Problem Statement

You are provided with a mock-up dataset that simulates a simplified fraud detection project you will encounter on the job on a daily basis.

The object is to develop a machine learning model to detect frauds using the information in the dataset. Each true positive benefits the companies by its transaction amount. Each false positive costs the company $200 because it takes time to investigate. There is no restriction on methodologies you can use. The success of your project will be evaluated based on the following criteria:

1. You need to perform data analysis on the raw data and present with at least one data visualization.
    - What insights can you derive from the raw data?
      We can infer that the number of negative classes are far higher (~95%) as compared to the positive classes (~5%). Average transaction amount is 110142.0. Overall, Wednesday has the highest number of transactions and also the highest count of fraud labels.

    - Why is the project worthwhile?
      This project is highly critical in the sense that it saves a huge deal of money for both the bank and it's customers by aiding in identifying and mitigating the fraudulent transactions.

    - What are the potential challenges you may be facing during the model development process?
      Some of the challenges include: imbalanced nature of the dataset, presence of missing values in multiple columns, presence of categorical columns which needs to be encoded.

2. Define proper performance metrics and provide business and statistical rationales of your selections.
    - Why are these metrics appropriate?
      Average transaction amount is 110142.0 which is way higher than the cost of reviewing the false positives. So, it is evident that the cost of false negative is much higher than cost of a false positive(200 Dollars to investigate). Recall is more important than precision for this analysis. We cannot use accuracy here because, the data is imbalanced. The accuracy will be 95 percent even if we predict everything as not-fraud.

3. Derive at least one new features from the data.
    - Why can these features potentially help your model?
      I derived quite a few new features for this analysis:
        i. ***Day/Month of the transaction: To look for patterns in fraudulent transaction as a function of day and month. Since, date as such doesn't add any value to our model***

ii. ***Days_bw_creation_and_transaction: This feature is the count of days between the account creation date and the transaction date. The idea was that, newer the account, higher could be the possibility of a fraudulent transaction which was observed in the analysis too.***
iii. ***Company and Industry: These two features were extracted from the beneficiary feature(which had 1531 unique values rendering it to be less important). We see that on an average, Retail, Health care, Entertainment industries have the highest fraud contribution. whereas gardening, agriculture and utilites have the least.***
iv. ***Customer tenure: This feature was created to inspect the relationship between tenure and fraud. It's measured as the number of days between current date and account opening date. But, given the dataset is old, it is unlikely to add value.***

4. Set up an experimental framework and perform hyperparameter optimization.
    - Does your design lead to a good approximation to the performance in production in the future?
      I've tried different sampling techniques and the design currently involves finding the best parameters through GridSearch of multiple classifiers. This is a good starting point for a Machine Learning problem. This can be further made better by stacking multiple classifiers, or finding the latent representations through deep learning and then leveraging a classifier to make predictions.
5. Perform a post-modeling analysis to convince the interviewers about the usability of your model.
    - What is the model performance?
      Model's performance on the test data are as follows:

      ```
      Test Precision: 0.6666666666666666
      Test Recall: 0.9259259259259259
      Test ROC AUC Score: 0.9547635533204573
      ```

    - What are the key features that drives the predictions?
      Based on the model, the top ten features driving this predictions are:

Feature	Importance
Days_bw_creation_and_transaction	0.325118
col_77	0.134765
col_7	0.078898
col_85	0.037604
col_9	0.033727
col_79	0.030102
col_110	0.027704
col_10	0.026047
col_90	0.025105
col_98	0.023761

- What threshold do you propose to use for the model and why?
  The following is the optimal threshold: 0.1167
  This was the best threshold for which the recall value is balanced with the appropriate false negative rate.