



# PREDICTING THE “FRAUD IN AUTO INSURANCE CLAIMS” & PATTERN EXTRACTION

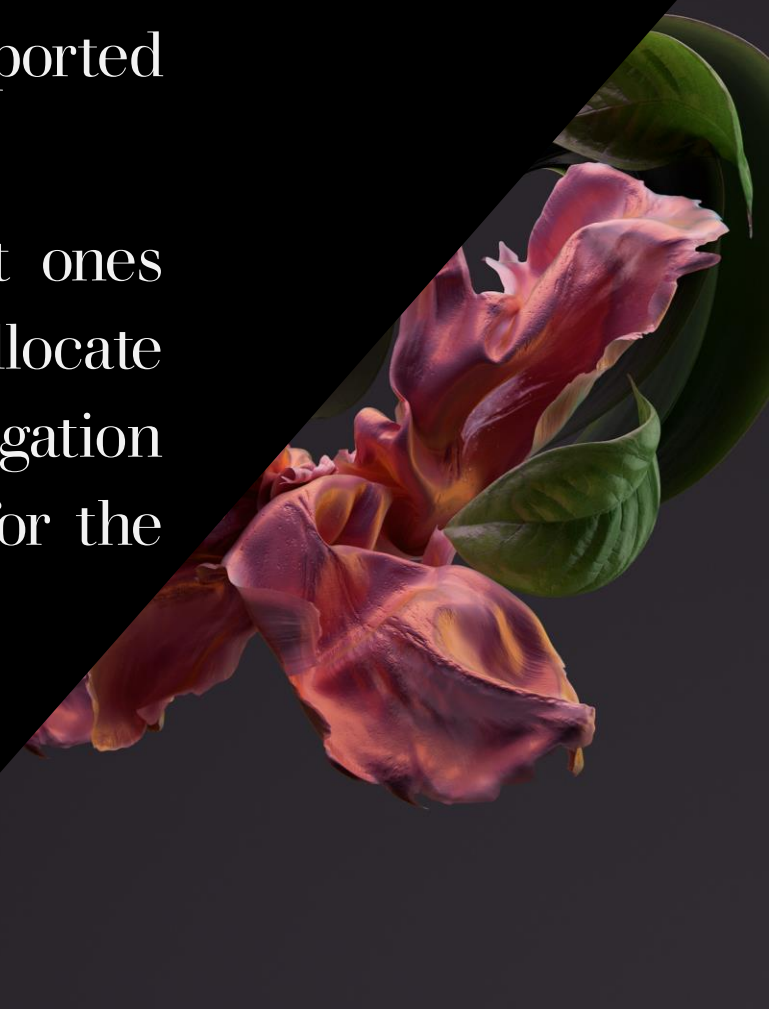
---

Presented by:-

Nischal.B.S



# Problem statement:

- A major general insurance company has a business problem with significant number of claims being reported are fraudulent in nature and it is leading to leakages.
  - So, the Insurer decided to predict the fraudulent ones before even processing the claims to allocate costs appropriately, to keep the thorough investigation process in place and to design proper action plan for the claims
- 

# AGENDA

- Create an analytical and modelling framework to predict the fraud in auto insurance claims based on the demographic, policy, claim, and vehicle related features provided in the datasets and also generate the top 20 patterns for fraud on target attribute

# About the Datasets

Demographics Data : Consists of Personal information related to the customer. Country, InsuredAge, InsuredGender, InsuredEducationLevel

Policy Information : These files consist of the customer auto insurance policy information, connected to the claim with the insurance company

Claim Information : These files consist of the details about the insurance claim, that the customer applied.

Data of Vehicle: These files consist of the details about the Vehicle, connected to the policy.

Fraud Data : Fraud is committed or not.



# Data Preparation

- Read the data set which is in the csv format and merge them

```
=pd.read_csv()
```

```
=pd.merge(result_,Test_Policy, on="CustomerID")
```

- Extracting the features from VehicleAttribute –

VehicleModel,

VehicleYOM,

VehicleID

By completing the above two steps we have the train and test dataset

---



# Missing value Treatment

Many attributes have missing values which are represented by various notations-

--- “???” , “?” , “-5” , “NA” , “MISSINGVAL” , “-1”

Classify if the attribute with the missing value is categorical or numerical  
- Numerical

Impute the numerical column which is Normally distributed with median value and Non-Normal distribution with mean.

