Credit Card Fraud Detection Case Study

Problem Statement:

The objective of this case study is to identify the fraudulent transactions accurately and build a model to figure out better ways to reduce frauds in order to Increase the company's performance and thereby ensuring customer satisfaction.

The dataset includes credit card transactions made by European cardholders over a period of two days in September 2013. It is found that the dataset is highly imbalanced i.e. out of 2, 84,807 transactions, 492(0.172% of total transaction) were fraudulent.

Approach:

Since it is a classification problem on structured dataset, we decided to build classification models such as random forest, XGboost as they are highly performing models and don't consume much time to execute comparatively to other models.

Steps:

1. **Data Understanding:**
    - The dataset present were 2, 84,807 transactions with 31 columns. Apart from Time and amount all the other features (from V1 to V28) are obtained using PCA for privacy purpose.
    - There are no null values present in the dataset.
    - The dataset is highly imbalanced with 0.172% of fraudulent transactions out of the total transactions.

2. **Exploratory Data Analysis:**
    - Performed both univariate and bivariate analysis and checked if there is any skewness in the data.
    - SMOTE and ADASYN test to deal with the class imbalance.

3. **Feature Scaling:**
    - Normalized both the time and amount column within a particular range using standarscalar.

4. **Model Building:** Building all the models on the dataset would take huge amount of time and resources, so based on the data type, let's figure out the modeling technique to implement.
    - Logistic Regression: It works best on linearly separable data and which needs to be interpretable. But most of the data will have overlap between classes and so using logistic regression is not a good decision.
    - Decision Tree: The problem with decision trees is that we exactly don't know where to stop. This would lead to overfitting when the tree fits all the samples in the training data perfectly.
    - Random Forest: It works best for structured/uncorrelated data and could take care of overfitting. For our imbalanced data, this would give better result.
    - XGBoost: It is an extension of gradient boosting with parallel processing, better optimization and regularisation. And so it is more accurate and finds the best fit.

So, for the credit card fraud detection data type, choosing random forest and XG boost would be more appropriate.

5. **Model Comparison:**
    - for comparison, AUC-ROC curve is used to find out the performance of the model i.e. TPR is high and PFR is low(misclassifications are low)
    - Since, the model requires the analysis of fraudulent transactions in order to reduce to it, we would take precision into consideration and evaluate accordingly.