**Credit Risk Prediction Using Machine Learning**

**1. Introduction**

Accurate credit risk assessment is crucial for financial institutions to manage loan portfolios, reduce default rates, and ensure sustainable credit operations. The goal of this project is to develop a supervised learning framework that classifies applicants into *low-risk* (good credit) and *high-risk* (bad credit) categories using the German Credit dataset. Additionally, we analyze key influencing factors and propose recommendations to enhance credit evaluation processes.

**2. Dataset Overview**

* **Dataset**: German Credit Data (1,000 entries)
* **Attributes**: Includes demographic, financial, and loan-related variables such as Age, Sex, Credit amount, Duration, Saving accounts, etc.
* **Target Variable**: Risk is not explicitly provided and is derived using domain knowledge.

**3. Methodology**

**3.1 Data Preprocessing**

* **Handling Missing Values**:  
  Missing entries in Saving accounts and Checking account were imputed with 'unknown' to preserve the sample size.
* **Feature Encoding**:  
  Categorical features were transformed using label encoding or one-hot encoding, depending on the model.
* **Redundancy Removal**:  
  Dropped Unnamed: 0 as it was a non-informative index column.
* **Feature Scaling**:  
  Numerical features were standardized to improve convergence for certain algorithms

**3.2 Target Variable Construction**

In the absence of a labeled Risk column, a proxy risk label was generated using a rule-based heuristic:

*credit\_threshold = df['Credit amount'].median()*

*duration\_threshold = df['Duration'].median()*

*df['Risk'] = (*

*(df['Credit amount'] > credit\_threshold) &*

*(df['Duration'] < duration\_threshold)*

*).astype(int)*

* **Label 0**: Low risk
* **Label 1**: High risk

**Rationale**: High credit amounts over shorter durations often represent risky loans due to higher monthly repayment burdens.

**3.3 Model Selection**

We implemented and compared the performance of the following classification models:

|  | **Justification** |
| --- | --- |
| Logistic Regression | Baseline; interpretable |
| Random Forest Classifier | Handles feature interactions; robust to noise |
| XGBoost Classifier | Gradient boosting; high performance with complex data |

**3.4 Evaluation Metrics**

Model performance was evaluated using:

* **Accuracy**
* **Precision**
* **Recall**
* **F1-Score**
* **Confusion Matrix**

*from sklearn.metrics import classification\_report, confusion\_matrix*

*print(classification\_report(y\_test, y\_pred))*

*sns.heatmap(confusion\_matrix(y\_test, y\_pred), annot=True, fmt='d')*

**4. Results**

**4.1 Model Performance (Random Forest Example)**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 91% |
| Precision | 83% |
| Recall | 71% |
| F1-Score | 76% |

These results indicate that the model is particularly effective in distinguishing between high- and low-risk applicants, with a strong balance between precision and recall.

**4.2 Feature Importance**

The Random Forest model highlighted the following as most influential in credit risk classification:

| **Feature** | **Importance** |
| --- | --- |
| Credit amount | High |
| Duration | High |
| Age | Moderate |
| Checking account | Moderate |
| Purpose | Low |

Visualization:

*sns.barplot(x=model.feature\_importances\_, y=X.columns)*

**5. Discussion**

The constructed heuristic target variable provided a logical and data-driven approach for initiating the classification process in the absence of true risk labels. Among the tested models, Random Forest provided the best balance between accuracy and interpretability.

However, the current target definition is a simplification and may not fully capture the nuanced criteria used by financial institutions. The inclusion of external credit scores, employment history, and repayment history could significantly enhance prediction quality in a production setting.

**6. Conclusion**

* A machine learning pipeline was successfully developed to classify credit risk using features from the German Credit dataset.
* The most impactful predictors were Credit amount and Duration, which are intuitive given their influence on repayment capacity.
* Despite the lack of labeled ground-truth risk outcomes, a rule-based proxy approach provided meaningful classification results.
* The framework is modular and can be extended or retrained with new data or real-world risk labels.