-Categorical

Impute the Categorical\_\_\_\_\_column with mode.

---

## Creating Dummy variables with Nominal Attributes and Label encoding Attributes that are ordinal

- Nominal Attributes-

'TypeOfIncident', 'TypeOfCollission', 'SeverityOfIncident',  
'AuthoritiesContacted', 'IncidentState', 'IncidentCity',  
'PropertyDamage', 'Witnesses',  
'InsuredOccupation', 'InsuredHobbies',  
, 'InsurancePolicyState',  
'Policy\_CombinedSingleLimit',  
'PolicyAnnualPremium', 'VehicleModel',  
'VehicleYOM'

- Ordinal Attributes :

'PoliceReport', 'InsuredGender', 'ReportedFraud'

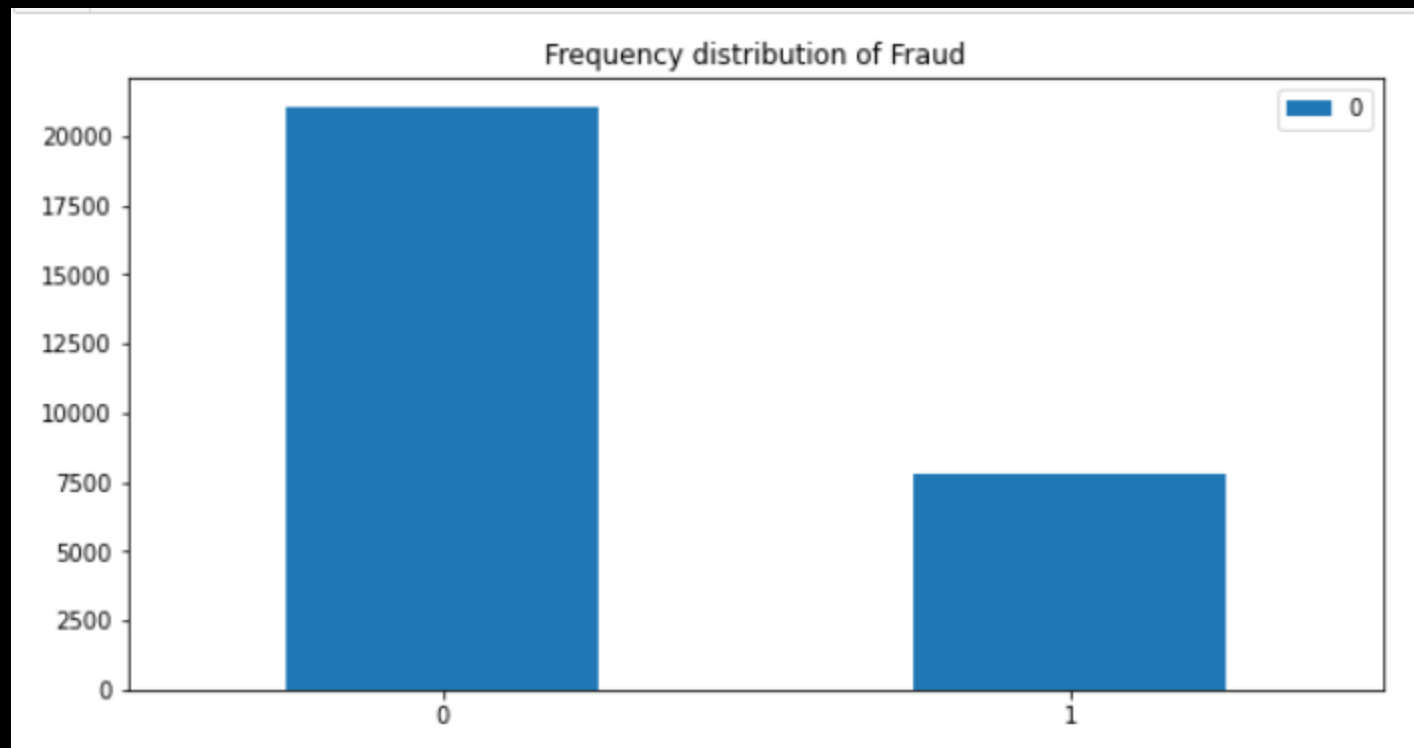
## Check for correlation and colinearity

Heat Map- To drop columns which are highly correlated.  
Any value with above 60% was dropped.

