CREDIT RISK PREDICTION PROJECT REPORT

- INDHU PRIYADHARSHINI S

1. Problem Statement

Financial institutions frequently encounter challenges in accurately evaluating the creditworthiness of loan applicants. Inaccurate assessments can lead to significant financial losses and increased risk exposure. This project aims to develop a machine learning-based solution that predicts whether an applicant is likely to be a good or bad credit risk based on their profile attributes, thereby assisting financial institutions in making informed and strategic lending decisions.

2. Abstract

This project presents a web-based Credit Risk Prediction system built using machine learning algorithms and deployed through Streamlit for interactive use. Leveraging the German Credit Dataset, we performed comprehensive data preprocessing, exploratory data analysis (EDA), model training, and evaluation across various classification models. The final deliverable is a user-friendly web application that allows users to input applicant details and obtain real-time credit risk predictions.

3. Dataset Description

We utilized the German Credit Dataset, comprising 1000 samples, with the following key features:

- **Age** (numeric)

- **Sex** (categorical: male, female)

- **Job** (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)

- **Housing** (categorical: own, rent, free)

- **Saving accounts** (categorical: little, moderate, quite rich, rich)

- **Checking account** (numeric, balance in Deutsch Marks)
- **Credit amount** (numeric, in Deutsch Marks)
- **Duration** (numeric, loan term in months)
- **Purpose** (categorical: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

The target variable is the applicant's **credit risk** (Good or Bad).

4. Methodology

Data Preprocessing

- Imputed missing values in Saving accounts and Checking account features by assigning appropriate categories ('unknown').
- Performed label encoding to convert categorical features into numerical format.
- Applied StandardScaler to normalize numerical attributes, ensuring consistency across features.

Exploratory Data Analysis (EDA)

- Conducted detailed visualizations and analysis to understand feature distributions and relationships with the target variable.
- Key insights identified:
- Applicants aged below 25 exhibited a slightly higher tendency toward bad credit risk.

- Higher credit amounts combined with longer loan durations correlated strongly with bad credit outcomes.

- A 'little' balance in checking accounts was a common trait among bad credit applicants.

Model Training and Evaluation

We evaluated multiple classification algorithms, including:

- **Random Forest Classifier**

- **XGBoost Classifier**

- **Linear Regression** (although Linear Regression is generally more suited for continuous output, it was adapted here for comparative purposes)

Performance was assessed using:

- **Accuracy**

- **ROC AUC Score**

- **Precision**, **Recall**, and **F1 Score**

Best Performing Model

The **XGBoost Classifier** emerged as the best-performing model, achieving:

- **Accuracy**: **76%**

- **ROC AUC Score**: **0.775**

Its robustness and superior generalization ability made it the ideal choice for deployment.

5. Streamlit Web Application

A fully interactive web application was developed using **Streamlit**, providing a seamless experience across three major sections:

- **Home**: Introduces the project and displays key insights and visualizations.
- **EDA**: Presents data exploration findings through intuitive plots and graphs.
- **Predict**: Allows users to input applicant information and obtain real-time credit risk predictions.

Key Features:

- Simple and clean user interface for ease of use.
- User input forms capturing all essential applicant details.
- On-the-fly data preprocessing and scaling using the pre-fitted scaler.
- Instantaneous display of the prediction result and probability, supporting confident decision-making.

6. Conclusion

This project successfully demonstrates the application of machine learning in the domain of credit risk assessment. Among the evaluated models, the **XGBoost Classifier** exhibited superior predictive performance and was chosen for deployment. The resulting **Streamlit web application** is intuitive, efficient, and offers real-time risk predictions, making it a valuable asset for financial institutions aiming to enhance their loan approval processes and mitigate credit risks.