**Credit Card Default Prediction**

**Overview**

Credit card fraud is a significant challenge in the financial industry, impacting both financial institutions and customers. As fraudsters develop more sophisticated techniques, it becomes crucial to implement advanced fraud detection models. This project aims to predict credit card payment defaults using machine learning models, leveraging customer transaction history and financial attributes to assess the likelihood of default.

**Objectives**

The primary goal of this project is to develop a robust credit card fraud detection model that can accurately identify transactions at risk of default. By leveraging machine learning techniques, the model analyzes historical transaction patterns and payment behaviors to predict the likelihood of future defaults.

**Features**

* **Data Preprocessing:** Handling missing values, feature scaling, and encoding categorical variables.
* **Exploratory Data Analysis (EDA):** Understanding data distribution and feature correlations.
* **Model Training:** Implementing Logistic Regression, Random Forest, and XGBoost.
* **Model Evaluation:** Assessing performance using precision, recall, confusion matrix, and AUC-ROC score.
* **Model Deployment:** Saving the best-performing model for future inference.

**Dataset Overview**

The dataset used for this project is the **UCI Credit Card dataset**, obtained from the UCI Machine Learning repository. It consists of customer credit card transaction history and financial attributes from Taiwan between April 2005 and September 2005.

**Dataset Information**

* **Number of Entries:** 30,000
* **Number of Features:** 25
* **Target Variable:** default.payment.next.month (1 for default, 0 for non-default)

**Key Features:**

* **Demographics:** AGE, SEX, MARRIAGE, EDUCATION
* **Credit History:** LIMIT\_BAL, PAY\_0 to PAY\_6 (payment history for six months)
* **Transaction Details:** BILL\_AMT1 to BILL\_AMT6 (bill amounts), PAY\_AMT1 to PAY\_AMT6 (payment amounts)

**Installation**

To set up and run this project on your local machine, follow these steps:

1. Clone the repository:

git clone https://github.com/your-username/credit-card-default.git

1. Navigate to the project directory:

cd credit-card-default

1. Create a virtual environment and activate it:
2. python -m venv venv

source venv/bin/activate # On Windows: venv\Scripts\activate

1. Install dependencies:

pip install -r requirements.txt

1. Place the dataset in the data/ directory.

**Model Training**

Run the following script to preprocess data, train models, and evaluate performance:

python train.py

**Benchmark Models**

To establish a baseline for fraud detection, three machine learning models were implemented:

1. **Logistic Regression**
2. **Random Forest**
3. **XGBoost**

**Model Training Pipeline:**

* **Data Preprocessing:** Handling missing values, feature scaling, and encoding categorical features.
* **Feature Engineering:** Selecting the most relevant features for training.
* **Model Training:** Applying cross-validation for improved generalization.
* **Model Evaluation:** Assessing accuracy, precision, recall, F1-score, and ROC-AUC.

**Model Evaluation**

Performance metrics of the trained models:

**Logistic Regression:**

* High recall for non-defaults: **0.97**
* AUC-ROC score: **0.7103**
* Weak recall for defaults: **0.24**

**Random Forest:**

* Improved recall for defaults: **0.35**
* AUC-ROC score: **0.7750**

**XGBoost:**

* Best AUC-ROC score: **0.7795**
* Improved precision for defaults: **0.61**
* Recall for defaults: **0.37**

**Key Insights**

* All models effectively identify non-defaulting customers but struggle with default recall, which is critical for mitigating credit risk.
* **XGBoost** showed the best overall performance but still requires further optimization.

**Recommendations**

To improve the fraud detection model:

1. **Enhance Feature Engineering** – Extract meaningful features from transaction patterns.
2. **Balance the Dataset** – Use oversampling/undersampling techniques like SMOTE.
3. **Adjust Classification Thresholds** – Tune the decision threshold to improve recall for defaults.

**Importance of Fraud Detection**

Credit card fraud detection is vital for financial institutions to:

* Mitigate financial losses
* Prevent unauthorized transactions
* Maintain trust and confidence in the financial system

**Deployment**

To use the trained model for predictions:

python predict.py

**Conclusion**

The benchmark models (Logistic Regression, Random Forest, and XGBoost) with cross-validation provide a solid foundation for evaluating credit card fraud detection. Future work should focus on improving recall for defaults through better feature engineering and threshold adjustments.

🚀 **Next Steps:** Explore hyperparameter tuning, deep learning models, and real-time fraud detection pipelines!