KBC NORTH MAHARASHTRA UNIVERSITY, JALGOAN
Z.B. PATIL COLLEGE, DHULE
Department of STATISTICS

"AUTO INSURANCE FRAUD DETECTION"

BY
DYNANESHWAR KOTE
AJAY PATIL
SACHIN WANKHEDE

Guided By
MS. PRATIKSHA
RUIKAR MAM

INTRODUCTION

- An improper activity committed by individuals in order
  to gain benefit.
- There are various types of frauds viz. Health care, Agricultural frauds but we are focusing on Auto Insurance Fraud Detection.

What is Auto Insurance Fraud Detection

- The insurance industry is concerned with the detection of fraudulent behavior with insurance company due to vehicles . The number of automobile claims involving some kind of suspicious circumstance is high and has become a subject of major interest for companies . By building a classification model auto insurance fraud can be detected.

Need of Auto insurance fraud detection

- India is one of the biggest market for insurance industries all over the world, yet it is not free from risks.

- Indian Insurance Industry looses around $6 billion every year to this insurance frauds.

- Hence there is an urgent need to develop a capability which can help companies identify whether the given insurance claim is fraud or genuine with high degree of accuracy and with less amount of time.

- This will also help in maintaining the customers satisfaction and also the trust towards the insurance company.

OBJECTIVE

- To minimize number of fraud claim cases.
- To build a classification methodology to determine whether a customer is placing a fraudulent insurance claim or not .
- To provide quickness & high accuracy for claiming process.
- To reduce the amount of financial loss of company due to such illegals frauds.

Methodolog y

- We use machine learning & their algorithm using python.
- The data used for this study is secondary data. It is extracted and compiled from Kaggle website
- The data is then preprocessed and after training , the data is modeled using Xgboost classifier

  and we predict given claim is fraud

Importing Libraries

- In this step we import all necessary libraries required in our project

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Load the Dataset

- After importing all our required libraries then we load the Dataset.
- The Dataset we used in this project is a publicly available dataset taken from Kaggle

```
data = pd.read_csv("insuranceFraud.csv")
```

Basic Operation on Dataset.

- Here we perform some basic operation on our dataset to check whether our dataset working

```
data.head()
```

```
data.info()
```

```
data.describe()
```

```
data.isnull().sum()
```

Checking the Null values

```
data.isna().sum()
```

```
months_as_customer               0
age                              0
policy_csl                       0
policy_deductable                0
policy_annual_premium            0
umbrella_limit                   0
insured_sex                      0
insured_education_level          0
insured_occupation               0
insured_relationship             0
capital-gains                    0
capital-loss                     0
incident_type                    0
collision_type                 178
incident_severity                0
authorities_contacted            0
incident_hour_of_the_day         0
number_of_vehicles_involved      0
property_damage                360
bodily_injuries                  0
witnesses                        0
police_report_available        343
total_claim_amount               0
injury_claim                     0
property_claim                   0
vehicle_claim                    0
fraud_reported                   0
```

Data cleaning

- Cleaning missing values using categorical imputer

```python
from sklearn_pandas import CategoricalImputer
imputer = CategoricalImputer()
```

```python
data['collision_type']=imputer.fit_transform(data['collision_type'])
data['property_damage']=imputer.fit_transform(data['property_damage'])
data['police_report_available']=imputer.fit_transform(data['police_report_available'])
```

Extracting categorical data data

```
cat_data = data.select_dtypes("object").copy()
```

```
cat_data.head()
```

insured_sex	insured_education_level	insured_occupation	insured_relationship	incident_type	collision_type
MALE	MD	craft-repair	husband	Single Vehicle Collision	Side Collision
MALE	MD	machine-op-inspct	other-relative	Vehicle Theft	Rear Collision
FEMALE	PhD	sales	own-child	Multi-vehicle Collision	Rear Collision
FEMALE	PhD	armed-forces	unmarried	Single Vehicle Collision	Front Collision
MALE	Associate	sales	unmarried	Vehicle Theft	Rear Collision

Encoding

- In this step we perform label encoding on categorical variables in the dataset

```
cat_data["policy_csl"] = cat_data["policy_csl"].map({ '100/300':1,"250/500":2, '500/1000':3})
cat_data["insured_sex"] = cat_data["insured_sex"].map({ "FEMALE": 0 ,"MALE": 1})
cat_data["insured_education_level"] =cat_data["insured_education_level"].map({'JD' : 1, 'High School' : 2,
cat_data["incident_severity"] = cat_data["incident_severity"].map({"Trivial Damage":1 , "Minor Damage":2 ,
cat_data["property_damage"] = cat_data["property_damage"].map({"NO": 0 , "YES": 1})
cat_data["police_report_available"] = cat_data["police_report_available"].map({"NO":0 , "YES":1})
cat_data["fraud_reported"] = cat_data["fraud_reported"].map({"N":0 , "Y":1})
```

