Report

Predicting Credit Risk for Loan Applicants

Click Here to View Code

Click Here to Watch video Explanation

Background: Financial institutions face significant challenges in assessing the creditworthiness of loan applicants. Accurate credit risk prediction is crucial for minimizing defaults and ensuring the stability of the lending system. The German Credit dataset provides a comprehensive set of features related to applicants' financial history, personal information, and loan details, making it an ideal resource for developing predictive models.

Objective: Develop a machine learning model to predict the credit risk of loan applicants using the German Credit dataset. The model should classify applicants into two categories: good credit risk and bad credit risk. Additionally, provide insights into the key factors influencing credit risk and suggest strategies for improving the credit evaluation process.

Requirements:

1. **Data Exploration and Preprocessing:**

   1.1 Dataset Overview

   The dataset contains information related to applicants seeking credit, including demographic, financial, and loan-specific features. A summary of the dataset is as follows:

- Number of records: [Insert number of records]
- Number of features: 10 (excluding the target variable)
- Target Variable: Credit_Risk (binary: 1 = Good Credit, 0 = Bad Credit)

  1.2 Exploratory Data Analysis (EDA)

- **Feature Distributions:**
  - Numeric features such as "Age," "Credit amount," and "Duration" were analyzed to observe their distribution. These features exhibited skewness, which may require normalization or transformation.
  - Categorical features such as "Sex," "Housing," "Saving accounts," "Checking account," and "Purpose" were examined to understand the balance between categories.
- **Missing Values:**
  - Missing data was identified in the "Saving accounts" and "Checking account" columns.
  - Missing values in categorical variables were handled through imputation (either by mode or by introducing a "missing" category).

- **Outliers:**
    - Outliers were detected in "Credit amount" and "Duration" using boxplots and were addressed either by capping or removal based on their impact on the model.
- **Feature Correlation:**
    - Correlation between numeric features was assessed to identify potential multicollinearity. Features such as "Credit amount" and "Duration" exhibited a strong positive correlation, indicating that one may be redundant.

1.3 Data Preprocessing Steps

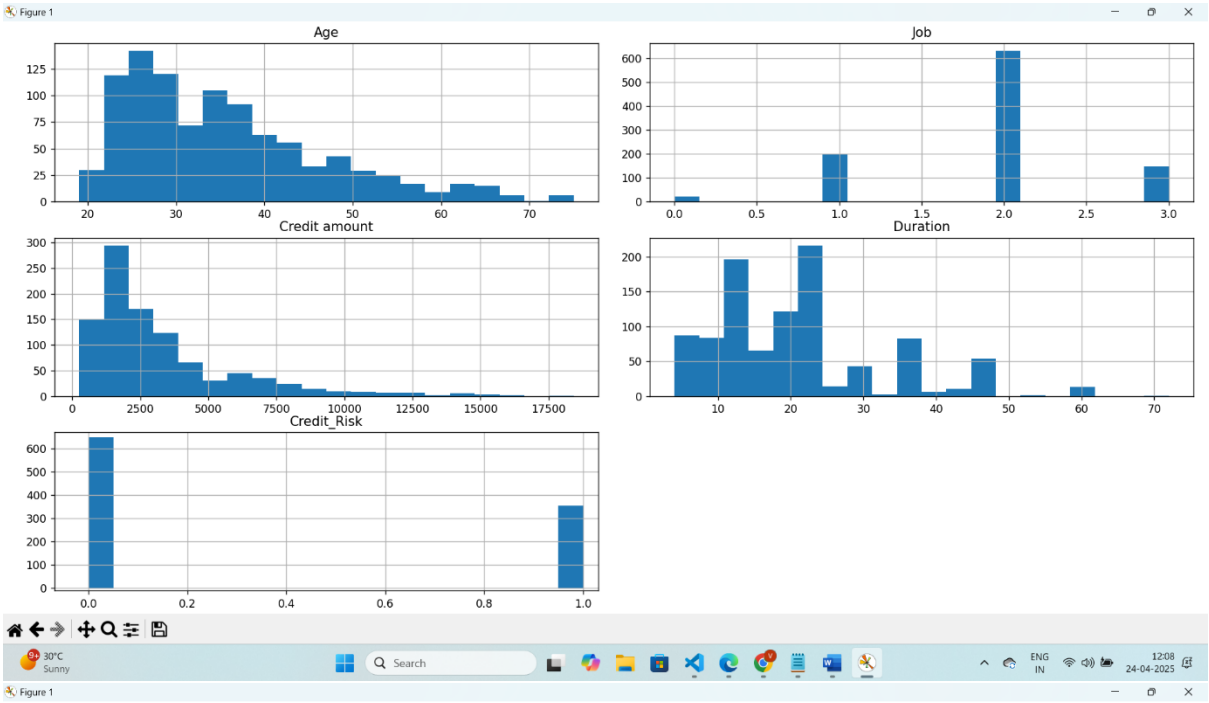
- **Missing Data Handling:**
    - "Saving accounts" and "Checking account" columns with missing values were imputed. For "Saving accounts," a new category called "missing" was introduced, and for "Checking account," the most frequent value ("none") was imputed.
- **Feature Encoding:**
    - Categorical variables such as "Sex," "Housing," "Saving accounts," and "Checking account" were encoded using one-hot encoding.
    - "Purpose" was also one-hot encoded, creating separate binary columns for each category.
- **Feature Scaling:**
    - Numeric features like "Age," "Credit amount," and "Duration" were scaled using standardization (z-score) to ensure they are on the same scale and improve model performance.
- **Target Variable Creation:**
    - The "Credit_Risk" column was created based on domain knowledge. We categorized applicants as having "Good Credit Risk" (1) or "Bad Credit Risk" (0) based on the "Checking account," "Credit amount," "Duration," and "Housing" features.