Variance inflation factor- to check for co linearity .Any value above 10 was dropped



# Over Sampling to remove unbalanced dataset



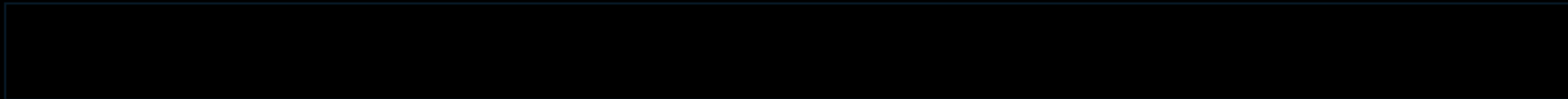
## Standardizing the numerical columns (knn,svm,NB)

Standardization is the process of bringing all the columns in the dataset to the same scale .

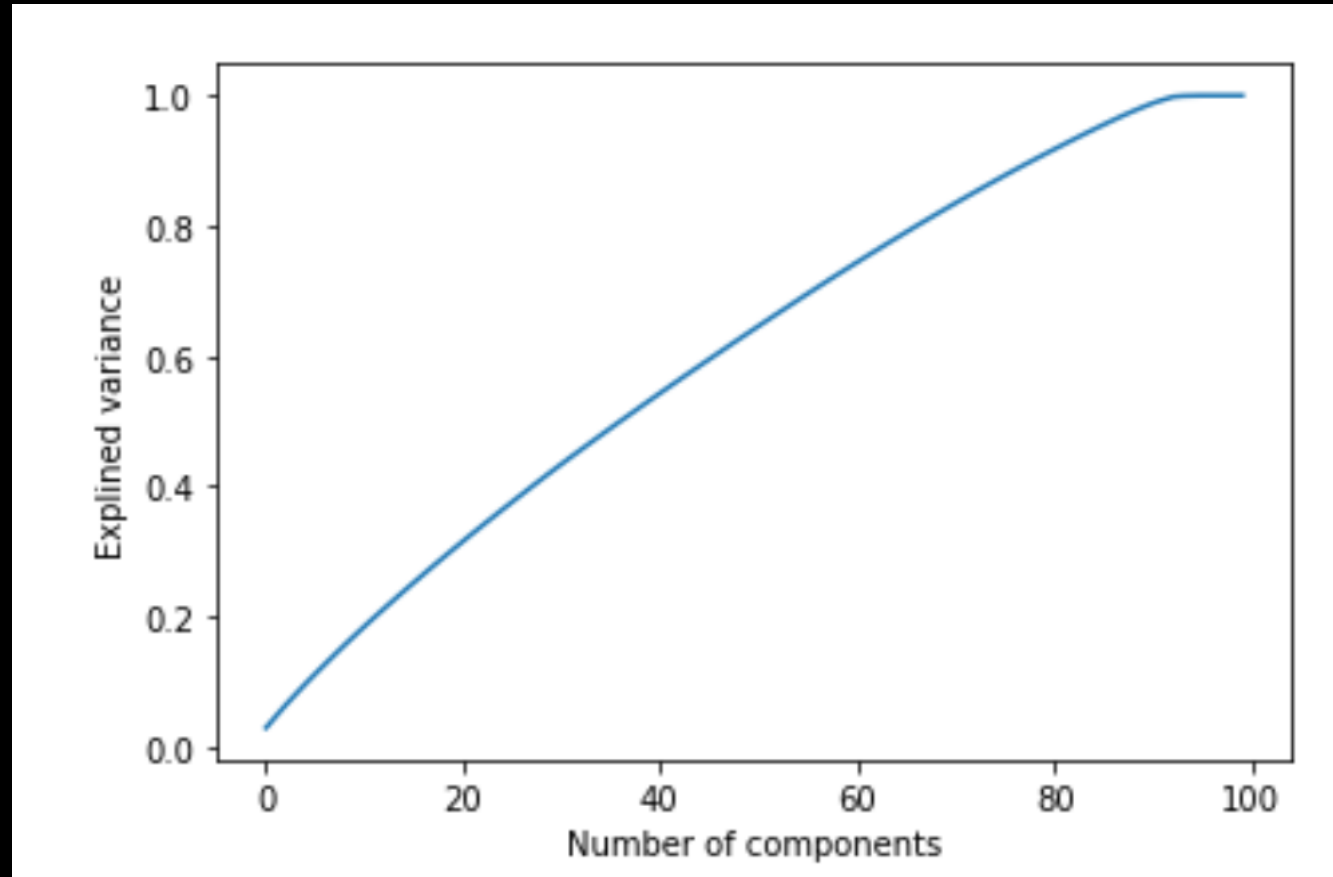
This is done by taking the mean as zero and calculating the deviation (spread) from Zero for all the numerical columns.

