Project Report on Credit Card Fraud Detection using Machine Learning 1. Introduction Credit card fraud is a major issue in today's digital economy, causing billions of dollars in financial losses worldwide. Detecting fraudulent transactions quickly and accurately is critical for both financial institutions and customers. This project aims to build a machine learning model to detect fraudulent credit card transactions by handling the class imbalance problem using sampling techniques. 2. Objective The main goal of this project is to develop a predictive system that can:
- Accurately identify fraudulent transactions.
- Handle severe class imbalance in the dataset.
- Compare the effectiveness of under-sampling and over-sampling strategies. 3. Dataset Description - **Source**: Kaggle – Credit Card Fraud Detection Dataset.
- **Records**: 284,807 transactions.
- **Features**:
  Time: Seconds elapsed between each transaction.
  V1–V28: Anonymized numerical features (derived from PCA).
  Amount: Transaction amount.
  Class: Target variable (0 = Normal, 1 = Fraud).
- **Imbalance**: Fraud cases ≈ 492 (0.17%), Genuine cases ≈ 284,315 (99.83%). 4. Data Preprocessing 1. **Data Cleaning**: Checked for missing values (none found).
2. **Feature Scaling**: StandardScaler applied to the Amount and Time features.
3. **Class Balancing**:
  - Under-Sampling: Reduced majority (non-fraud) samples to match minority (fraud).
  - Over-Sampling: Used SMOTE (Synthetic Minority Over-sampling Technique) to increase fraud samples. 5. Modeling Approach The following machine learning models were applied:
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier

Train-Test Split: 70% training, 30% testing.
Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC. 6. Results **Without Sampling**: High Accuracy (>99%) but very poor Recall for fraud detection due to imbalance.

Under-Sampling: Accuracy dropped due to reduced data size. Recall improved significantly but at the cost of higher false positives.

Over-Sampling (SMOTE): Balanced performance with high Recall and reasonable Precision. XGBoost gave the best results with high ROC-AUC score (>0.98).

Key Finding: Over-Sampling with advanced classifiers (Random Forest/XGBoost) provides the most reliable fraud detection system. 7. Discussion **Trade-offs**:
- Under-Sampling: Faster but loses genuine transaction data.
- Over-Sampling: More computationally expensive but preserves data distribution.

In fraud detection, Recall is more important than Accuracy, since missing fraudulent transactions (False Negatives) is more costly than incorrectly flagging genuine ones. 8. Conclusion The project successfully demonstrated how machine learning models can be used to detect fraudulent transactions.
Handling class imbalance is critical for building effective fraud detection systems.
The best performing approach in this project was Over-Sampling (SMOTE) + XGBoost, achieving strong Recall and ROC-AUC values.