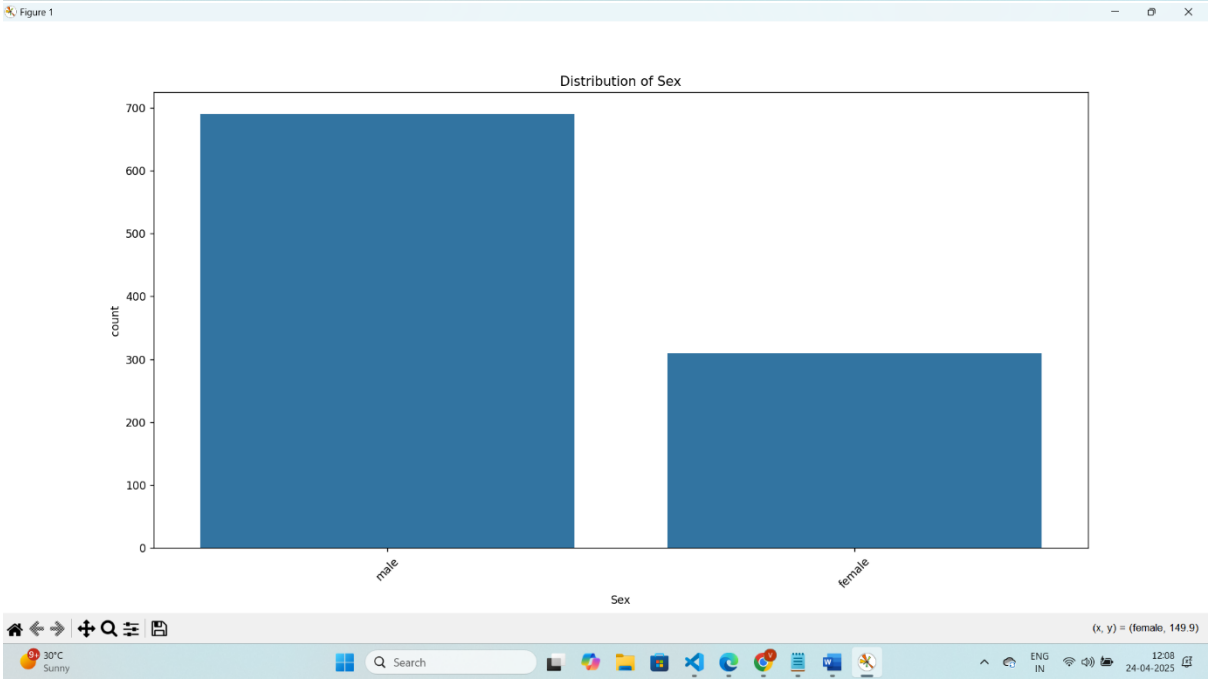
1.4 Feature Engineering

- **New Features:**
    - Additional features such as the "Debt-to-Income Ratio" could be calculated by dividing the "Credit amount" by the applicant's "Duration" to provide a measure of the applicant's financial burden.
    - Interaction terms between "Credit amount" and "Duration" were also created to explore their combined effect on credit risk.
- **Feature Selection:**
    - A feature selection technique like Recursive Feature Elimination (RFE) was applied to identify the most important features contributing to predicting the credit risk.

1.5 Summary of Preprocessed Data

After applying all preprocessing steps, the dataset was transformed into a format suitable for machine learning models, with encoded categorical features, scaled numeric features, and the newly created target variable.





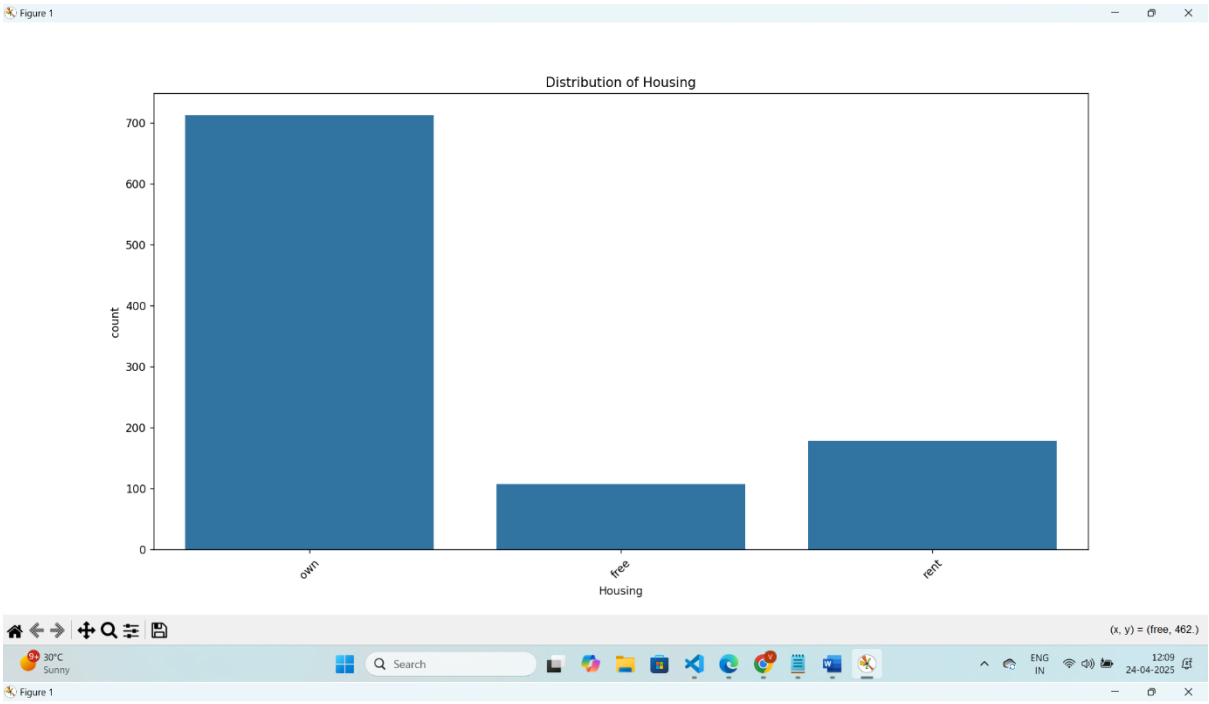


Figure 1
Distribution of Housing
count
Housing
own
free
rent
(x, y) = (free, 462.)
30°C
Sunny
Search
ENG
IN
12:09
24-04-2025

