***Credit Card Fraud Detection Project Report***

**1. Introduction**

This project aims to develop a model for detecting fraudulent credit card transactions. The dataset used is highly imbalanced, with a significantly higher number of legitimate transactions compared to fraudulent ones.

**2. Data Preprocessing**

* **Dataset Loading:** The credit card transaction data is loaded into a pandas DataFrame using pd.read\_csv('creditcard.csv').
* **Data Exploration:** Basic data exploration is performed using functions like head(), tail(), info(), and isnull().sum() to understand the data structure, data types, missing values, and basic statistics.
* **Handling Class Imbalance:** The dataset is highly imbalanced. To address this, undersampling is applied to create a balanced dataset for model training. This is done by randomly selecting a subset of legitimate transactions equal in size to the number of fraudulent transactions.
* **Data Splitting:** The dataset is split into features (X) and target (Y), where X contains all columns except 'Class' (the target variable) and Y contains the 'Class' column. The dataset is further split into training and testing sets using train\_test\_split with a test size of 20% and stratification to ensure similar class distributions in both sets.

**3. Model Training and Evaluation**

* **Logistic Regression:** A Logistic Regression model is trained on the training data using model.fit(X\_train, Y\_train).
* **Random Forest:** A Random Forest model is trained on the training data using model.fit(X\_train, Y\_train).
* **Model Evaluation Metrics:** Accuracy score, precision, recall, and F1-score are used to evaluate the models' performance.
* **Accuracy on Training and Testing Data:** Accuracy is calculated and printed for both the training and testing sets for both models.
* **Visualization of Accuracy:** Bar charts are generated to compare the training and testing accuracy of both models.

**4. Prediction Function**

* A function named predict\_fraud is defined to predict whether a transaction is fraudulent or legitimate.
* This function takes transaction data as input, converts it into a pandas DataFrame, uses the trained model (model) to make a prediction, and returns 'Fraud' or 'Legit' accordingly.

**5. Results and Discussion**

* The accuracy of the models on the training and testing data is reported.
* The Random Forest model demonstrates better performance than the Logistic Regression model based on the reported metrics and visualization.
* The predict\_fraud function provides a way to apply the trained model for real-time fraud prediction.

**6. Conclusion**

* The project successfully demonstrates building a credit card fraud detection model using machine learning.
* Undersampling is used to handle class imbalance, and model evaluation is performed using relevant metrics.
* The Random Forest model appears to be a more effective choice compared to the Logistic Regression model for this particular problem based on the presented analysis.

**7. Future Improvements**

* Exploring alternative techniques for handling class imbalance, such as oversampling or SMOTE.
* Fine-tuning model hyperparameters using techniques like grid search or cross-validation.
* Evaluating the model's performance on a larger and more diverse dataset.
* Implementing the model for real-time fraud detection in a credit card transaction system.