**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.

**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 **precision of 73%** at identifying fraud cases **Recall of 74%** and **Accuracy of about 74%** 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, showing how robust our model's parameter settings are. With this model we also achieved similar results **precision of 74%** at identifying fraud cases **Recall of 73%** and **Accuracy of about 73%** 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. Decision Trees excel in revealing decision pathways and are an excellent choice for gaining insights into the factors that contribute to fraud. We also had similar results with an **Accuracy of 73%.**

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. To optimize our model's performance, we iterated over various values in search for the optimal alpha parameter, a critical component of Ridge Regression's regularization. Our iterative approach involved evaluating various alpha values to identify the best-performing configuration. With this model we also had similar results with it being **73% accurate**.

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 fell short of expectations with an accuracy of about 50 % in identifying fraud and a f1 score of 50.

Approach 6-

**Catboost:** We chose a CatBoost classifier, due to its ability in handling categorical data and delivering high-quality predictions. CatBoost's innovative approach to gradient boosting minimizes the need for extensive data preprocessing, allowing us to focus on the core of our fraud detection task. We also had similar results with an **accuracy of 73%** and similar f1 score.

**Conclusion:**

In conclusion, this data science project focused on developing a fraud detection model based on categorical features. Despite the absence of clear predictors in the categorical columns, our model, which included data scaling and dimensionality reduction, demonstrated a promising **accuracy of 73%** in identifying both legitimate and fraudulent transactions.