PROJECT REPORT

CREDIT RISK PREDICTION

Submitted by,

Nitheesh Srinivaasan R

Table of Contents

S.no	Section Title	Page no.
1	Problem Statement	1
2	Abstract	1
3	Dataset Description	1
4	Methodology	1
5	Streamlit Web Application	2
6	Key Realationship Found	3
7	Conclusion	

1. Problem Statement

Financial institutions often face challenges in accurately assessing the creditworthiness of loan applicants. Misjudging credit risk can lead to significant financial losses. The aim of this project is to develop a machine learning-based application that predicts whether a given loan applicant is likely to be a **good** or **bad** credit risk based on their profile.

2. Abstract

In this project, we built a web-based Credit Risk Prediction system using machine learning techniques and Streamlit for UI deployment. We utilized the **German Credit Dataset**, performed data preprocessing, exploratory data analysis (EDA), model training, and evaluation using multiple classification algorithms. The final output is an interactive web application that enables users to predict credit risk by inputting applicant details.

3. Dataset Description

The German Credit Dataset includes 1000 samples with the following features:

- **Age** (numeric)
- **Sex** (categorical)
- **Job** (0 to 3)
- **Housing** (own, rent, free)
- **Saving accounts** (little, moderate, quite rich, rich)
- **Checking account** (little, moderate, quite rich, rich)
- **Credit amount** (numeric)
- **Duration** (numeric)
- **Purpose** (car, education, business, etc.)

Target variable: **Risk** (Good/Bad credit risk)

4. Methodology

1. **Data Preprocessing**:

    - Filled missing values in 'Saving accounts' and 'Checking account' with 'unknown' or appropriate category.
    - Encoded categorical variables using label encoding.
    - Scaled numerical features using `StandardScaler`.

2. **Exploratory Data Analysis (EDA)**:

    - Visualized distribution and relationships between features and target.
    - Identified trends such as:

        - Younger individuals (<25) tend to have slightly higher bad credit risk.
        - High credit amount and long durations are associated with bad credit risk.
        - 'little' checking account balance is a common trait in bad credit cases.

3. **Model Training and Evaluation**:

    - Tested multiple algorithms:

        - **Random Forest**
        - **XGBoost Classifier**
        - **K Nearest Neighbours**
        - **Decision Tree Classifier**
        - **Support Vector Classifier (SVC)**
    - Performance Metrics:
        - Accuracy
        - ROC AUC Score
        - Precision, Recall, and F1 Score
4. **Best Performing Model**:
    - **XGBoost Classifier**
        - **Accuracy**: 0.76
        - **ROC AUC Score**: 0.77

5. Streamlit Web Application

- Built using `streamlit` for a user-friendly interface.
- App has three tabs:

    - **Home**: Overview and visual plots
    - **EDA**: Displays plots related to data analysis
    - **Predict**: Allows user input and performs risk prediction using trained XGBoost model

Features:

- Input fields for all applicant details.
- Data is preprocessed and scaled using saved scaler.
- Model prediction and probability of good credit shown in the output.

6. Key Relationships Found

- **Credit Amount vs Duration**: Higher credit amounts over long durations correlate with bad credit risk.
- **Age vs Credit Risk**: Young and very old age groups show higher risk.
- **Account Types**: Individuals with 'little' in both checking and saving accounts are more prone to be risky.
- **Purpose**: Certain purposes like 'business' or 'repairs' are more often linked with bad credit outcomes.

7. Conclusion

This project successfully demonstrates the application of machine learning to predict credit risk. The XGBoost classifier outperformed other models and was chosen for deployment. The Streamlit app is intuitive and provides quick, accurate predictions to assist decision-makers in finance.