Credit Card Fraud Detection Project Report

1. Project Overview

The goal of this project is to develop a machine learning model that can detect fraudulent credit card transactions from a dataset. Fraudulent transactions are a growing concern in the financial sector, and detecting them accurately can help reduce financial losses. In this project, I employed two machine learning models — **Logistic Regression** and **Random Forest** — to predict whether a given transaction is fraudulent or genuine.

Dataset Description

The dataset used for this project is from Kaggle's Credit Card Fraud Detection Dataset, which contains anonymized features related to credit card transactions. Key features include:

- **Time**: Time elapsed between the current transaction and the first transaction in the dataset.
- **V1-V28**: Anonymized features generated via PCA transformation to protect sensitive information.
- **Amount**: The amount of the transaction.
- **Class**: The target variable (1 for fraudulent, 0 for genuine).

The dataset contains over 280,000 transactions, with a significant class imbalance (only about 0.17% of the transactions are fraudulent).

---

2. Data Preprocessing

2.1. Data Cleaning

- The dataset was first loaded and examined for missing values or inconsistencies.
- No missing values were detected in the dataset, so no imputation or data removal was necessary.

2.2. Data Splitting

- The dataset was split into training and test sets using an **80-20 split**. The training set was used to train the machine learning models, and the test set was used to evaluate the models' performance.

2.3. Handling Class Imbalance

- Given the significant class imbalance (fraudulent transactions constitute less than 1% of the data), various techniques were explored to address this, such as **SMOTE (Synthetic Minority Over-sampling Technique)**, which was used to oversample the minority class (fraudulent transactions) in the training set.

---

3. Machine Learning Models

3.1. Logistic Regression

- Logistic Regression is a simple linear model that predicts the probability of the occurrence of an event (in this case, fraud) based on input features.
- The model was trained on the preprocessed data and tuned using **cross-validation** to find the best hyperparameters.

3.2. Random Forest

- Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. This method is robust and works well for high-dimensional datasets like ours.
- Random Forest was trained with a default hyperparameter setup and tuned for optimal performance.

---

4. Model Evaluation

The performance of the models was evaluated using several metrics, including **Accuracy**, **Precision**, **Recall**, **F1-Score**. Since the dataset is highly imbalanced, **F1-Score** was chosen as the primary evaluation metrics, as they provide a better understanding of the model's performance on the minority class (fraudulent transactions).

4.1. Evaluation Metrics

- **Accuracy**: Proportion of total correct predictions (both fraudulent and non-fraudulent).
- **Precision**: Proportion of predicted fraudulent transactions that were actually fraudulent.
- **Recall**: Proportion of actual fraudulent transactions correctly identified.
- **F1-Score**: Harmonic mean of precision and recall, which balances both.

4.2. Results

Logistic Regression

- **F1-Score**: 54%
- **Precision**: 39%
- **Recall**: 88%

Despite achieving a relatively decent Recall, the F1-Score of Logistic Regression is lower due to its inability to capture fraudulent transactions effectively. This is likely because Logistic Regression struggles with the class imbalance and the complex nature of fraud detection.

Random Forest

- **F1-Score**: 86%
- **Precision**: 86%
- **Recall**: 87%

Random Forest significantly outperformed Logistic Regression. With a higher F1-Score, precision, and recall, the Random Forest model is better at identifying fraudulent transactions. It benefits from its ensemble nature, allowing it to handle the data's complexity and class imbalance more effectively.

Comparison of Results:

Model	F1-Score	Precision	Recall
Logistic Regression	54%	39%	88%
Random Forest	86%	86%	87%

---

5. Conclusion

The **Random Forest** model outperforms the **Logistic Regression** model in detecting fraudulent credit card transactions. With an **F1-Score of 86%**, it is more effective at balancing precision and recall, making it a better choice for fraud detection in this highly imbalanced dataset.

Key Takeaways:

- **Logistic Regression** provided a baseline, but its performance was limited due to the class imbalance and its simplicity.
- **Random Forest**, as an ensemble model, performed significantly better, achieving high precision and recall, and is more suitable for the fraud detection task.
- Given the class imbalance, using **SMOTE** helped to improve model performance by oversampling the minority class (fraudulent transactions).

---

6. Future Work

There are several potential improvements for this project:

- **Hyperparameter Tuning**: Further tuning of model parameters (especially for Random Forest) using grid search or random search can potentially improve the results.
- **Ensemble Methods**: Combining the predictions from multiple models (e.g., stacking Logistic Regression, Random Forest, and XGBoost) could enhance performance.
- **Advanced Models**: Trying more complex models such as **XGBoost**, **LightGBM**, or **Neural Networks** might yield even better results.
- **Real-time Detection**: Implementing the model into a real-time fraud detection system where transactions can be monitored and flagged for fraud in real time.