**Title: Fraud Detection Model using Categorical Features**

**Team Name - HIT**

**Overview:**

In this data science project, we aimed to develop a fraud detection model using categorical features to identify fraudulent and non-fraudulent transactions. The project involved analyzing the distribution of fraud and non-fraud cases across categorical columns, scaling the data using Standard Scaler, applying dimensionality reduction techniques, and ultimately building a classification model.

**Data Exploration:**

During the initial data exploration phase, we observed that fraudulent cases appeared to be uniformly distributed among the various classes in the categorical columns. There were no clear or glaring predictors that could easily distinguish fraudulent from non-fraudulent transactions. Surprisingly, a similar uniform distribution was observed for non-fraudulent cases.

**Data Preprocessing:**

To prepare the data for modeling, we performed the following preprocessing steps:

* Standard Scaling: We used the Standard Scaler to standardize the numerical features, ensuring that they all have a mean of 0 and a standard deviation of 1. This step was essential for models sensitive to feature scales.
* Dimensionality Reduction: We applied dimensionality reduction techniques to reduce the complexity of the dataset while retaining as much information as possible. This step helped to mitigate the curse of dimensionality and improve model performance.
* Data preparation and categorical preprocessing with Target Encoding: we created a small sample from the entire dataset to fit our encoding model. This is to prevent data leakage.

**Hypothesis testing:**

We performed Identification of suitable predictors using Analysis of Variance (ANOVA), Chi Squared Contingency test and Pearson's Correlation coefficient

For the ANOVA test, we chose a critical value of 0.05. If the p-value is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis, otherwise we accept the null hypothesis and reject the alternative hypothesis. From the results, it was clear that no statistically significant relationship exists between the numeric features and the target variable.

The results of the Pearson's correlation coefficient further corroborated the findings from the result of the ANOVA Test.

Using Chi Squared Contingency Test only one feature ("Country Code") appears to be correlated with our target variable. This show that 99% of the columns in this dataset cannot be used to predict the Target as no relationship exists.

**Data preparation and categorical preprocessing with Target Encoding**

Through our EDA and statistical tests, many of the domain specific features (especially the categorical ones) that one might conceive would give good prediction on Fraudulent transactions actually show zero correlation to the target variable ("Fraudulent Flag"). I believe that this is caused in part by the high cardinality within the categorical features. High cardinality hampers the performance of machine learning models on this dataset as they can barely make connections with relevant features. This will lead the model to opt for guessing (which is characterized by an accuracy of about 50% or less) rather trying to understand the data and we actually ran in to this issue during some of our first trials.

For example, the "IP Address" column should actually be a very good predictor of a transaction's validity, as fraudsters often make fraudulent transaction remotely through computers not registered with the user's account. However, the cardinality of the "IP Address" is extremely high as the column contains

The same is true for the "Authentication Method" column as fraudsters are more likely gain access to the user's account and authenticate their transactions through weaker methods like Geolocation verification as they can easily mask thier location with a VPN. So, under ideal circumstances it follows that weaker authentication methods should have more cases of fraudulent transaction. However, as a result of the 39 unique values, that column has a very high cardinality.

Even with the current state of the dataset, Target encoding can help concretize expected relationships between the categorical features and the target variable. In target encoding, an encoder is trained on a small random sample of the dataset, then transforms the categorical labels in the original dataset with the probability a transaction in that category being Fraudulent. This approach will help the Machine learning models make better associations with relevant categorical datasets.

In our implementation, we will be assuming that the distribution of fraudulent cases within the dataset are a rough estimate of what happens in the real world.

**Model Building:**

Approach 1-

**Plain logistic regression classifier**: We chose to use logistic regression in our fraud classification project because of its simplicity and efficiency, align perfectly with the demands of real-time fraud detection. Its interpretable nature facilitates clear insight into the factors influencing fraudulent activities. In terms of data preprocessing, we employed Standard Scaler to standardize feature scales, ensuring uniformity across our dataset. Additionally, Principal Component Analysis (PCA) was leveraged for dimensionality reduction, enhancing computational efficiency and preventing overfitting. With this model we were able to achieve an **Accuracy of about 50 %** at identifying fraud and non-fraud cases.

Approach 2-

**Xgb Classifier:** We chose to use a XGBoost classifier in our fraud classification project to try achieve optimal performance. XGBoost, a powerful ensemble learning algorithm, offers remarkable versatility and efficiency, making it an excellent choice for intricate tasks like fraud detection. To fine-tune our model, we diligently applied both random search and grid search hyperparameter optimization techniques. Remarkably, both methods yielded comparable results, with the model having **50% accuracy** also at identifying fraud and non-fraud cases.

Approach 3-

**Plain decision tree classifier:** We chose this model for the simplicity and interpretability of a plain Decision Tree classifier. We also had similar results with an **Accuracy of 50%.**

Approach 4-

**Ridge regression classifier and grid search:** We used Ridge Regression because of its regularization properties. Ridge Regression offers us a balanced blend of simplicity and flexibility, allowing us to navigate the intricacies of fraud detection effectively. The results remained unchanged at **50% also.**

Approach 5-

**Autoencoder representations feeding a decision tree:** This approach trains an autoencoder which is a feed forward network whose input and output layers have the same shape and whose goal is to try and accurately transform the data into a richer representation. It also had an **accuracy of about 50 %** in identifying fraud and a f1 score of 50.

**In all it appears that most of our models have similar accuracy and F1 scores, this shows how data can hamper a model's performance. With Better data we believe that any of these models can be used to reasonable distinguish between fraudulent and non-Fraudulent transactions.**