assessment**, providing a scalable and data-driven approach to managing loan default risks effectively.