Problem Statement: Predict whether credit card transactions are fraudulent or not

Data Understanding: The Data usually contains transactions made by credit cards of cardholders. Such data often consists of both numerical data and categorical with noisy information. The dataset from Kaggle https://www.kaggle.com/mlg-ulb/creditcardfraud presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. Such dataset is highly imbalanced with more positive classes(fraud) and very less (0.17%) of negative classes. We also need to identify which features would be good for predictions and important relationships among them to develop a good model.

Data Preprocessing: If the features are continuous values, we can normalize the dataset. After exploring the data distribution, we can handle missing values and outliers. We can apply PCA to reduce noisiness in the dataset and reduce dimensions(features) for visualization. For unbalanced dataset, there are many approaches. The common approaches are oversampling or under-sampling. We can also identify relevant features using various techniques (like L2 Regularization). We can also create new features which can improve the model prediction.

Data Modeling: Once relevant features are identified, we can start with baseline model and then try different models to get better results. Logistic regression or Multi-layer perceptron models can be configured and trained using pre-processed data. We can split the training data into validation set and test set. We will do validation test on validation set. Based on the validation, we can hyper-tune the parameters configured with respect the models and predict on test set. We also need to handle prediction of global transactions which could be false positives on basis on specific scores or features.

Data Visualization: Projecting our test set predictions into a two-dimensional space, we can produce a scatter plot showing the clusters of fraudulent and non-fraudulent transactions. To show the patterns, we can plot a histogram of a customer's transaction features which shows maximum fraud cases. This will help the company to get a better idea whether to confirm/ discard the credit card transaction being fraudulent.