CREDIT CARD FRAUD DETECTION

Credit cards are very commonly used banking instruments and fraud in this area is taking different forms too despite various strigent methods implemented by Banks. Hence,the current case study dealing with Credit card fraud detection helps to uncover probable fraudulent attempts beforehand and thus avoiding huge financial and reputational loss for the financial institutions.

In the current case study, the dataset consists of credit card transactions spread across two days. This data is highly imbalanced with around 90% of data belonging to non-fraudulent cases.This is one of the challenging tasks to handle without which the prediction results will be inaccurate.

The very first step in the case study would be to load the data and understand the features. One important point to note here is that the dataset is already transformed using PCA (Principal component Analysis) which resulted in the Gaussian variables. On the PCA transformed features, a through exploratory data analysis will be conducted including univariate and bivariate analysis. The data will then be checked for skewness and distribution of classes across different variables and handled accordingly.

Secondly, as we saw that the data is highly imbalanced, techniques like SMOTE (Synthetic Minority Oversampling Technique) and Adaptive Synthetic Sampling (ADASYN) will be used to balance the data. As SMOTE uses K nearest neighbours to create random synthetic samples, experiments will be performed to identify the correct value of k (an odd value) to avoid the problem of underfitting and overfitting.

Further, the data will be split into train and test to train the model and evaluate the performance on different subsets of data. Depending on the type of data, several classification models like Logistic Regression,SVM,Decision trees,Random forests will be built and performance will be compared using K-fold validation.

If the data is linearly separable, a logistic regression model will give the best result. Along with this, the model will be highly interpretable. On the other hand, models like Decision trees and random forest work well in scenarios where data classes are not linearly separable. As Decision Trees check the data in many ways, therefore one of its major disadvantages is that it can easily overfit. To avoid such scenarios of overfitting and underfitting, hypermeter tuning comes to rescue.

Thus,the models with best performance results will be built using the hyperparameter tuning of model parameters to reach the best AUC (Area Under Curve) for the classification problem. Example for SVM, parameters C and gamma will be tuned, and when using Random Forests, parameters like max_features, max_depth, min_samples_leaf and min_samples_split will be tuned individually to get the range of values where the model is performing better. Grid search and k-fold cross validations to obtain the range of hyperparameter value on which the model might perform well. As the data is highly imbalanced, stratified k-Fold Cross Validation will ensure that each fold is representative of all the strata of data.

Metrics like precision, recall, F1 score are usually used to measure the performance of classification models. In our case as we want to save the banks from high value fraudulent transactions, we will focus on high recall in order to detect actual fraudulent transactions.

****