Credit Card Fraud Detection

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  - We are trying to create a solution for credit card fraud detection.
- What industry/realm/domain does this apply to?
  - This solution would primarily apply to the financial and banking industry but can also apply to the retail and consumer industry as well.
- What is the motivation behind your project?
  - The goal is to detect credit card fraud based on historical spending patterns. If a credit card company detects credit card fraud early and accurately, then their customers are more likely to trust the company and continue business with them.

Data Understanding

- What data will you collect?
  - We will be using secondary data to solve our business problem.
- Is there a plan for how to get the data?
  - We will be using a dataset that already exists and is available on Kaggle. , will download a copy for our use
    - Fraud Detection | Kaggle
- Are the features that will be used described clearly?
  - There is a data dictionary for this data set that clearly describes the features. Thus, we do have all the relevant features for our modelling

Data Preparation

- What kind of preprocessing steps do you foresee?
  - Data Cleaning/Anonymization: With the assumptions of Garbage in, Garbage out (GIGO),this step seeks to improve the quality and accuracy of the data by removing or correcting errors, duplicates, outliers, handling missing values and any other process to ensure better data quality. With our data containing PII, there will be a need for us to anonymize PII to avoid breaking the ethics and privacy laws surrounding PII data
  - Feature Scaling: To avoid certain features from dominating the ML process, this step will seek to bring features into similar scales. Common techniques we anticipate using will be the Standardization technique
  - Feature Encoding: Since most ML techniques operate on numerical data, we will use one-hot encoding to transform the categorical features into numerical representations
  - Train-Validation-Test Split: To ensure optimal results from our ML modelling, we will split our data set into separate sets into train, validation, and testing. The train data will be used to train the model, validation data will be use to tune the hyperparameters and evaluate model performance and the test data will be used for the final evaluation after model development
  - Handling Imbalanced Data: Our data is highly imbalanced and with most ML techniques configured by default to handle balanced data, we are more likely to get a misleading

result if the data Imbalanced is not handled. Hence, we will try different techniques such as SMOTE, ADASYN, etc. to address this issue

- What are some of the cleaning/pre-processing challenges for this data?
    - Data Anonymization: Our credit card contains sensitive and private information and anonymizing the data is crucial to ensure compliance to data protection regulations
    - Imbalanced Data: As common with most classification problems, the credit card data is highly imbalanced and hence we will have to find the optimal way to oversample/balance the data for modelling
    - Feature Engineering: The credit card data contains raw transactional amount, timestamp, codes, etc. Capturing relevant information from some of these features will be challenging, hence the need to apply aggregations or creating new features based on domain knowledge can be very useful

Modeling

- What modeling techniques are most appropriate for your problem?
    - We will be experimenting with different ML techniques to determine the most suitable one. Since this is a classification problem, we will be experimenting with:
        - Logistic Regression
        - Random Forest
        - Gradient Boosting
        - Support Vector Machine (SVM)
        - Ensemble Methods like Voting Classifier
- What is your target variable?
    - Is fraud or not fraud

- Is this a regression or a classification problem?
    - This is a Binary Classification Problem: classifying between one of 2 distinct classes or categories (Is fraud or not fraud)

Evaluation

- What metrics will you use to determine success?
    - For this business problem, the success of the model will be determined by
        - Precision
        - Recall
        - F1- Score
        - ROC- AUC Score
- What modeling algorithms are you planning to use?
    - We will be experimenting with different ML techniques to determine the most suitable one. Since this is a classification problem, we will be experimenting with:
        - Logistic Regression
        - Random Forest
        - Gradient Boosting
        - Support Vector Machine (SVM)
        - Ensemble Methods like Voting Classifier