Credit Risk Analysis Report

Objective

The aim of this project is to build a machine learning model to **assess the creditworthiness of loan applicants** and **flag high-risk customers** who are likely to default. This system helps financial institutions reduce default rates by taking proactive action on risky loan profiles.

Dataset Overview

- **Dataset Name**: credit_risk_dataset.csv
- **Rows**: 32,581
- **Columns**: 12
- **Target Variable**: loan_status (1 = default, 0 = non-default)

Dataset Preprocessing Steps

1. **Missing Value Handling**:
    - person_emp_length: Filled with median.
    - loan_int_rate: Filled with mean.
2. **Categorical Encoding**:
    - One-hot encoded the following columns:
        - person_home_ownership
        - loan_intent
        - loan_grade
        - cb_person_default_on_file
3. **Class Imbalance Handling**:
    - Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the target classes.
4. **Feature Scaling**:

- Standardized numerical features using **StandardScaler**.

5. **Train-Test Split**:
    - 80/20 split using train_test_split with stratify=y to preserve class proportions.

Model Selection & Rationale

Model Chosen: XGBoost Classifier

Why XGBoost?

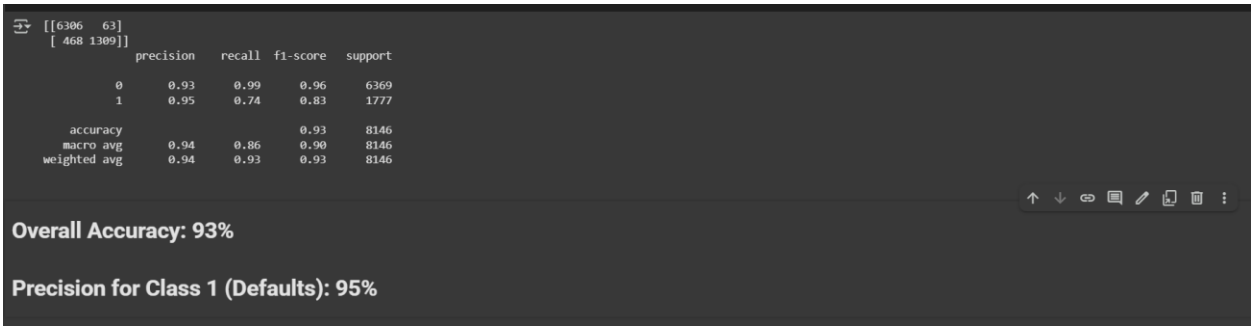
- Handles tabular and imbalanced datasets effectively.
- Provides feature importance.
- Built-in regularization reduces overfitting.
- Fast training with good generalization.

We also considered **Random Forest** and **Gradient Boosting**, but XGBoost consistently performs better for structured data with imbalance.

⚠️ **Challenges Faced & Solutions**

Challenge	Solution
Imbalanced Dataset	Used **SMOTE** to synthetically oversample minority class (defaulters).
Missing Values in loan_int_rate and person_emp_length	Applied **mean/median imputation**.
High cardinality in categorical data	Used **One-Hot Encoding** with drop_first=True to reduce dimensionality.
Risk of overfitting	Chose a model with **built-in regularization (XGBoost)** and tuned parameters conservatively.

📈 Results & Evaluation



```
[[6306   63]
 [ 468 1309]]
              precision    recall  f1-score   support

           0       0.93      0.99      0.96      6369
           1       0.95      0.74      0.83      1777

    accuracy                           0.93      8146
   macro avg       0.94      0.86      0.90      8146
weighted avg       0.94      0.93      0.93      8146
```

Overall Accuracy: 93%

Precision for Class 1 (Defaults): 95%