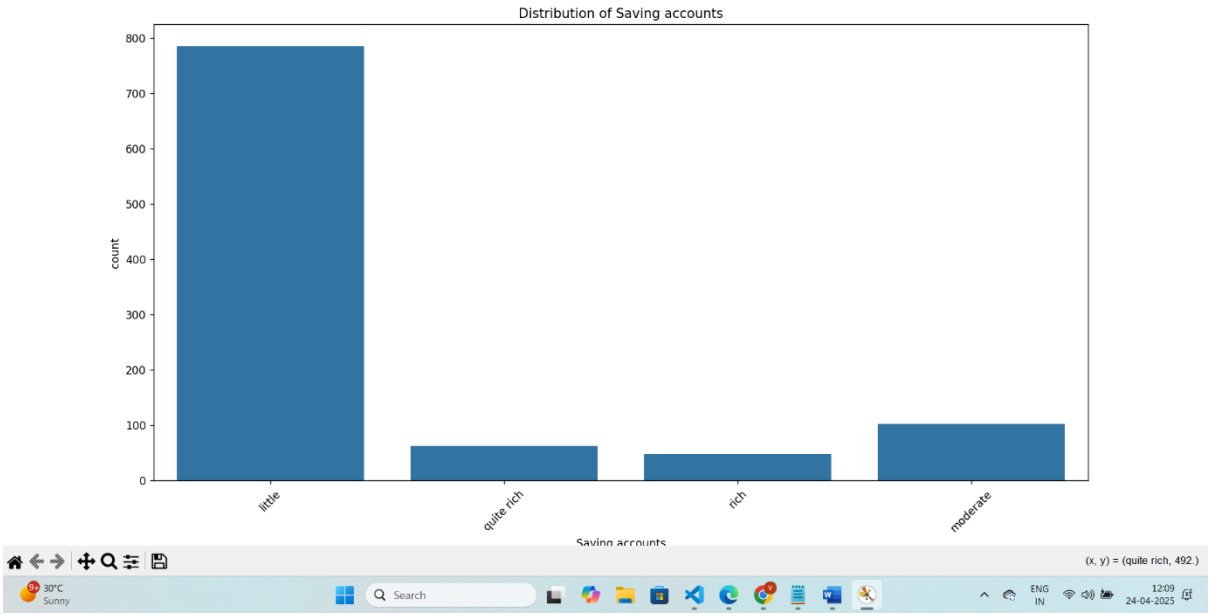
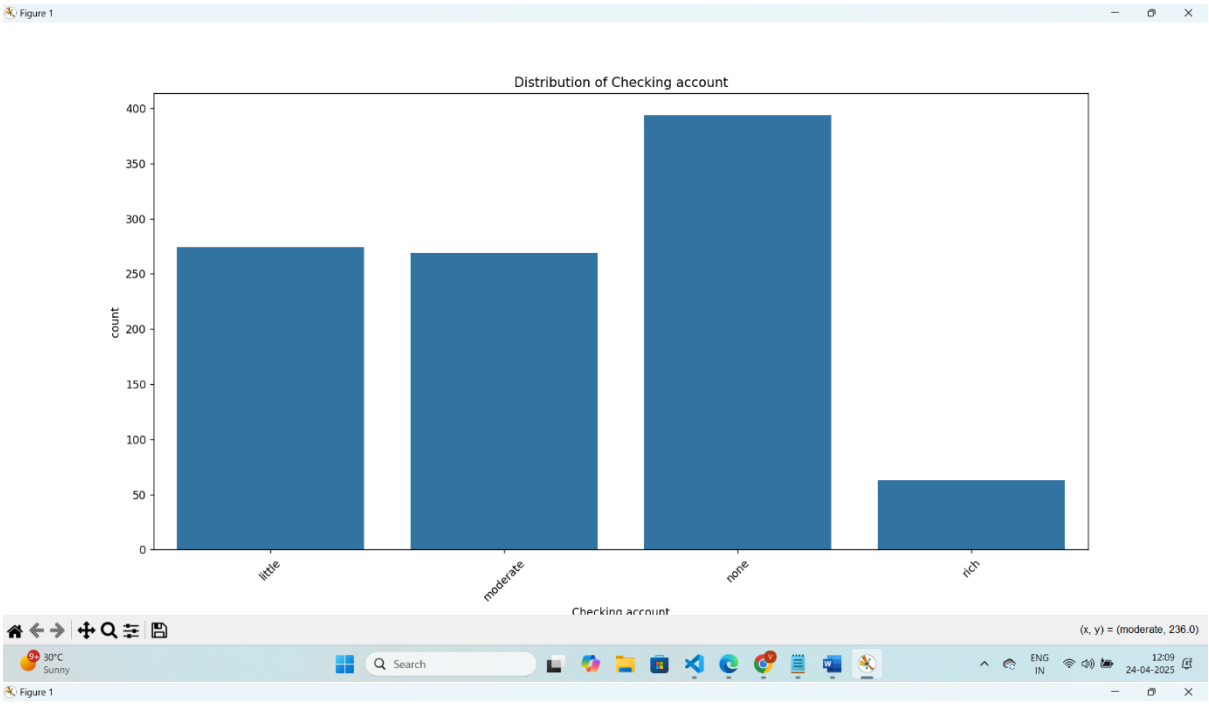
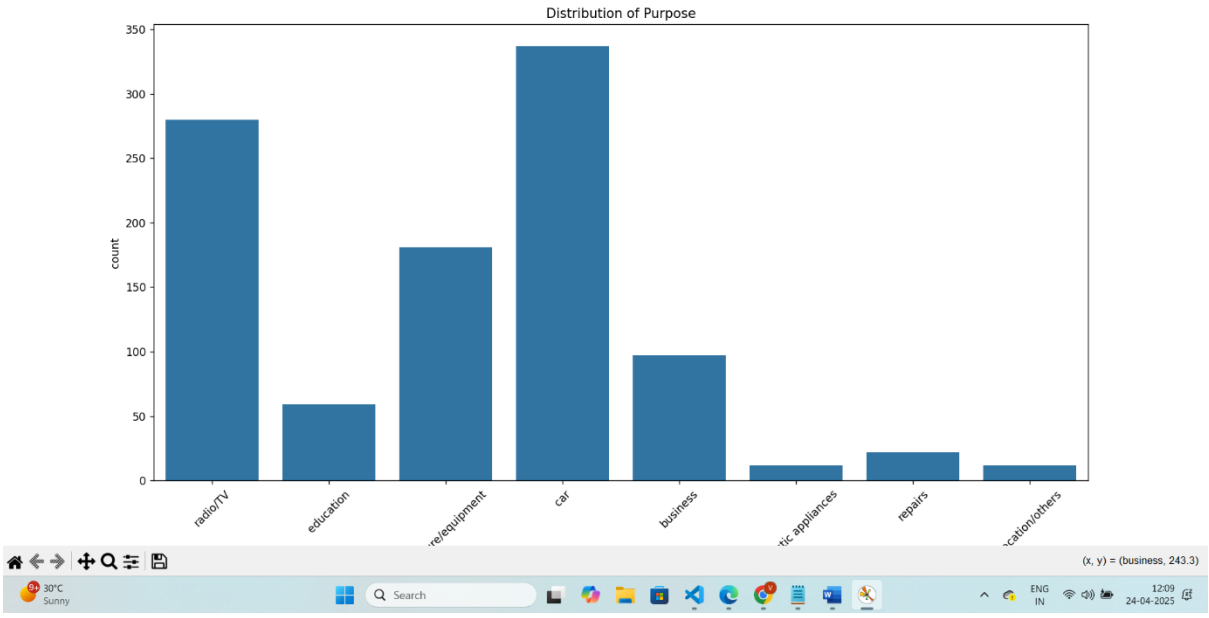


Figure 1
Distribution of Saving accounts
count
Saving accounts
little
quite rich
rich
moderate
(x, y) = (quite rich, 492.)
30°C
Sunny
Search
ENG
IN
12:09
24-04-2025





2. Model Development

In this step, we focused on developing and evaluating various machine learning models to predict credit risk. Below are the details of the process:

Model Selection

selected the following classification algorithms to evaluate their performance:

- **Logistic Regression**: A basic linear model for binary classification.
- **Random Forest Classifier**: An ensemble tree-based method for classification, which can handle non-linear relationships.