The columns we more large range is given more weights to bring it on the same scale of other columns with less range.

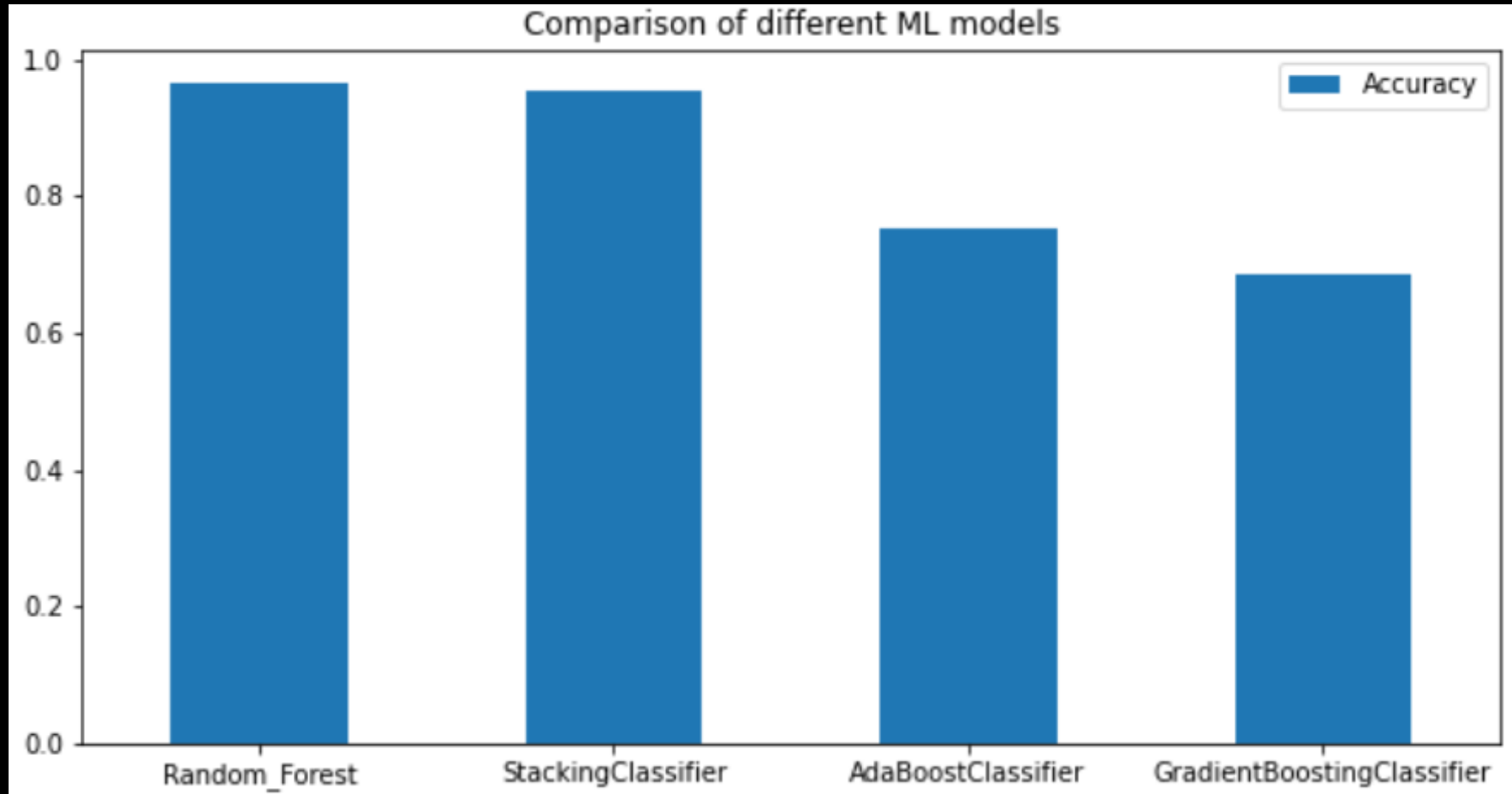
Standardization is done after splitting the data to train test split



