Loan Default Prediction System – Major Project Report

Student Name: *Virendra Mahajan*
Course: Data Science
Project Category: Major Project

---

1. Introduction

Banks and financial institutions face high financial risk when customers fail to repay loans. Predicting loan default accurately helps lenders make informed decisions, reduce losses, and improve risk management.

This project focuses on building a **Loan Default Prediction System** using machine learning. The system analyzes applicant information such as income, credit score, loan amount, employment history, and more to predict if a customer is likely to default.

This end-to-end project includes:
Data preprocessing
Exploratory data analysis (EDA)
Model building (Logistic Regression, Random Forest)
Model comparison
Feature importance analysis
Evaluation (Accuracy, F1, ROC-AUC, Confusion Matrix)
A working **Streamlit web app** for real-time prediction

---

2. Objective

- The main objectives of this project are:
- To analyze loan applicant data and identify key factors that influence default.
- To build a machine learning model that predicts whether a loan applicant is likely to default.
- To compare multiple machine-learning models and select the best-performing one.
- To develop a user-friendly **Streamlit app** that allows real-time loan default prediction.
- To produce insights and visualizations that help stakeholders make informed decisions.

---

3. Methodology

3.1 Dataset Used

Dataset: loan_data_sample.csv
The dataset contains essential features such as:

age

income

loan_amount

term

credit_score

employment_years

home_ownership

purpose

default

3.2 Tools and Technologies

Component	Technology Used
Programming	Python
Data Handling	Pandas, NumPy
ML Models	Logistic Regression, Random Forest
Visualization	Matplotlib, Seaborn
Model Persistence	Joblib
Web App	Streamlit
Environment	Virtual Environment (venv)

---

4. Code & Implementation Details

4.1 Preprocessing

File: src/preprocess.py
Steps performed:

Removed irrelevant columns (loan_id)

Filled missing values

Standardized numerical features

Encoded categorical variables using OneHotEncoder

Saved clean data:
data/processed/loan_data_processed.csv

4.2 Model Training

File: src/train_model.py

Two models were trained:

Logistic Regression

Random Forest Classifier

Evaluation done using:

Accuracy

F1-Score

ROC-AUC

Output:
Best model saved as src/models/model.joblib

4.3 Model Comparison

File: src/compare_models.py

Generated comparison table:

Model	Accuracy	F1 Score	ROC-AUC
Logistic Regression	**0.75**	**0.7826**	**0.7348**
Random Forest	0.525	0.5777	0.6174

Logistic Regression is the best-performing model.

Saved as:
reports/model_comparison.csv

4.4 Feature Importance

File: src/feature_importance.py

Identified the top factors influencing loan default.

Saved files:
reports/feature_importance.csv
src/figures/feature_importance.png

4.5 Evaluation

File: src/evaluate.py

Results on 200-sample test dataset:

Accuracy: 91%

F1 Score: 0.91

ROC-AUC: 0.985

Confusion Matrix Saved:
src/figures/confusion_matrix.png

---

5. Results & Observations

Key Insights:

Logistic Regression outperformed Random Forest on this dataset.
The most important features were:

Credit Score

Annual Income

Loan Amount

Employment Years
Higher credit score = lower default risk
High loan amount + low income = higher default probability

Model Performance Summary

The model shows strong performance with high accuracy and ROC-AUC, making it suitable for real-world loan risk assessment.

---

6. Streamlit Web App

File: src/app.py

The app allows users to input:

Age

Income

Loan Amount

Term

Credit Score

Employment Years

Home Ownership

Purpose

It then predicts:
Default or **Not Default**

---

7. Conclusion

The Loan Default Prediction System successfully predicts loan default using machine-learning techniques. Logistic Regression proved to be the most reliable model for this dataset.

Key learnings:

Importance of data preprocessing

Handling categorical + numerical features

Evaluating multiple models

Deploying ML models in real-time using Streamlit

This project aligns with Data Science major project objectives by covering the complete ML lifecycle: Data → Model → Evaluation → Deployment.

---

8. Links, Datasets & Files Used

Resource	Location
Dataset	data/loan_data_sample.csv
Processed Dataset	data/processed/loan_data_processed.csv
Model	src/models/model.joblib
Confusion Matrix	src/figures/confusion_matrix.png
Feature Importance Graph	src/figures/feature_importance.png
Code Files	src/*.py
Comparison Results	reports/model_comparison.csv
Feature Importance CSV	reports/feature_importance.csv
