Credit Risk Prediction System – Project Summary

Problem Statement

The objective of this project is to predict the likelihood of loan default using borrower and loan-related features. Accurate credit risk prediction helps financial institutions minimize losses and make data-driven loan approval decisions.

Dataset Explanation

The dataset contains borrower demographics, financial attributes, and loan characteristics. Key features include personal attributes (age, income, employment length), loan attributes (interest rate, loan-to-income ratio), and one-hot encoded categorical variables such as loan grade, loan intent, and home ownership status. The target variable is loan_status, indicating default or non-default.

Preprocessing Steps

Categorical variables were converted into numerical form using one-hot encoding. Feature and label alignment was ensured using index-based selection. The dataset was split into training and testing sets in a 70:30 ratio. Final feature ordering was preserved using a separate columns.json file.

Visualizations

Exploratory data analysis included bar plots for feature importance, ROC curves, confusion matrices, and distribution plots for key numerical features such as income and interest rate.

Algorithms Used

Several machine learning models were explored including Logistic Regression, K-Nearest Neighbors, Decision Trees, and XGBoost. XGBoost was selected as the final model due to its superior performance on tabular data.

Performance Comparison

Models were evaluated using accuracy, classification reports, and AUROC. XGBoost achieved the highest AUROC score, indicating better discrimination between defaulters and non-defaulters.

Deployment Details

The trained XGBoost model was serialized using pickle and deployed as a Streamlit web application. The app supports CSV uploads and manual feature input. Deployment was carried out on Streamlit Cloud with dependencies managed via a requirements.txt file.

Learning Outcomes and Challenges Faced

The project provided hands-on experience with end-to-end machine learning pipelines, feature engineering, model evaluation, and cloud deployment. Challenges included handling feature mismatches, pickle errors, environment dependency issues, and deployment debugging.