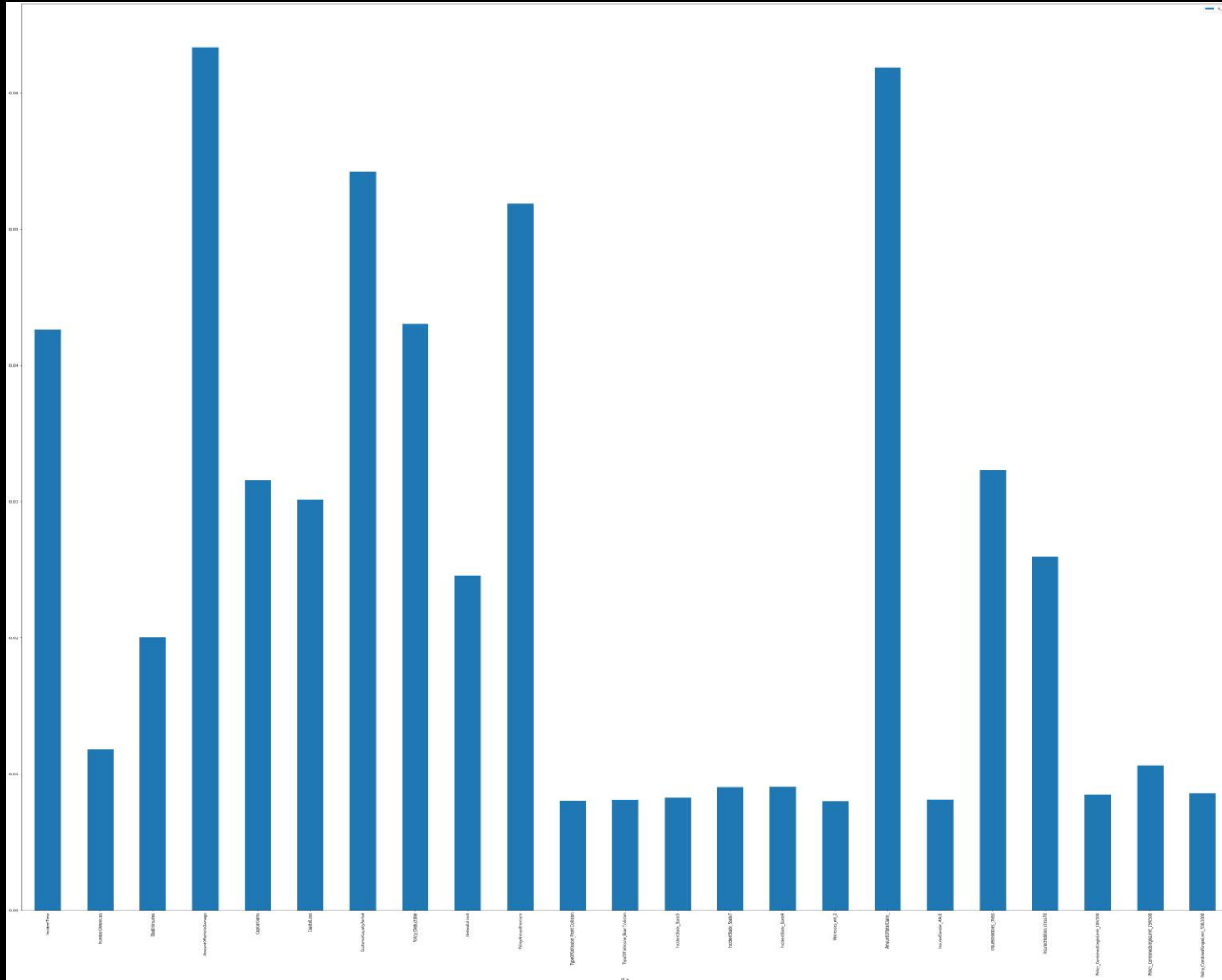
# Applying PCA to check for Dimensionality reduction



