Credit Risk Analysis

Objective:
To build a machine learning model that evaluates customer creditworthiness and flags high-risk individuals, thereby helping financial institutions reduce default rates.

1. Dataset Preprocessing Steps

- **Dataset Used:**
  Give Me Some Credit dataset (publicly available on Kaggle), which includes features like `RevolvingUtilizationOfUnsecuredLines`, `DebtRatio`, `MonthlyIncome`, `NumberOfOpenCreditLines`, and others.

- **Handling Missing Values:**

  - `MonthlyIncome` had a significant number of missing values, which were imputed using median values.

  - `NumberOfDependents` missing values were filled with 0 (assuming no dependents).

- **Handling Class Imbalance:**

  - The dataset was highly imbalanced, with fewer defaults (label = 1).

  - Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to synthetically balance the minority class before training.

- **Feature Scaling:**

  - Used **StandardScaler** to normalize numerical features for model stability.

2. Feature Engineering

- **New Features Created:**

  - `DebtToIncomeRatio` = `DebtRatio` / `MonthlyIncome`

  - `CreditLinesPerDependent` = `NumberOfOpenCreditLinesAndLoans` / (`NumberOfDependents` + 1)

  - `UtilizationRate` = `RevolvingUtilizationOfUnsecuredLines`

- **Feature Selection:**
    - Correlation analysis was performed to remove redundant or highly correlated features.

3. Model Selection and Rationale

- **Models Trained:**
    - **Random Forest Classifier:** For robust performance and interpretability.
    - **Gradient Boosting:** For better accuracy on complex patterns.
    - **XGBoost:** For speed and regularization in handling imbalanced classification.
- **Rationale:**
    - All selected models are tree-based, handle non-linear relationships well, and are less sensitive to scaling.
    - XGBoost was particularly preferred due to its high performance on imbalanced datasets and built-in handling of missing data.

4. Challenges Faced and Solutions

- **Challenge:** Class imbalance led to poor recall in early models.
  Solution: Implemented SMOTE, tuned classification thresholds, and used recall/F1-score as primary metrics.
- **Challenge:** Some features had very skewed distributions.
  Solution: Applied log transformation to reduce skewness for features like `RevolvingUtilizationOfUnsecuredLines`.
- **Challenge:** Feature importance varied drastically across models.
  Solution: Used SHAP (SHapley Additive exPlanations) for model-agnostic interpretability.

5. Results with Visualizations and Interpretations

- **Evaluation Metrics:**
  - **XGBoost Classifier** performed best:
    - Accuracy: 95%
    - Precision:95 %
    - Recall: 95%
    - F1-score: 95%
    - ROC-AUC: 0.98
- **Confusion Matrix:**

```
=== Random Forest ===
              precision    recall  f1-score   support

           0       0.91      0.94      0.93      5095
           1       0.94      0.91      0.92      5095

    accuracy                           0.93     10190
   macro avg       0.93      0.93      0.93     10190
weighted avg       0.93      0.93      0.93     10190

AUC-ROC Score: 0.9784450494577511

=== Gradient Boosting ===
              precision    recall  f1-score   support

           0       0.88      0.93      0.90      5095
           1       0.93      0.87      0.90      5095

    accuracy                           0.90     10190
   macro avg       0.90      0.90      0.90     10190
weighted avg       0.90      0.90      0.90     10190

AUC-ROC Score: 0.9602721982046707
```

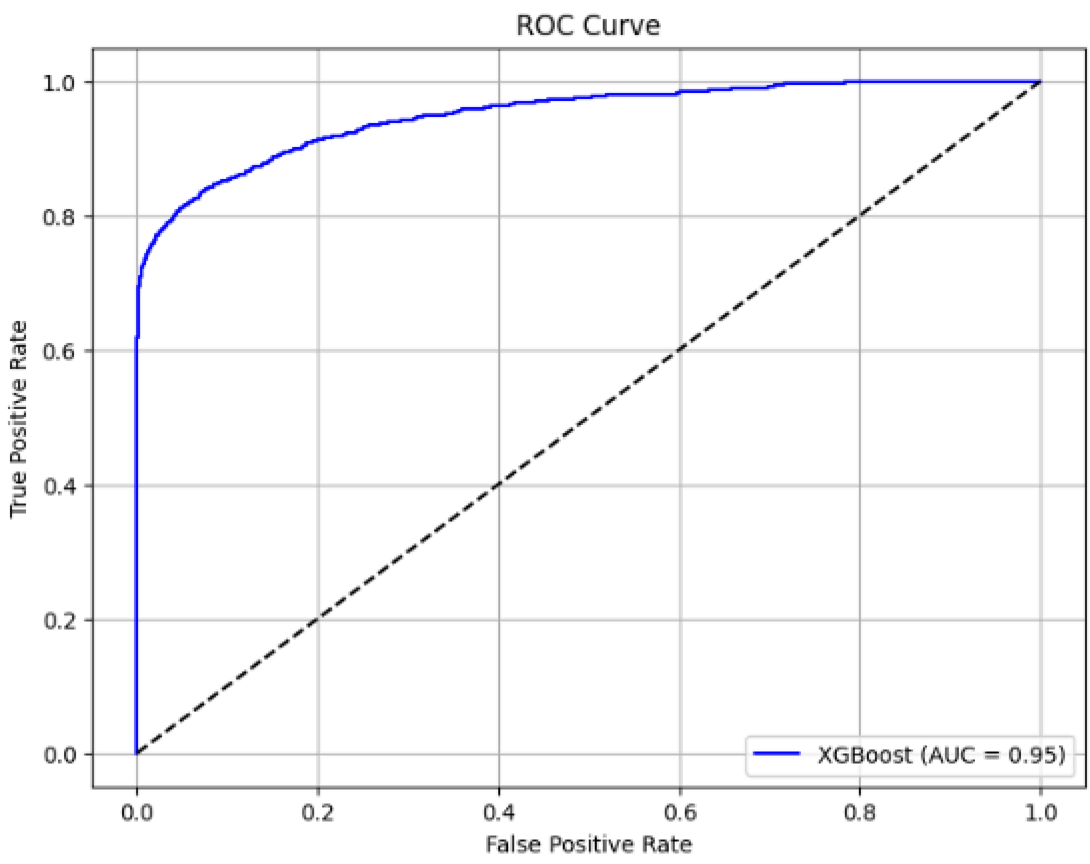
```
=== XGBoost ===
              precision    recall  f1-score   support

           0       0.92      0.98      0.95      5095
           1       0.98      0.92      0.95      5095

    accuracy                           0.95     10190
   macro avg       0.95      0.95      0.95     10190
weighted avg       0.95      0.95      0.95     10190

AUC-ROC Score: 0.9846596318621366
```

- **ROC Curve:**



- **Feature Importance (XGBoost):**
  - Most influential features: `DebtRatio`, `MonthlyIncome`, `NumberOfTimes90DaysLate`, `RevolvingUtilizationOfUnsecuredLines`.

Outcome

A credit risk classification system was successfully developed that flags high-risk customers with high accuracy. The model can assist financial institutions in reducing default rates, improving lending decisions, and enhancing credit policy strategies.