Title: Fraud Detection Model using Categorical Features

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 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:

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.

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.

Model Building:

Approach 1- Plain Logistic Regression classifier: With this model we got a model that's 71% accurate at identifying both legitimate and fraudulent transactions.

The model has an f1 score of 71

Approach 2- : Xgb classifier: With this model we got equivalent results to the others, it was 70% accurate at identifying both legitimate and fraudulent transactions. We used grid search for hyper parameter tuning

We also tried using random search, but it gave comparable results at 68%.

Approach 3 - 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 4- Ridge regression and ridge search: With this model we got comparable results to the first two, it was 70% accurate at identifying both legitimate and fraudulent transactions.

The model has an f1 score of 70.

Approach 5- CatBoost model: This model gave similar results to the decision tree moddel with an accuracy of 50% and f1 score of 50 also.

From our trials the most efficient models were the xgb classifier using grid search and the plain regression classifier both having accuracies of 70% and 71% and f1 scores of 70 and 71 each.

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 rate of 71% in identifying both legitimate and fraudulent transactions. Further refinements and enhancements can be explored to improve model performance and adapt to evolving fraud patterns.