Loan Default Prediction Report

Course:AI

Submitted by: Malk Emad Abdallah

Baraa magdy ahmed

Mariam mohamed mahmoud

Talal mohamed saad

Instructor: Eng_Abdullah wagih

Date: 9 JUL 2025

1. DEFINE PROBLEM (NON-TECHNICAL)

The objective is to predict whether a loan applicant is likely to default, using historical data. This assists financial institutions in making informed decisions and reducing the risk of bad loans.

2. DATASET DESCRIPTION

The dataset used in this analysis consists of historical loan data with features that describe both the applicant and the loan. Typical features include:

- **Personal Information**

person_age: Age of the applicant (in years).
person_gender: Gender of the applicant (male, female).
person_education: Educational background (High School, Bachelor, Master, etc.) person_income:Annual income of the applicant (in USD).
person_emp_exp: Years of employment experience.
person_home_ownership: Type of home ownership (RENT, OWN, MORTGAGE).

- **Loan Details**

loan_amnt: Loan amount requested (in USD).
loan_intent: Purpose of the loan (PERSONAL, EDUCATION, MEDICAL, etc.).
loan_int_rate: Interest rate on the loan (percentage).
loan_percent_income: Ratio of loan amount to income.

- **Credit & Loan History**

cb_person_cred_hist_length: Length of the applicant's credit history (in years).
credit_score: Credit score of the applicant.

previous_loan_defaults_on_file: Whether the applicant has previous loan defaults (Yes or No).Target Variable

- **Outcome Variable:** A binary target column indicating whether the loan was **defaulted (1)** or **not defaulted (0)**.

3. DATA PREPROCESSING (PLANNED OR ONGOING)

- **Missing Values:** None found — all columns are complete.
- **Encoding:** Categorical columns such as `person_gender`, `person_education`, etc., will need encoding.
- **Scaling:** Required for numeric columns (`person_income`, `loan_amnt`, `loan_int_rate`, etc.)
- **Balancing:** To be checked — class imbalance in `loan_status` may require SMOTE or similar.
- **Splitting:** Data will be divided into training and testing sets.

4. DATA ANALYSIS

Exploratory Data Analysis (EDA) helps understand relationships and distributions.

- **Univariate Analysis** – Histograms of individual features.
- **Bivariate Analysis** – Relationships with target variable.
- **Multivariate Analysis** – Correlation matrix, pair plots.

5. DATA PREPROCESSING

Steps included:

- **Encoding categorical variables** using label encoding.
- **Handling missing values** (none found).
- **Outlier handling** for variables like `loan_amnt` and `loan_int_rate`.

- **Train-Test Split:** 70% training, 30% test set.
- **Feature Scaling:** Used MinMaxScaler for algorithms sensitive to scale.

5. CLASSIFICATION MODELS USED

1.logistic regression

- Served as a **baseline model**.
- Simple and interpretable.

2. RANDOM FOREST CLASSIFIER

- Handled non-linearity and feature interactions well.
- Outperformed logistic regression.

3. XGBOOST CLASSIFIER

- **Best performing model** in terms of precision and recall.
- Used hyperparameter tuning via `RandomizedSearchCv`.

6. MODEL EVALUATION

METRICS USED

- **Accuracy**: Overall correctness.
- **Precision**: How many predicted defaulters were actual defaulters.
- **Recall**: How many actual defaulters were correctly identifi
- Confusion Matrix

6.EXPLORATORY DATA ANALYSIS (INITIAL OBSERVATIONS)

- Many applicants have **low experience** (`0-3 years`) and **modest incomes**, which may influence risk.
- **Default rate is higher for lower income and younger applicants.**
- **Most loans are for DEBTCONSOLIDATION, MEDICAL, or PERSONAL** needs.
- Some applicants have **previous defaults** recorded, which should correlate with higher risk.
- **No major missing values** were found, making the dataset relatively clean.

7. EXPECTED OUTCOME

The final model will help:

- **Identify high-risk applicants** before issuing loans
- Improve **approval workflows** and **reduce defaults**
- Support banks in developing data-driven credit policies

8. CONCLUSION

- The model can help banks identify high-risk applicants, minimize loan default rates, and optimize their lending strategies.
- **Credit score, income, and loan amount** are among the top predictors.
- **XGBoost is the most reliable model** for predicting loan defaults in this dataset.
- With proper preprocessing and tuning, machine learning can significantly enhance credit risk assessment.

---

9. FUTURE WORK

- Integration of real-time data (transactional behavior, account activity).
- Exploration of deep learning models for further performance gains.