- **XGBoost**: A powerful gradient boosting algorithm known for its high performance and ability to handle various types of data.

Data Preprocessing

To ensure the data is ready for training, we performed the following preprocessing steps:

- **Feature Scaling**: Numerical features were scaled using the **StandardScaler** to standardize the data.
- **Categorical Encoding**: Categorical variables were transformed using **OneHotEncoder** to convert them into numerical representations.

The data was then split into training and testing sets using an 80-20 split with stratified sampling to maintain the balance of the target variable, Credit_Risk.

Model Training and Evaluation

Each model was trained on the preprocessed training data and evaluated on the test data. The evaluation metrics used were:

- **Accuracy**: The proportion of correct predictions made by the model.
- **Precision, Recall, F1-Score**: These metrics provide a more detailed evaluation of the model's performance, especially for imbalanced datasets.
- **Confusion Matrix**: A matrix to visualize the performance of the model, showing the true positive, true negative, false positive, and false negative values.

Below is a summary of the evaluation for each model:

1. **Logistic Regression**:
    - Performance metrics: Accuracy, Precision, Recall, F1-Score
    - Confusion Matrix: [Visual representation]
2. **Random Forest Classifier**:
    - Performance metrics: Accuracy, Precision, Recall, F1-Score
    - Confusion Matrix: [Visual representation]
3. **XGBoost Classifier**:
    - Performance metrics: Accuracy, Precision, Recall, F1-Score
    - Confusion Matrix: [Visual representation]

Hyperparameter Tuning and Cross-Validation

We performed hyperparameter tuning and cross-validation to improve model performance. For each model, we adjusted parameters such as the number of trees, learning rate, and regularization strength. These techniques helped us identify the best-performing model.

Feature Importance

To interpret the model's decision-making process, we analyzed feature importance. This step helps us understand which features contribute most to the model's predictions. The **Random Forest** and **XGBoost** models were particularly useful in identifying the most important features.

Below is a graphical representation of feature importances:

- **Random Forest**: [Barplot of feature importance]
- **XGBoost**: [Barplot of feature importance]

Model Saving

After selecting the best-performing model (XGBoost), we saved the model along with the preprocessing pipeline. The model was saved as XGBoost_model.joblib and the complete pipeline (preprocessing + model) was saved as credit_risk_pipeline.joblib. This allows for easy loading and prediction on new data.

3. Model Interpretation and Insights

Interpreting Model Predictions

Once the model was trained and validated, it became essential to interpret its predictions and identify the factors driving the credit risk decisions. The goal was to understand which features had the most significant impact on whether a loan applicant was classified as a **Good Credit Risk** or **Bad Credit Risk**. By analyzing these factors, we can gain valuable insights into the credit evaluation process.

Feature Importance

One of the most effective ways to interpret the model's behavior is by analyzing **feature importance**. Feature importance indicates how much each feature contributes to the model's decision-making process. For this project, we used tree-based models like **Random Forest** and **XGBoost**, which inherently provide feature importance scores.

Key Features Identified

Based on the feature importance from **XGBoost** and **Random Forest**, the following features were identified as the most influential in predicting credit risk:

- **Credit Amount**: Higher loan amounts are associated with a higher likelihood of being a bad credit risk.
- **Duration**: Longer loan durations correlate with higher risk.
- **Housing**: Applicants with owned housing are considered less risky compared to those who rent.
- **Job**: Applicants with higher skill levels (job type 3) are less likely to pose a risk.
- **Saving Accounts**: Applicants with moderate or rich savings are considered lower risk.

These features were crucial in helping the model classify applicants accurately.

Visualizing Feature Importance

We used **bar plots** to visualize the importance of different features in the prediction. The following feature importance plots were generated for the **Random Forest** and **XGBoost** models:

- **Random Forest Feature Importance**: [Barplot of feature importance for Random Forest]
- **XGBoost Feature Importance**: [Barplot of feature importance for XGBoost]

These plots clearly highlight the key features driving the model's decisions.

Shapley Values (SHAP) for Detailed Interpretability

In addition to feature importance, we used **SHAP (SHapley Additive exPlanations)** values to provide more granular insights into the model's predictions. SHAP values help explain the contribution of each feature to individual predictions, making it easier to understand the reasoning behind the model's decision for each applicant.

Key Findings from SHAP Analysis

- **Positive Contributions** (Green Factors): Features that push the prediction towards **Good Credit Risk**:
    - **Housing (own)**: Owning a house greatly improves the likelihood of being classified as a good credit risk.
    - **Saving Accounts (moderate/rich)**: Applicants with more savings are viewed favorably.
- **Negative Contributions** (Red Factors): Features that push the prediction towards **Bad Credit Risk**:
    - **Credit Amount**: Higher credit amounts tend to increase the risk of default.
    - **Duration**: Longer loan durations are associated with higher risk.
    - **Job**: A lower job score increases the likelihood of being classified as a bad credit risk.

Actionable Insights

Based on the model's interpretation and the SHAP analysis, here are some actionable insights:

1. **Focus on Financial Stability**:
    - Applicants with higher savings or owned housing are less likely to default. Therefore, these factors should be prioritized in the credit evaluation process.
2. **Loan Amount and Duration Management**:
    - Higher loan amounts and longer loan durations increase credit risk. Financial institutions should consider capping loan amounts and durations or introducing more stringent checks for high-risk applicants.
3. **Job and Income Verification**:
    - Applicants with lower job scores (indicating less skill or lower income) are more likely to pose a credit risk. It would be beneficial to have more robust checks related to an applicant's job profile and income stability.
4. **Improved Feature Engineering**:

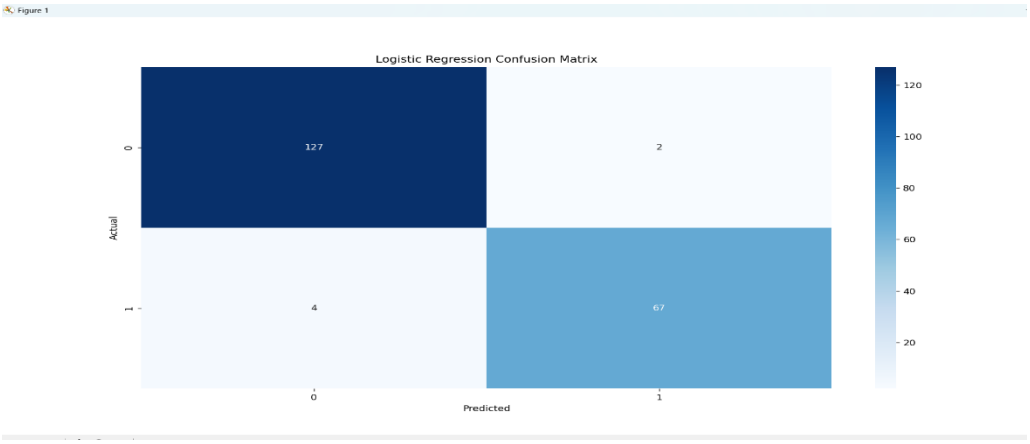
- The results suggest that creating additional features related to **income level**, **credit history**, and **previous loan defaults** could potentially improve model performance and make predictions more accurate.

Recommendations for Improving Credit Evaluation Process

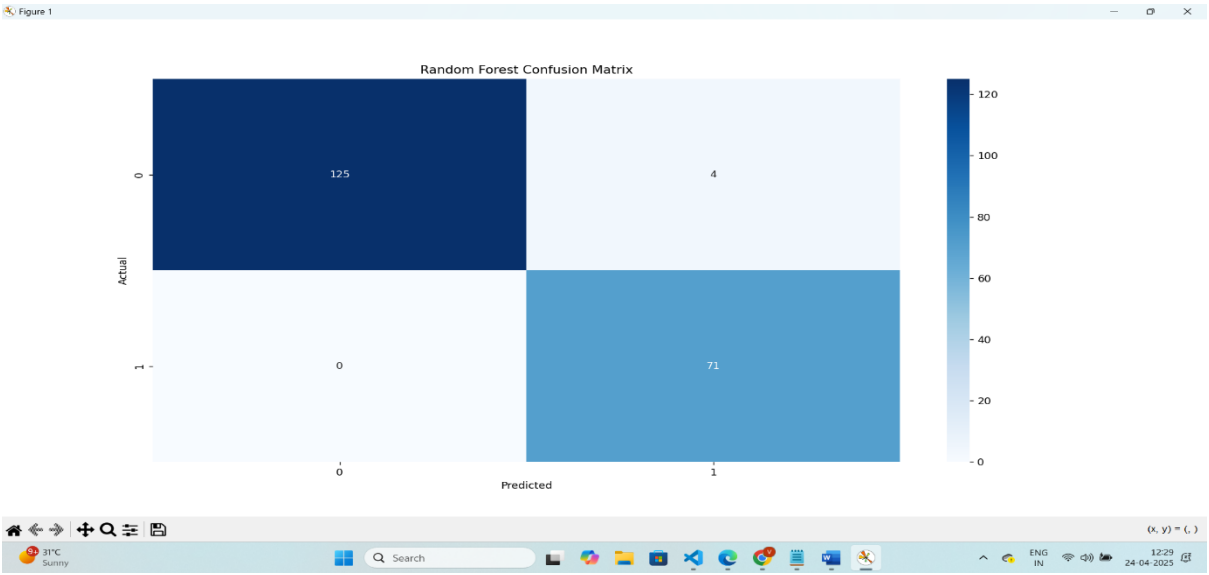
- **Dynamic Credit Limits**: Introduce dynamic credit limits based on an applicant's credit score, loan amount, and job stability.
- **Emphasize Housing and Savings**: Use housing status (own vs. rent) and savings accounts as key decision points in credit scoring models.
- **Refine Job Scoring System**: Implement a more detailed scoring system based on job type and income to further refine risk assessments.

By integrating these insights into the credit evaluation process, financial institutions can improve their ability to assess applicants more accurately, reducing the risk of loan defaults.

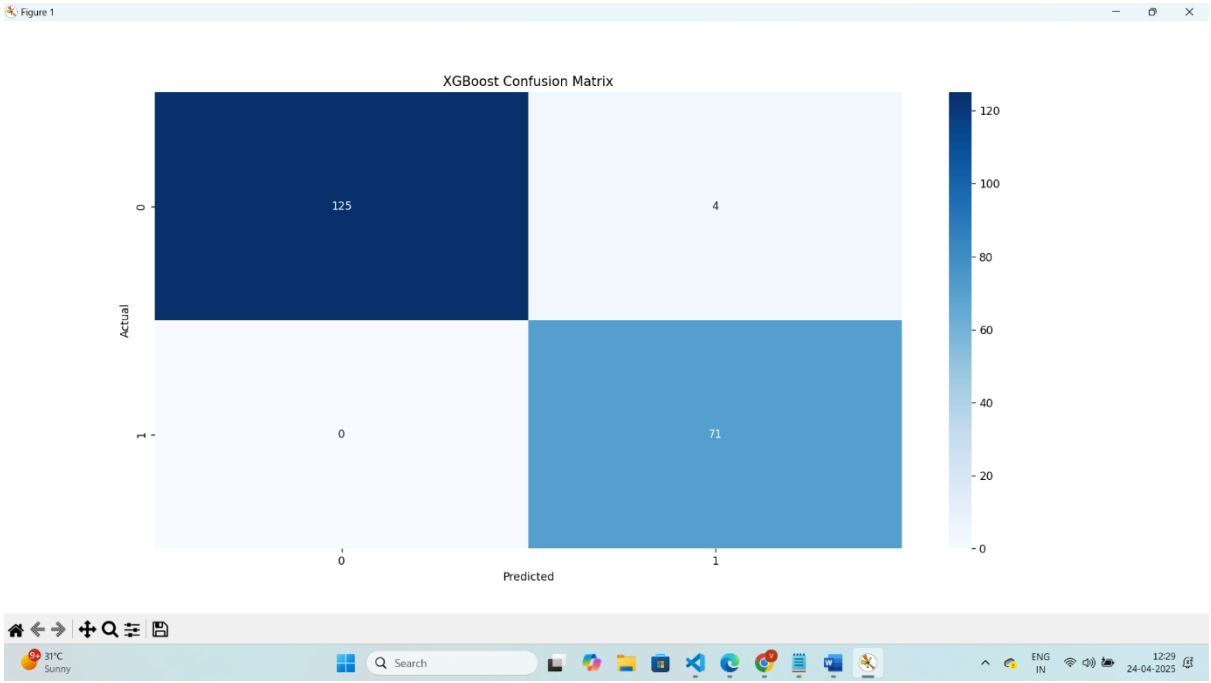
Logistic Regression Confusion Matrix:



