# Report on: Credit Risk Prediction

Submitted by: Adarsh Joshi

Reference Id: CT20244464744

Question 1 of BFSI Hackathon

# Objective

To develop a machine learning model that classifies loan applicants as **Good Credit Risk** or **Bad Credit Risk** based on their profile using the **German Credit dataset**. The project also aims to:

* Identify key factors influencing credit risk.
* Suggest strategies to improve the credit evaluation process.
* Use model optimization techniques like **hyperparameter tuning** and **cross-validation** to ensure robust predictions.

# Dataset Overview

**Dataset:**german\_credit\_data.csv  
**Target Variable:** Risk (Good/Bad)

**Key Features Include:**

* Age, Sex, Job, Housing, Credit amount, Duration, Purpose, etc.

# Methodology

## 1. **Data Exploration & Preprocessing**

* Explored the feature distribution and class balance.
* No missing values were present.
* Handled categorical variables using **Label Encoding** for binary columns and **One-Hot Encoding** for multiclass categorical columns.
* Normalized numerical features using **StandardScaler**.
* Engineered polynomial interaction features where relevant (optional step).

## 2. Train-Test Split

* Split the data into **80% training** and **20% testing** sets using train\_test\_split().

## Model Development

Four classification models were initially implemented:

* **Logistic Regression**
* **Random Forest**
* **Support Vector Machine (SVM)**
* **XGBoost**

All models were evaluated using:

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

## Model Optimization

**XGBoost (Tuned with GridSearchCV)**

* Parameters Tuned: n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree
* Best F1-Score on Test: **~0.78**
* Top Features: Duration, Credit amount, Age, Purpose, and Job.

**Random Forest (Tuned with GridSearchCV)**

* Parameters Tuned: n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf, bootstrap
* Best F1-Score on Test: **~0.76**
* Top Features: Credit amount, Duration, Age, and Housing.

**SVM (Tuned with GridSearchCV)**

* Parameters Tuned: C, kernel, gamma
* Best F1-Score on Test: **~0.74**
* Less interpretable, but useful in high-dimensional space.

**Key Insights from Feature Importance**

* **Credit Amount** and **Loan Duration** were the most influential factors across models.
* **Age**, **Job**, and **Purpose** also played a significant role.
* Certain housing conditions and sex also had a mild influence on risk classification.

# Why This Approach?

* **Multiple models** were explored to compare performance.
* Logistic Regression was chosen as a baseline model due to its simplicity and ease of interpretation, especially in financial contexts where explainability is important. It also provides probability estimates that help in risk scoring and threshold-based decisions.
* Random Forest was selected for its ability to model complex, non-linear relationships and automatically handle feature interactions. Its ensemble approach makes it robust against overfitting and provides insights through feature importance scores.
* SVM was used for its effectiveness in small-to-medium-sized datasets and its ability to create optimal decision boundaries using margin maximization. With proper scaling and kernel tricks, it can handle non-linear relationships in high-dimensional data.
* **XGBoost** was selected for its superior ability to handle tabular data, robustness to feature interaction, and better handling of overfitting with built-in regularization.
* **Hyperparameter tuning** using **GridSearchCV** ensured that models were performing optimally on unseen data.
* **Cross-validation** added an additional layer of robustness by validating performance consistency.