Credit Card Fraud Detection

MR. ASHUTOSH .P. GAIKWAD

INTRODUCTION

Credit card fraud has emerged as major problem in the electronic payment sector. In this survey, we study data-driven credit card fraud detection particularities and several machine learning methods to address each of its intricate challenges with the goal to identify fraudulent transactions that have been issued illegitimately on behalf of the rightful card owner. In particular, we first characterize a typical credit card detection task: the dataset and its attributes, the metric choice along with some methods to handle such unbalanced datasets. These questions are the entry point of every credit card fraud detection problem.

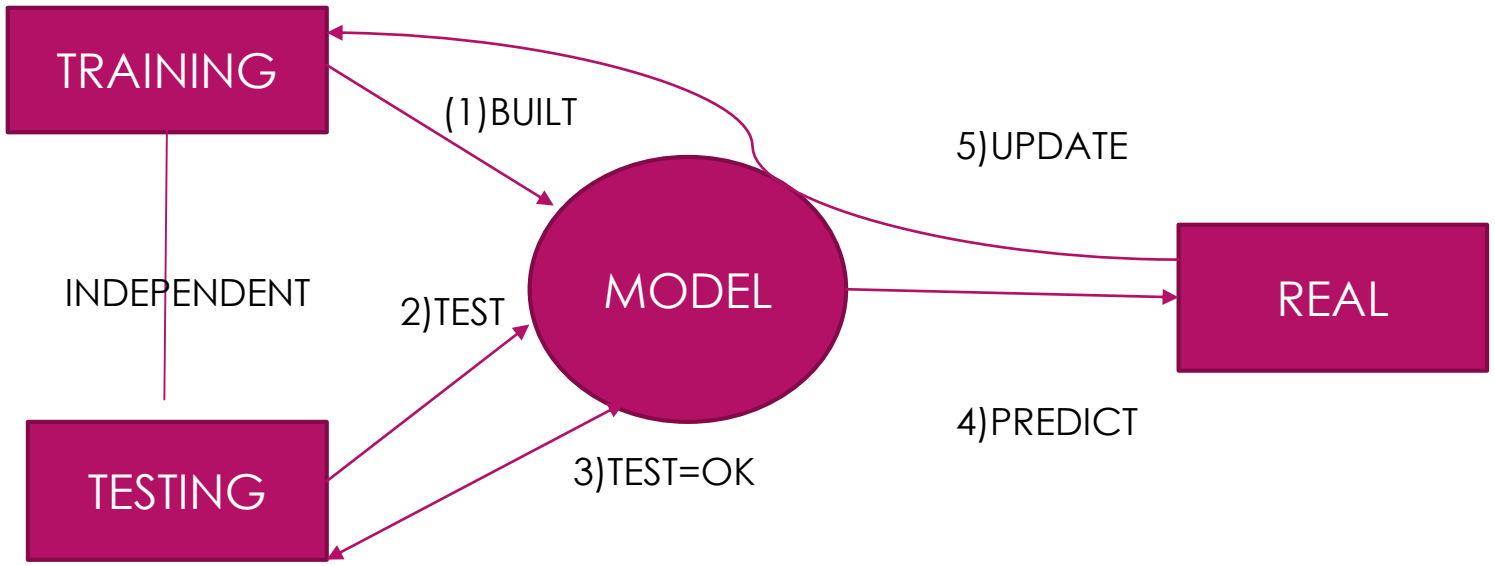
PROJECT OBJECTIVES

- To predict Credit Card fraud
- Highlighting the main variables/factors influencing Credit Card Fraud.
- Use various ML algorithms to build prediction models, evaluate the accuracy and performance of these models.
- Finding out the best model for our Credit Card Fraud Detection case & providing executive summary.

DATASET DESCRIPTION

- Source dataset is in csv format.
- Dataset contains 284807 rows and 31 columns.
- There is no missing values for the provided input dataset.
- Class is the variable which notifies whether a particular customer is fraudulent or not. And we will be developing our models to predict best Outcome.

FRAUD DETECTION PREDICTION MODEL



METHODOLOGIES

- EDA(Exploratory Data Analysis):There is no missing for the provided input dataset. All the data in dataset are categorical type and also Standard Scaled.
- Model building which includes defining the purpose if model, determine the model boundary, build the model, create an interface and export the model.
- Evaluating machine learning algorithm is an essential part of project.