# Building Classification Models

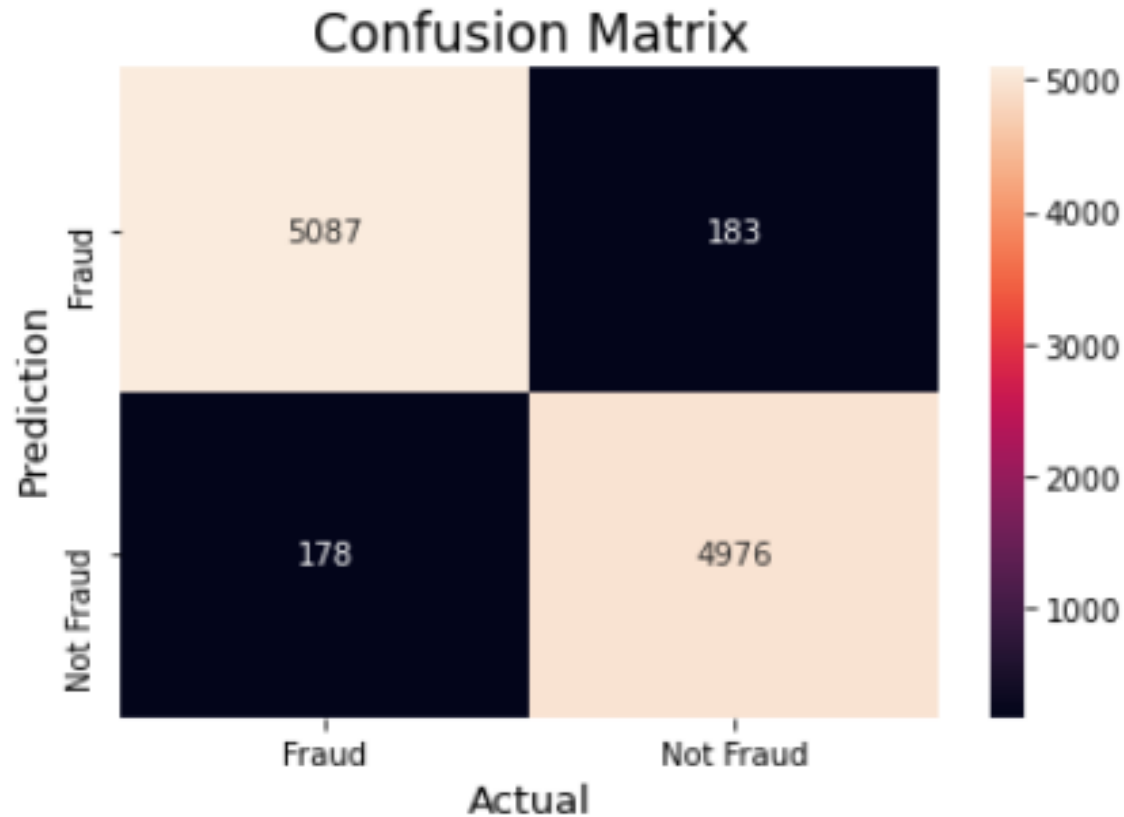


# Feature importance



|     | o_x                                | o_y      |
|-----|------------------------------------|----------|
| 0   | IncidentTime                       | 0.040785 |
| 1   | NumberOfVehicles                   | 0.013024 |
| 2   | BodilyInjuries                     | 0.018328 |
| 3   | AmountOfVehicleDamage              | 0.059694 |
| 4   | CapitalGains                       | 0.031036 |
| 5   | CapitalLoss                        | 0.029731 |
| 6   | CustomerLoyaltyPeriod              | 0.054508 |
| 7   | Policy_Deductible                  | 0.039963 |
| 8   | UmbrellaLimit                      | 0.024565 |
| 9   | PolicyAnnualPremium                | 0.055601 |
| 41  | Witnesses_wit_0                    | 0.009736 |
| 47  | AmountOfTotalClaim_                | 0.058830 |
| 90  | InsuredHobbies_chess               | 0.037992 |
| 91  | InsuredHobbies_cross-fit           | 0.029845 |
| 110 | Policy_CombinedSingleLimit_250/500 | 0.009122 |

# EVALUATION METRICS



|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.97      | 0.97   | 0.97     | 5270    |
| 1            | 0.96      | 0.97   | 0.96     | 5154    |
| accuracy     |           |        | 0.97     | 10424   |
| macro avg    | 0.97      | 0.97   | 0.97     | 10424   |
| weighted avg | 0.97      | 0.97   | 0.97     | 10424   |

THANK YOU !!!