DEEPANSHU GOYAL

*27th August 2020*

# SUPERVISED LEARNING APPROACH TO PREDICT CREDIT CARD FRAUDS

*Overview of a supervised learning approach and performance of different classification algorithms on credit card dataset available on kaggle*

## DATA OVERVIEW

The dataset contains transactions made by credit cards in September 2013 by European cardholders and present transactions that occurred in two days, where out of 284,807 there are 492 fraudulent transactions. This dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. A typical case of imbalance data and one of the important consideration to apply supervised learning models

There are 30 features in the dataset, principal components are shared as features to avoid sharing any confidentional & personal information

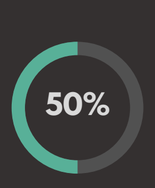
## SUPERVISED LEARNING APPROACH

### Exploratory Data Analysis

- There are no missing values in the dataset
- ~83% of the fraud transactions are less than £200
- ~90% of the genuine transactions are less than £200
- Fraud transactions are no different than genuine transactions and are of both high value and low value

### Train-Test Split (70-30%)

- Stratify splitting (Random Selection) with 70% data for training and 30% for testing



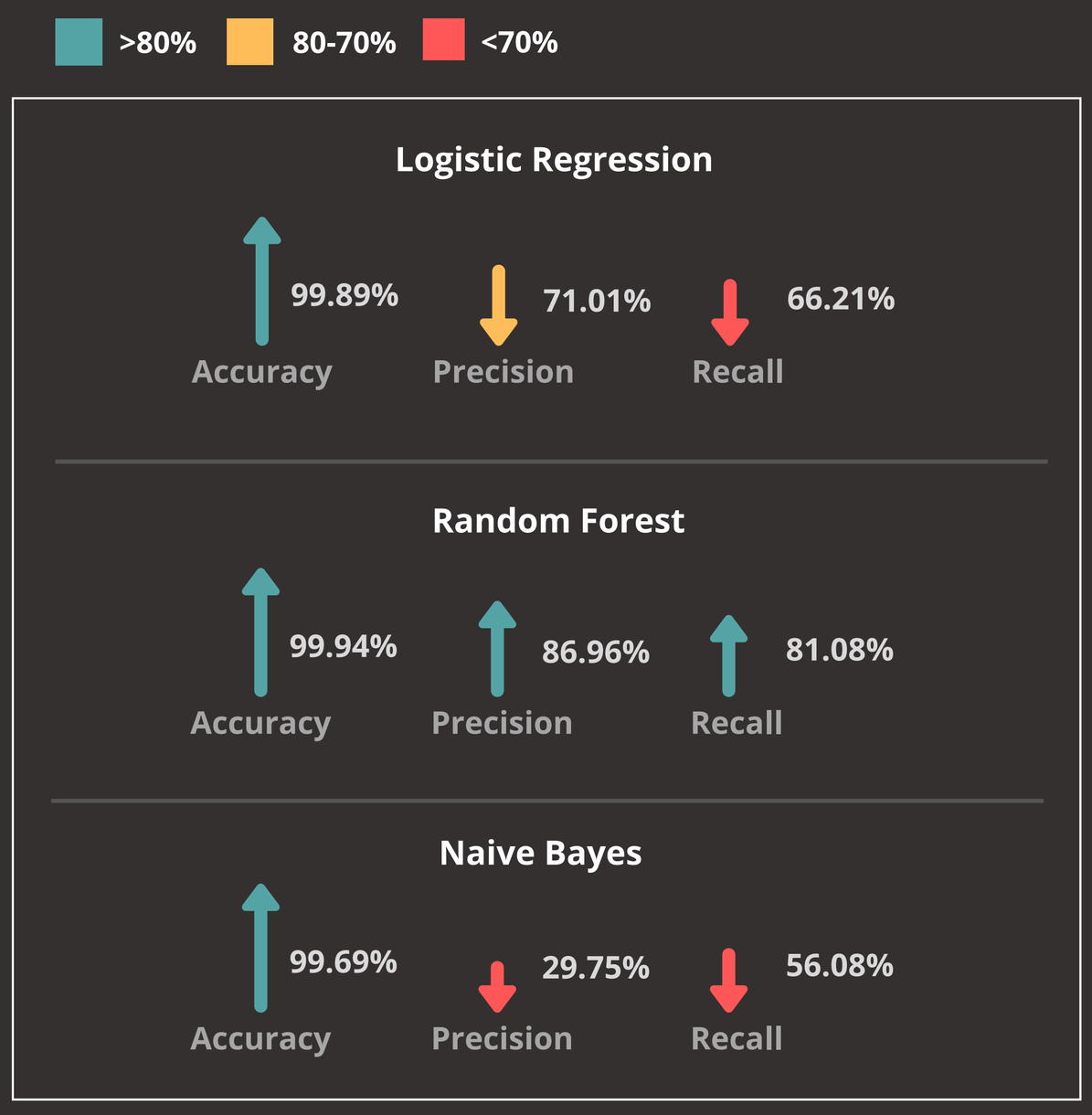

### Handle Imbalance Data

- Technique used
  - Random under-sampling
  - Random over-sampling
  - SMOTE over-sampling
  - SVMSMOTE over-sampling
  - Borderline SMOTE

### Modelling & Prediction

- Base accuracy is 99.83%. Selected model should have higher accuracy than this one
- ML Models used
  - Logistic Regression
  - Random Forest
  - Naive Bayes

### Classification Algorithm Performance & Comparison




- ***Random Forest with SMOTE oversampling is a clear winner***
- ***Considering a cost of £100 for a fraudulent transaction, company can save an average of £2M a year. (provided similar type of frauds happen as given in the training data)***

Datasource - https://www.kaggle.com/mlg-ulb/creditcardfraud