Credit Card Fraud
Problem understanding and Action plan
MADHAVAN RANGARAJAN (DS:C15)

Problem Statement: *To predict fraudulent credit card transactions with the help of machine learning models.*
The goal of the task is to classify the customer credit card transactions as fraudulent Vs non fraudulent.
Detecting and preventing fraud is a key responsibility of any financial institutions as its business performance depends on gaining and holding on to customers' trust. When fraud takes place customers not only lose their financial assets but also risk the exposure of sensitive information. In order to prevent this from happening, financial institutions track customers spending patterns in terms of the location in which they make their purchases, the average amount they spend, the platforms they usually make purchases on etc. They have systems in place where customers receive alerts of their card usage. However, fraudsters are persons that attempt to find ways around all the security systems. Among different ways of frauds, Skimming is the most common one, which is the way of duplicating of information located on the magnetic strip of the card. Apart from this, fraudsters manipulate and alter genuine cards/create counterfeit cards/steal cards.

Action plan
As a data scientist my goal is to build a supervised ML model that will help the financial institutions predict a transaction as fraud.

The data set for this project is of 150.0 MB size and has information of 2.85 million transactions of European card holders. The columns are components that have already been worked on with PCA, (to protect customer's privacy), hence the data is already transformed.

Pipeline to build a workable model:

Understand data: Extract the data, load and transform to get it ready to perform EDA

EDA: Perform extensive exploration to look for data types, data distribution and outliers. Since PCA has already been performed I am assuming missing data and the imputation have been taken care of, but we will find out when we check for it. Outlier detection will be one of the key indicators of fraud detection. Complete univariate and bivariate data analysis and resolve any skewness issues at this step.

Train-Test Split: Split the data and perform the suggested K-fold cross validation, specifically stratified K-fold as it guarantees "class ratio".
It's been suggested that we build a model before and after addressing data imbalance issues, which will help us understand the importance of balancing data and allow for the comparison of model performance.

Data imbalance: Several techniques can be used to address this issue but my intuition is that ADASYN will be most effective because of the following advantages it offers

- It lowers the bias introduced by the class imbalance.
- It adaptively shifts the classification decision boundary towards difficult examples

Model-building and hyperparameter tuning: Different models will be tried out including but not limited to random forest. The instructor suggests that we choose performing randomized grid search CV to validate the model performance this will be performed along with XGboost and other techniques.

Model Evaluation: I plan on using different model evaluation techniques to determine the metrics that reflect the business problem at hand. Since the data is imbalanced, performance metrics that depend on 0.5 as the threshold will not be effective. So, either precision or recall will be employed.