Credit Card Fraud Prediction Report

1. Background and Context to the Problem Statement

Credit card fraud is the use of credit cards without the owner's permission to make a purchase or gain access to funds, and involves crime. According to this report, card fraud losses increased by more than 10% globally from 2020 to 2021, emphasizing the seriousness of the problem and the need for an accurate model. Fraudulent transactions are rare in nature, and therefore constitutes a challenge for a model to learn appropriately. There is a cost associated with flagging too many transactions and a big one for giving false negatives.

2. Identification and description of data set, along with the source.

The dataset can be found in Kaggle. It is a simulated credit card transaction dataset containing legitimate and fraud transactions from 01/01/2019 to 12/31/2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants. The test dataset consists of 555719 observations, and the train data set 1048574. The dataset is quite imbalanced with only 0.6% of observations in the minority class (fraud). This implies that the majority of transactions are non-fraudulent, which can pose challenges in building a predictive model that accurately identifies fraud cases. The features data is:

trans_date_trans_time: The date and time when the transaction occurred. **cc_num:** The credit card number used in the transaction. **merchant:** The entity or store where the transaction took place. **category:** The category of product or service purchased. **amt:** The amount of money involved in the transaction. **first:** The first name of the cardholder. **last:** The last name of the cardholder. **gender:** The gender of the cardholder. **street:** The street address of the cardholder. **city:** The city where the cardholder resides. **state:** The state where the cardholder resides.	**zip:** The zip code of the cardholder's address. **lat and long:** The latitude and longitude coordinates of the **city_pop:** The population of the city where the cardholder resides. **job:** The occupation or job title of the cardholder. **dob:** The date of birth of the cardholder. **trans_num:** A unique identifier for each transaction. **unix_time:** The UNIX timestamp of the transaction time. **merch_lat and merch_long:** Latitude and longitude coordinates of the merchant. **is_fraud:**The target variable indicating whether the transaction is fraudulent (1) or not (0).

3. Proposed ML Techniques to solve the problem

This problems is a classification problem, and therefore we propose to explore the following type of models:

- Decision Trees
- Logistic Regression
- Support Vector Machines
- KNN
- Neural Network
- Ensemble Methods like:
    - Random Forests
    - XGBoost
    - AdaBoost
    - HistGradientBoost
- Feature Engineering
    - Principal Component Analysis (if needed)

For preprocessing, we would consider:

- Standardization/Normalization
- Encoding categorical variables
- Missing value imputation

Because this is a highly imbalanced dataset, we would also try:

- Oversampling the minority or undersampling the majority class
    - Random over/under sampling, SMOTE (Synthetic Minority Over-sampling Technique).
- Reconsider the evaluation metric of the predictive model
    - Use metrics like Precision, Recall, F1-Score, or Area Under the ROC Curve (AUC-ROC) to evaluate model performance more effectively
- Use Ensemble techniques
    - Use ensemble methods like Random Forest that naturally handle imbalance. Algorithms like AdaBoost or Gradient Boosting can be utilized to focus more on the minority class, improving their performance on imbalanced datasets.