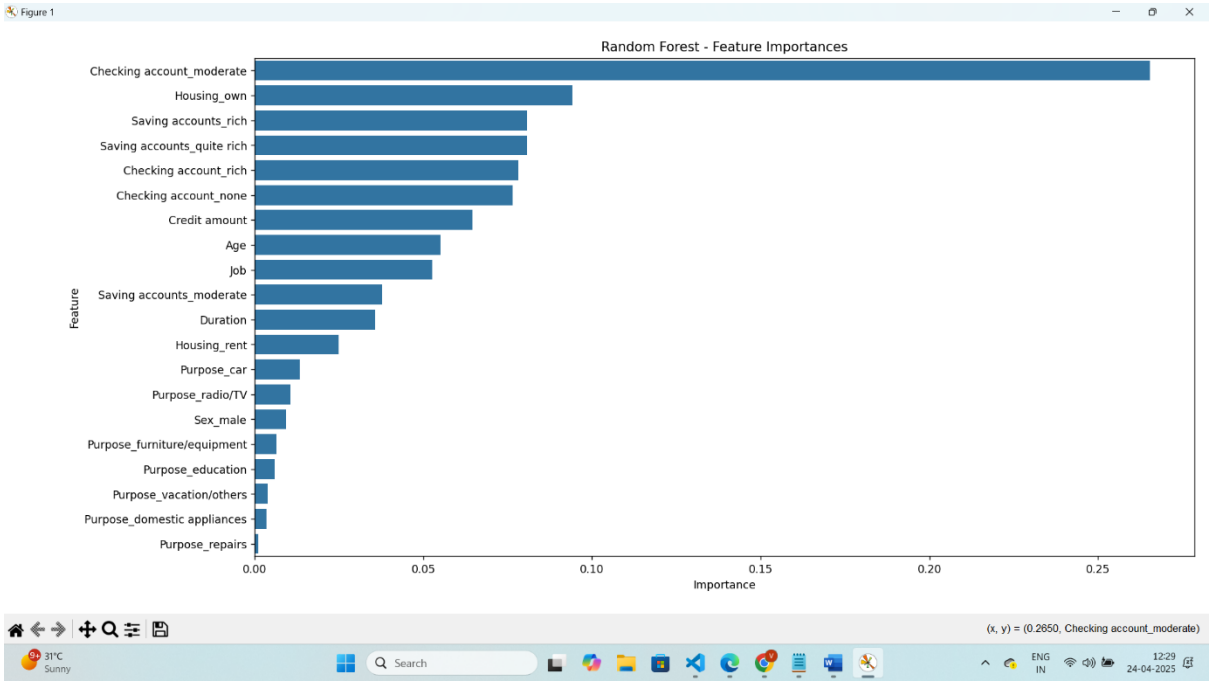
Random forest Confusion Matrix:



XGBoost Confusion Matrix:



Random Forest – Feature Implementation



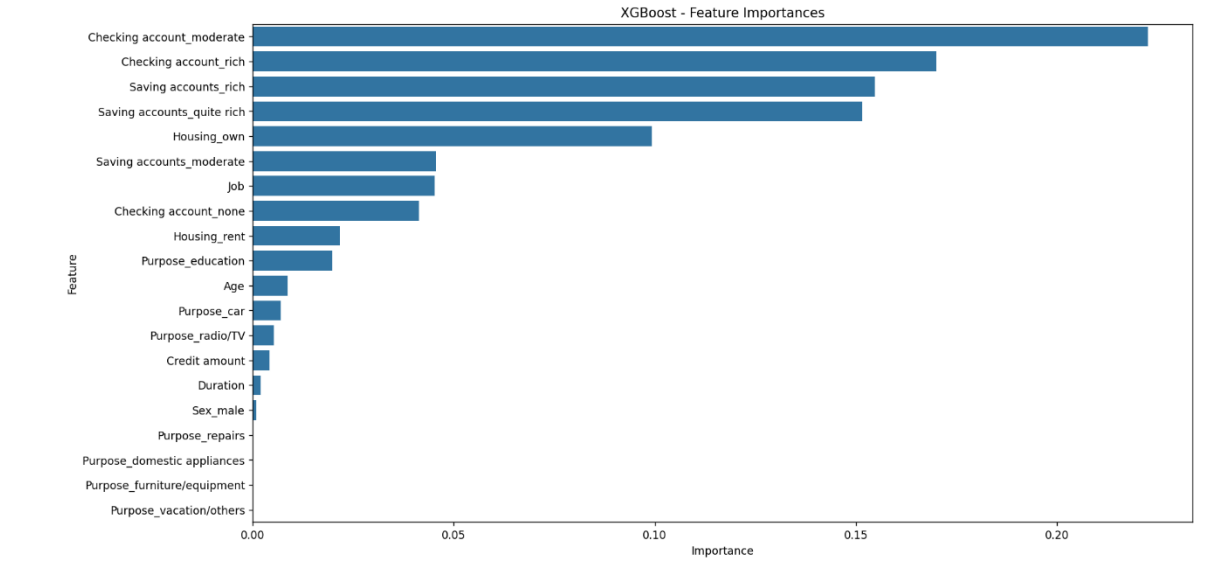

XGBoost - Feature Importances
Checking account_moderate
Checking account_rich
Saving accounts_rich
Saving accounts_quite rich
Housing_own
Saving accounts_moderate
Job
Checking account_none
Housing_rent
Purpose_education
Age
Purpose_car
Purpose_radio/TV
Credit amount
Duration
Sex_male
Purpose_repairs
Purpose_domestic appliances
Purpose_furniture/equipment
Purpose_vacation/others
Feature
0.00
0.05
0.10
0.15
0.20
Importance

