Credit Scoring Model

📌 Project Overview

This project aims to predict whether a customer is likely to default on their credit payment using machine learning techniques. A detailed analysis of financial data is performed to identify key factors influencing credit risk. Several classification models were trained and evaluated to determine the most accurate predictive approach.

🎯 Objectives

• Develop a credit scoring model using historical financial data
• Compare multiple classification algorithms to identify the best-performing model
• Handle class imbalance to improve model accuracy
• Evaluate model performance using industry-standard metrics
• Provide insights for potential future improvements and deployment strategies

📊 Dataset Information

The dataset used in this project contains customer financial records, including past payment history, credit limits, and other relevant features. The target variable indicates whether a customer defaulted on their payment.

Total Rows: 30,000
Total Columns: 24
Target Variable: 'default.payment.next.month'

🔍 Data Preprocessing

• Dropped irrelevant columns (e.g., 'ID')
• Checked and handled missing values (if any)
• Encoded categorical variables using Label Encoding
• Applied StandardScaler to normalize numerical features
• Used SMOTE to handle class imbalance and improve prediction fairness

📌 Exploratory Data Analysis (EDA)

EDA was performed to understand feature distributions and relationships. Key insights include:
• Distribution of the target variable (high imbalance observed)
• Correlation heatmap to identify relationships between features
• Visualization of important financial attributes (e.g., credit limit, past payments)

📌 Model Training & Comparison

Four machine learning models were implemented and compared:
✓ Logistic Regression
✓ Decision Tree Classifier

✓ Random Forest Classifier (Hyperparameter Tuned)
✓ XGBoost Classifier

Hyperparameter tuning was applied to Random Forest using GridSearchCV for optimal performance.

📊 Model Evaluation

Each model was evaluated based on multiple performance metrics:
• Accuracy
• Precision, Recall, and F1-Score
• ROC-AUC Score (for measuring overall performance)
• Confusion Matrix (for visualizing classification performance)

🏆 Results & Best Model Selection

• The Random Forest model with hyperparameter tuning achieved the highest accuracy and AUC-ROC score.
• Accuracy: 85%
• ROC-AUC Score: 0.92
• The model successfully differentiates between defaulters and non-defaulters.

🚀 Future Improvements

Although the model performed well, further improvements can be made:
• Implementing feature selection to remove less relevant features
• Exploring deep learning techniques for better predictive accuracy
• Deploying the model using Flask/Streamlit for real-world usage
• Testing alternative models like LightGBM and CatBoost

☑ Conclusion

This project successfully developed a credit scoring model to predict customer defaults. By applying data preprocessing techniques, handling class imbalance, and comparing multiple machine learning models, an optimized model was selected. The findings of this project can aid financial institutions in assessing credit risk effectively.