EXPOLATORY DATA ANALYSIS(EDA)

- Data visualization using seaborn and matplotlib
- Exploratory data analysis (EDA) is an approach to analyze data sets & to summarize their main characteristics, often with visual methods.
- A Statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis.

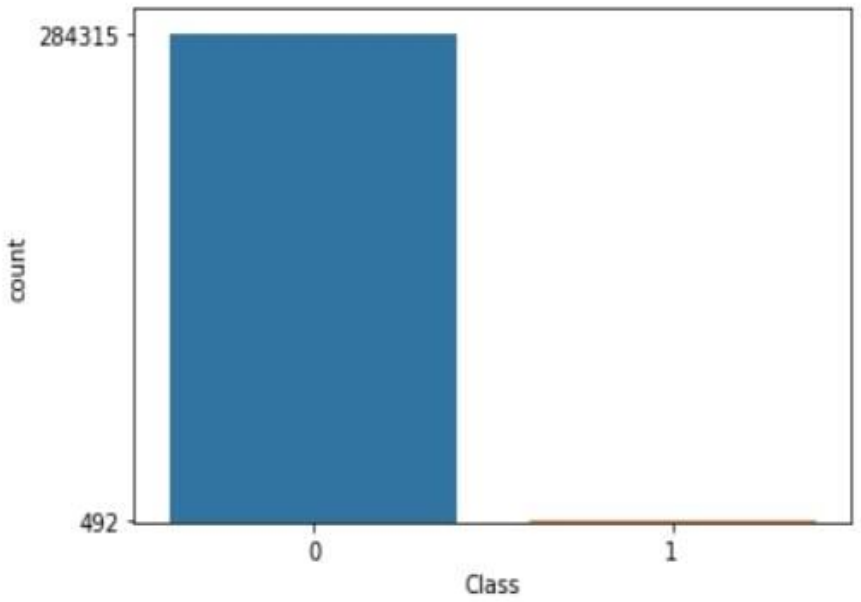
LIBRARIES & PACKAGES USED FOR DATASET

In [1]:
```
# importing all required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
# import all required packages
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix,classification_report
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import LinearSVC
from sklearn.svm import SVC
```

BAR GRAPH

```
In [163]: sns.countplot(data=df,x="Class")
          c=df["Class"].value_counts()
          plt.yticks(c)
          plt.show()
```



Bar Graph shows that the user number of users are getting fraudulent or not.

0 indicates that no fraud happened in credit card and 1 indicates there fraud happened in credit card.

SAMPLING TECHNIQUES

- Sampling techniques works on classification algorithm.
- To handle imbalance data in dataset we use sampling techniques
- There two types of sampling techniques under sampling and over sampling
- If you have basically reducing the majority class that is known as under sampling
- if you are increasing the minority class that is known as over sampling .

BEST SAMPLING TECHNIQUE FOR CREDIT CARD DATASET

After performing both the techniques on credit card dataset we have concluded that Random Over Sampling techniques is best for model. Random Oversampling includes selecting random examples from the minority class with replacement and supplementing the training data with multiple copies of this instance. In these example of credit card fraud detection number of fraudulent are less so instead of decreasing number of data by under sampling we use over sampler to get best possible result for these model

ENSEMBLING TECHNIQUES

Ensemble models in machine learning operate on various combination of the decisions from multiple models to improve the overall performance.

There are different types of Ensembling Technique :

1. Naive Aggregation method

a. Hard Voting b. Soft Voting

2. Bootstrapping

a. Bagging b. Pasting

3. Boosting Technique

a. ADA Boost b. Gradient Boost c. Extreme Gradient Boost (XGBoost)