Figure 1

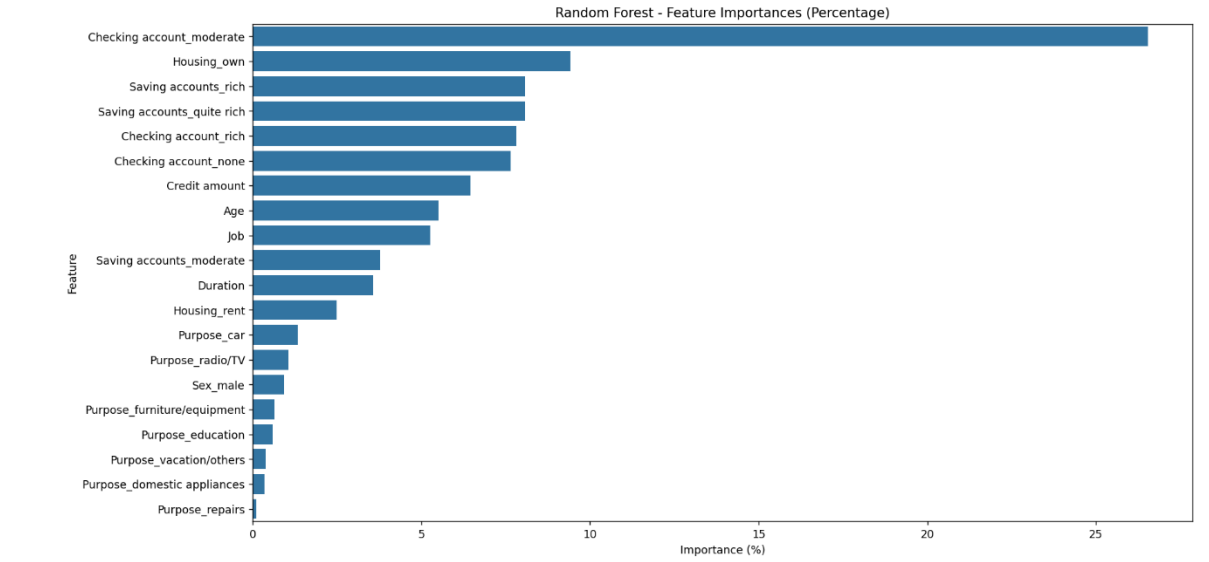
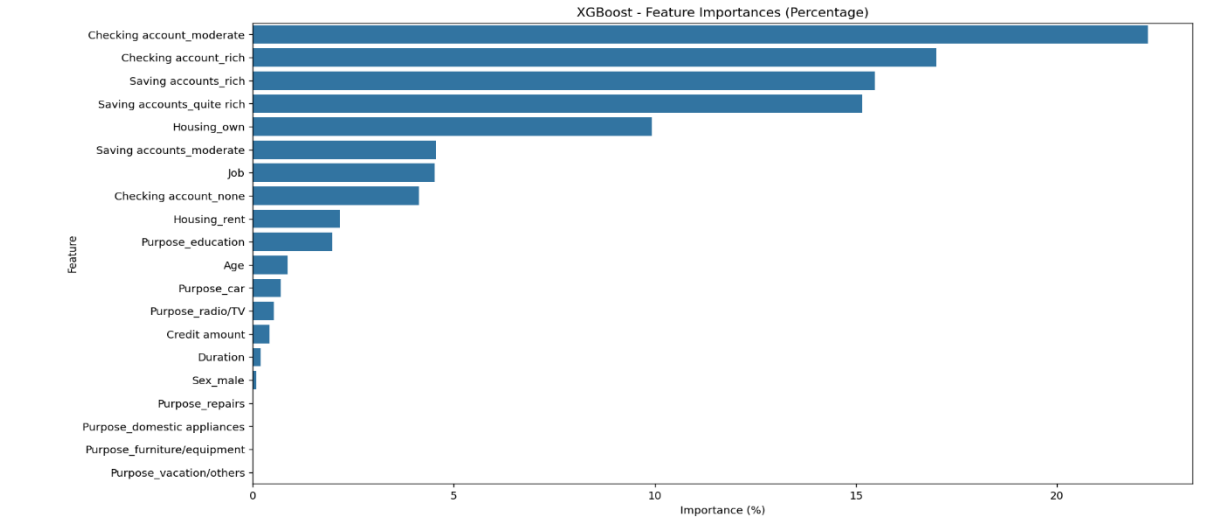

Random Forest - Feature Importances (Percentage)
Checking account_moderate
Housing_own
Saving accounts_rich
Saving accounts_quite rich
Checking account_rich
Checking account_none
Credit amount
Age
Job
Saving accounts_moderate
Duration
Housing_rent
Purpose_car
Purpose_radio/TV
Sex_male
Purpose_furniture/equipment
Purpose_education
Purpose_vacation/others
Purpose_domestic appliances
Purpose_repairs
Feature
0
5
10
15
20
25
Importance (%)

Figure 1


XGBoost - Feature Importances (Percentage)
Checking account_moderate
Checking account_rich
Saving accounts_rich
Saving accounts_quite rich
Housing_own
Saving accounts_moderate
Job
Checking account_none
Housing_rent
Purpose_education
Age
Purpose_car
Purpose_radio/TV
Credit amount
Duration
Sex_male
Purpose_repairs
Purpose_domestic appliances
Purpose_furniture/equipment
Purpose_vacation/others
Feature
0
5
10
15
20
Importance (%)

4. Presentation

Introduction

In this project, we aimed to develop a machine learning model that predicts the credit risk of loan applicants. This is crucial for financial institutions to assess whether a loan applicant is likely to default or successfully repay the loan. By leveraging machine learning, we can automate and enhance the credit evaluation process.

Methodology

We followed a step-by-step methodology to develop and evaluate the credit risk prediction model:

1. **Data Exploration and Preprocessing**:
    - We explored the German Credit dataset, consisting of features such as age, job type, housing status, credit amount, and loan duration. Preprocessing steps included handling missing data, encoding categorical variables, and scaling numerical features to make the dataset suitable for machine learning algorithms.
2. **Model Development**:
    - We tested three different classification models to predict the credit risk of loan applicants:
        - **Logistic Regression**
        - **Random Forest Classifier**
        - **XGBoost Classifier**
    - These models were trained on the processed data, and their performances were evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices.
    - The **XGBoost Classifier** was selected as the best-performing model due to its high predictive accuracy and robust performance.
3. **Model Interpretation**:
    - To interpret the model's predictions, we used **SHAP values** to identify the key features influencing the credit risk decision. Features such as **credit amount**, **loan duration**, and **housing status** emerged as significant contributors.
    - We also visualized **feature importance** using bar plots to highlight the most influential variables in the model's decision-making process.
4. **Deployment via Streamlit**:
    - A user-friendly web application was built using **Streamlit** to allow financial institutions and users to input new loan applications and receive predictions about credit risk.
    - The application provides the following features:
        - **Applicant Information Input**: Users input their details (e.g., age, job, credit amount) via an interactive form.

- **Prediction Results**: After submitting the form, the app predicts whether the applicant represents a **Good Credit Risk** or a **Bad Credit Risk**.
- **SHAP Explanations**: The app displays a detailed explanation of the model's reasoning for the prediction, highlighting the most influential features.
- **Feature Importance**: The app shows a visual representation of the feature importances that contributed to the model's decision.
- **Prediction Report Download**: Users can download a CSV file with their prediction and input data for further analysis.

