# CREDIT CARD FRAUD DETECTION SUMMARY

**VIDULA AROLKAR**

For many banks, retaining high profitable customers is the number one business goal. Banks need to be cautious about their customers' transactions, as they cannot afford to lose their customers' money to fraudsters. With the rise in digital payment channels, the number of fraudulent transactions is also increasing with new and different ways.

**PROBLEM STATEMENT:**

- The main objective of this case study is **to detect if the credit card transactions are fraudulent or not**.
- To determine how these frauds are affecting the bank's business.
- We would use Machine Learning models to achieve this.

**DATA:**

- The data set contains **2,84,807 total records**. Out of which **492 records are fraudulent**.
- The data collected is for the transactions done in 2 days.
- There are **28 Principal components** and only Numeric data due to Confidentiality.
- The other columns include **Time, Amount** and target column, **Class**.
- Amount is the transaction amount by the customer.
- Class depicts if the transaction was fraud (1) or not a fraud (0).
- As the data shows, there is high imbalance in the dataset (0.172% being fraudulent).

**PIPELINE FOR MODELLING:**

1. First step would always be Loading the data and check for Missing values if any.
2. Make a sanity check on the data (Outliers detection, Duplicate records, any other anomalies)
3. Exploratory Data Analysis:
    a) **Visualizing the data for Skewness.**
    b) **Check the spread of the data using Histograms, Scatterplots.**
    c) **Check outliers with Boxplots.**
    d) **Since the data is PCA transformed, we can assume that it is normalized.**
    e) **Univariate and Bivariate Analysis on the data to find correlations between the features.**
4. Once the data is clean and in proper format, we can split the data as Train and Test for Model building.
5. Since the data is highly imbalanced, some techniques like **SMOTE or Oversampling, under sampling** should be done for balancing.
6. First I would do modeling on the Imbalanced data, check for Accuracy and other metrics. This would make us understand how bad the imbalanced data is performing.
7. And then again we build the model for balanced dataset and check performance.
8. For better performance of the model, hyper parameters should be tuned for each of the model used.
9. For Hyper parameter tuning, K-fold cross validation or **Stratified K - fold** is better option to go for.
10. For Basic understanding, I would first go for **Logistic Regression and KNN model** (Simple Models).
11. After checking the performance of these 2 models, **Random Forest, XGBoost** would be the next models to train.
12. For Model evaluation, Accuracy may not be the best metric as the data is highly imbalanced.
13. **Precision and Recall** and **F1 Score** can be preferred as the performance metrics for each of the above mentioned models.
14. So we can say **FP (Transactions in real are not fraudulent, but predicted as frauds)** and **FN (Transactions are fraudulent in Real but predicted as non-fraudulent)** values should be as low as possible to reduce Type 1 and Type 2 errors.
15. TPR and FPR would give a better threshold value for Sensitivity – Specificity Tradeoff. **ROC curve** can be used for this.
16. To check which model is performing better, we would compare the performance metrics of each model with every other model. Benchmark would be Logistic Regression.