# Credit Card Fraud Analysis & Prediction

- Aditya P. Kulkarni - ID : 1330344

# Credit Frauds - Facts

- Billions of dollars of loss result from fraudulent credit card transactions annually
- Efficient fraud detection algorithms using advanced machine learning can help reduce these losses
- Fraud detection algorithm design is challenging due to non-stationary data distribution, imbalanced classes, and continuous transaction streams
- Public data is scarce for confidentiality reasons, making it difficult to determine the best approach
- The objective is to train a machine learning algorithm to predict fraudulent transactions using different sampling techniques to address class imbalance.

# Dataset - Overview

- The dataset includes transactions made by European cardholders in September 2013 with 492 frauds out of 284,807 transactions.
- We have 492 frauds out of 284,807 transactions in our dataset of transactions that took place over the course of two days.
- The dataset is very skewed, with frauds making up 0.172% of all transactions in the positive class.
- Unfortunately, due to confidentiality issues, we do not have access to the original features and more background information about the data.
- Link to dataset : https://drive.google.com/file/d/1n4VQFy6450LYOGvoYDJvQ_XTPkNyd5hP/view?usp=share_link

# Dataset - Metadata & Pre-processing

- **Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are ‘Time’ and ‘Amount’.**
- **Feature ‘Time’ contains the seconds elapsed between each transaction and the first transaction in the dataset.**
- **The feature ‘Amount’ is the transaction cost.**
- **Feature ‘Class’ is the response variable and it takes value 1 in case of fraud and 0 otherwise.**
- **Analyzed data for any missing transactions(values) using “*colSums(is.na(df))*”**
- **Checked if the columns Class is imbalanced using *“table(df$Class)”***

# Data Visualization - EDA

- Distribution of class labels.
- Distribution of time of transaction by class.
- Distribution of transaction amount by class.
- Feature correlation using pearson correlation method.

## Model Performance and top 3 features

| Model | Auc-Roc |
|---|---|
| XGBoost | 97.1% |
| Logistic regression | 97.1% |
| Random Forest | 97.7% |

| Feature | Percentage |
|---|---|
| V1 | 38% |
| V2 | 18% |
| V3 | 10% |

# Conclusion

- The project focuses on handling unbalanced datasets, like the fraud credit card transaction dataset where instances of fraudulent cases are few compared to normal transactions.
- Accuracy is not an appropriate measure of model performance for imbalanced data, so the metric AUC-ROC, and F1 score is used.
- Different methods of oversampling or undersampling the response variable are used to train the model better.
- The oversampling technique works best on the dataset, and significant improvement in model performance is achieved over imbalanced data.
- The Random Forest model achieved the best score of 0.977, but both random forest and logistic regression models performed well too.
- Tuning Random Forest model parameters can potentially improve performance further, but the project highlights the importance of effective sampling, modeling, and predicting with an imbalanced dataset.

- THANK YOU -