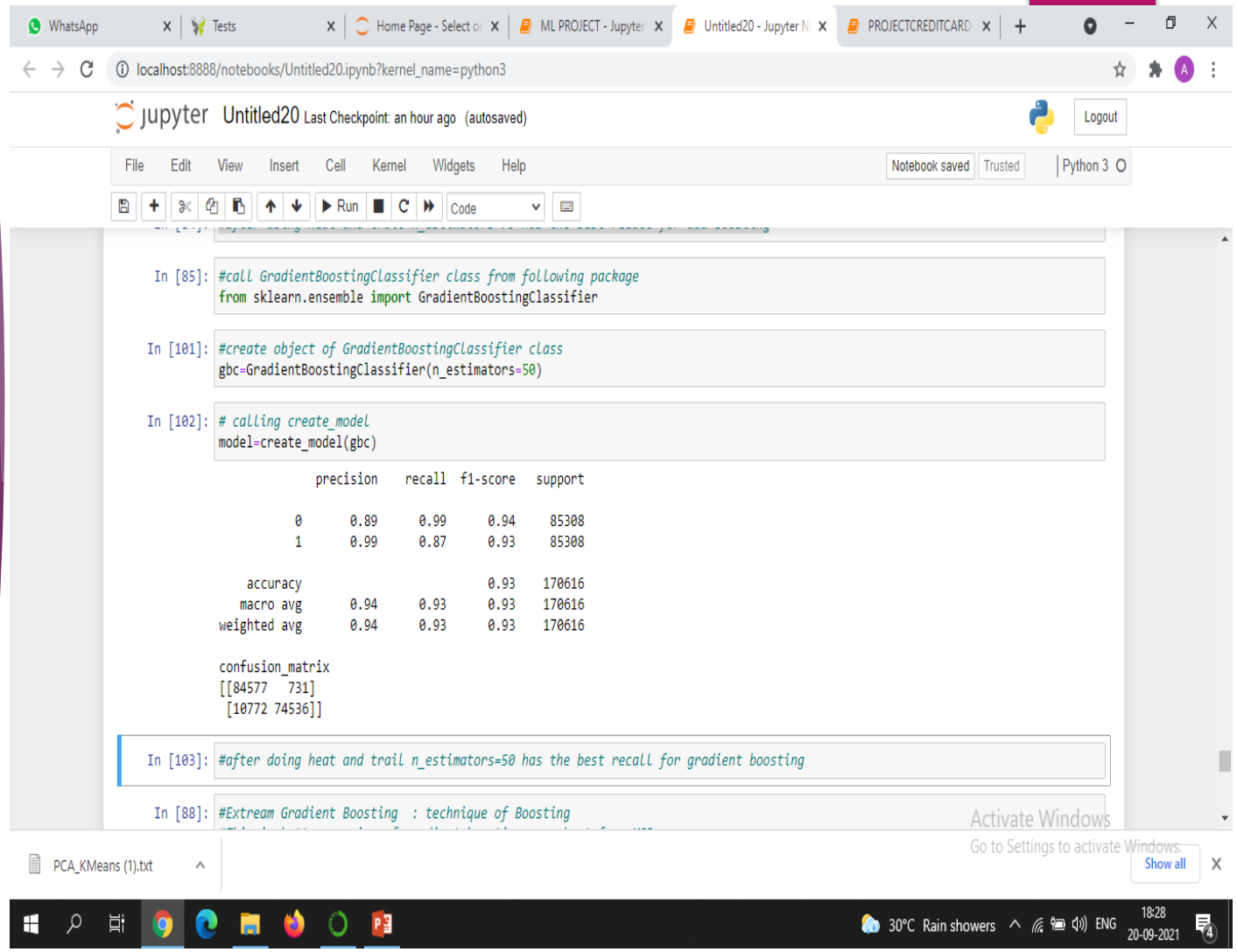
4. Stacking

ACCURACY OF VARIOUS MODELS AFTER ENSEMBLING

MODEL	ACCURACY
ADA BOOSTING	93%
PASTING(LOGISTICREGRESSION)	92%
GRADIENT BOOSTING	93%

BEST ENSEMBLING TECHNIQUE RESULT

After applying all the emsembling techniques we have concluded that GRADIENT BOOST is best ensembling techniques for this dataset



METRICS EVALUATION:

CONFUSION MATRIX

84577	**731**
10772	74536

EXPERIMENTAL RESULTS AND DISCUSSION

Fraud is considered as a positive class and legal as negative class and hence the meaning of the terms TP, TN, FP and FN are defined as follows:

- True Positive (TP) = Number of fraud transactions predicted as fraud
- True Negative (TN) = Number of legal transactions predicted as legal
- False Positive (FP) = Number of legal transactions predicted as fraud
- False Negative (FN) = Number of fraud transactions predicted as legal

CONCLUSION

After apply all the techniques such as EDA ,SAMPLING AND ENSEMBLING we have concluded that GRADIENT BOOST is best techniques with best recall and accuracy then other ensembling techniques. We can conclude that as the technology is developing day by day there are also fraudsters developing. Hence it is everyone's responsible to update about the technology and use it in a correct way. We should know about the Do's and Don'ts about the credit card before we start to use it and act accordingly to avoid any serious issues.