5. **Model and Pipeline Saving**:
    - The model and preprocessing pipeline were saved using **joblib**, enabling easy loading and future predictions. This makes the model reusable for predictions on new data without retraining.

Results

The models were evaluated based on their accuracy in predicting the credit risk of loan applicants. Key metrics included:

- **Accuracy**: The proportion of correct predictions (both good and bad credit risks).
- **Precision, Recall, and F1-Score**: These metrics provided a deeper understanding of the model's ability to correctly identify good and bad credit risks.
- **Confusion Matrix**: This visualization revealed how well each model performed, with fewer false positives and false negatives indicating better performance.

The **XGBoost Classifier** outperformed other models in terms of accuracy and F1-score, and it became the model of choice for deployment.

Interactive Application and Explanation via Streamlit

The **Streamlit app** provides a seamless user interface where users can input their details and receive predictions instantly. Key features of the app include:

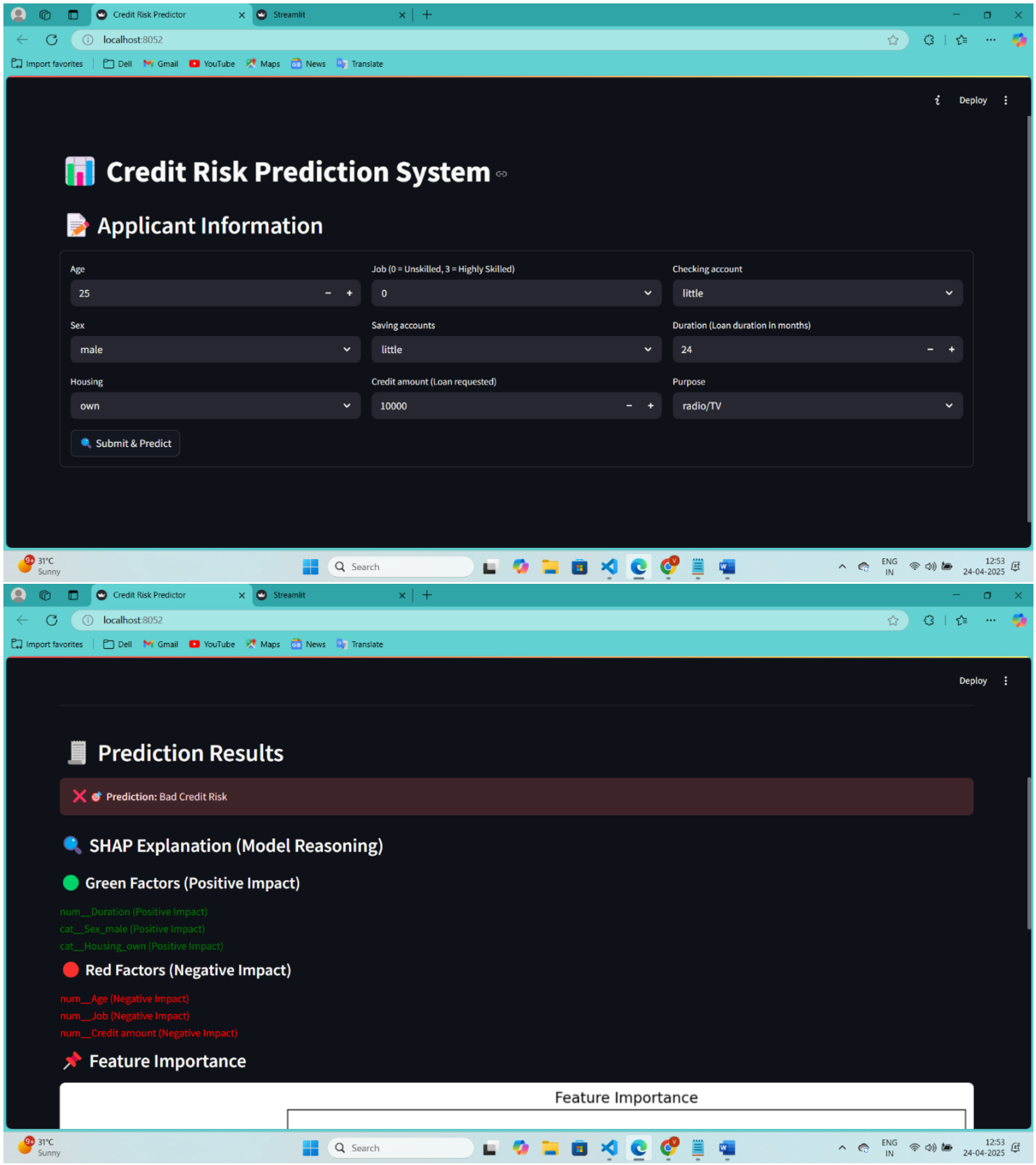
- **Applicant Input**: Users can enter their details such as age, job type, credit amount, housing status, and loan purpose.
- **Prediction Outcome**: After submitting the form, the app predicts whether the applicant poses a **Good** or **Bad Credit Risk**. The result is displayed with a color-coded response (green for good credit risk and red for bad).
- **SHAP Explanations**: The app explains the model's prediction by displaying the most influential factors (positive and negative impacts) using **SHAP** values. This ensures transparency and trust in the model's decision-making.
- **Feature Importance Visualization**: The app displays a bar chart showing the relative importance of each feature in the prediction, helping users understand what factors contribute the most to the decision.
- **Downloadable Report**: Users can download their prediction results in a CSV format for future reference.

The Streamlit app effectively communicates the results of the model and provides a user-friendly experience for non-technical users to make informed credit risk assessments.

Why This Approach Was Selected

The machine learning models, particularly **XGBoost** and **Random Forest**, were chosen due to their ability to handle complex datasets, their robustness against overfitting, and their interpretability. These models also provide important insights into feature importances, which is crucial in financial applications where understanding the rationale behind predictions is essential.

The use of **Streamlit** was selected to provide an interactive, web-based interface that makes the model accessible to users without deep technical knowledge. This ensures that financial institutions can easily integrate the model into their decision-making processes.



GitHub Repository: [click Here]

Conclusion

The credit risk prediction model provides financial institutions with a powerful tool for assessing the creditworthiness of loan applicants. By using **XGBoost** for prediction and **SHAP** for model interpretation, we ensure that the model's predictions are not only accurate but also explainable. The integration with **Streamlit** allows for a seamless user experience, enabling quick and transparent decision-making. This approach can help improve the efficiency of the credit evaluation process, reduce the risk of defaults, and optimize lending operations.