Catagorical Data after Encoding

policy_csl	insured_sex	insured_education_level	incident_severity	property_damage	police_report_available	fraud_reported
2	1	6	3	1	1	1
2	1	6	2	0	0	1
1	0	7	2	0	0	0
2	0	7	3	0	0	1
3	1	5	2	0	0	0

Combining categorical & numerical data

```
final_data =pd.concat([num_data,cat_data],axis=1)
```

```
final_data.head()
```

	months_as_customer	age	policy_deductable	umbrella_limit	capital-gains	capital-loss	incident_hour_of_the_day	number_of_vehicles_involved
0	328	48	1000	0	53300	0	5	1
1	228	42	2000	5000000	0	0	8	1
2	134	29	2000	5000000	35100	0	7	3
3	256	41	2000	6000000	48900	-62400	5	1
4	228	44	1000	6000000	66000	-46000	20	1

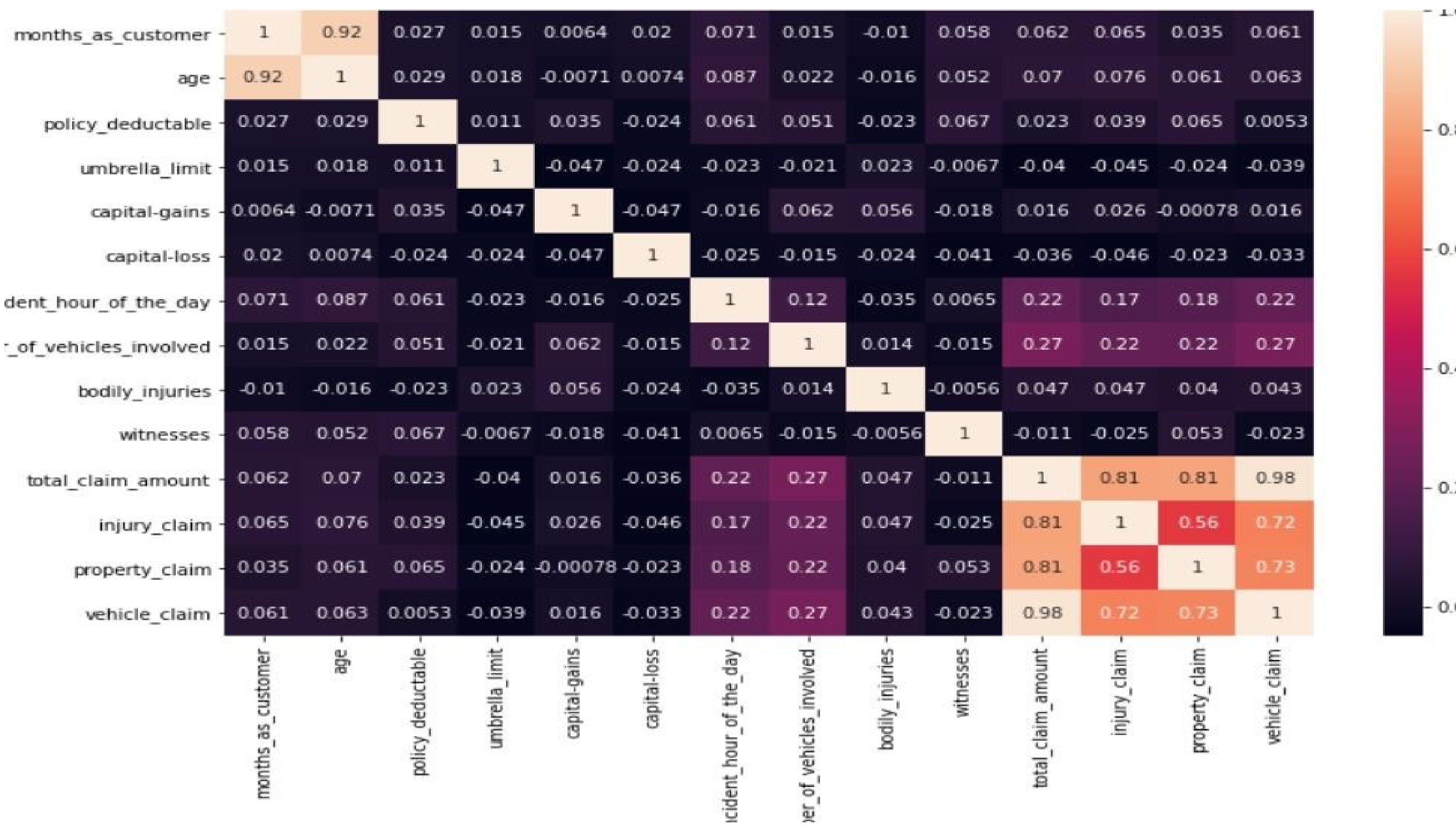
Separating Feature column and Target Column

```
#removing target column from feature column
x=final_data.drop("fraud_reported",axis= 1)
```

```
#making feature column
y=final_data["fraud_reported"]
```

Checking Multicollinearity using Heatmap

- Here we plot heatmap showing relation between the variables



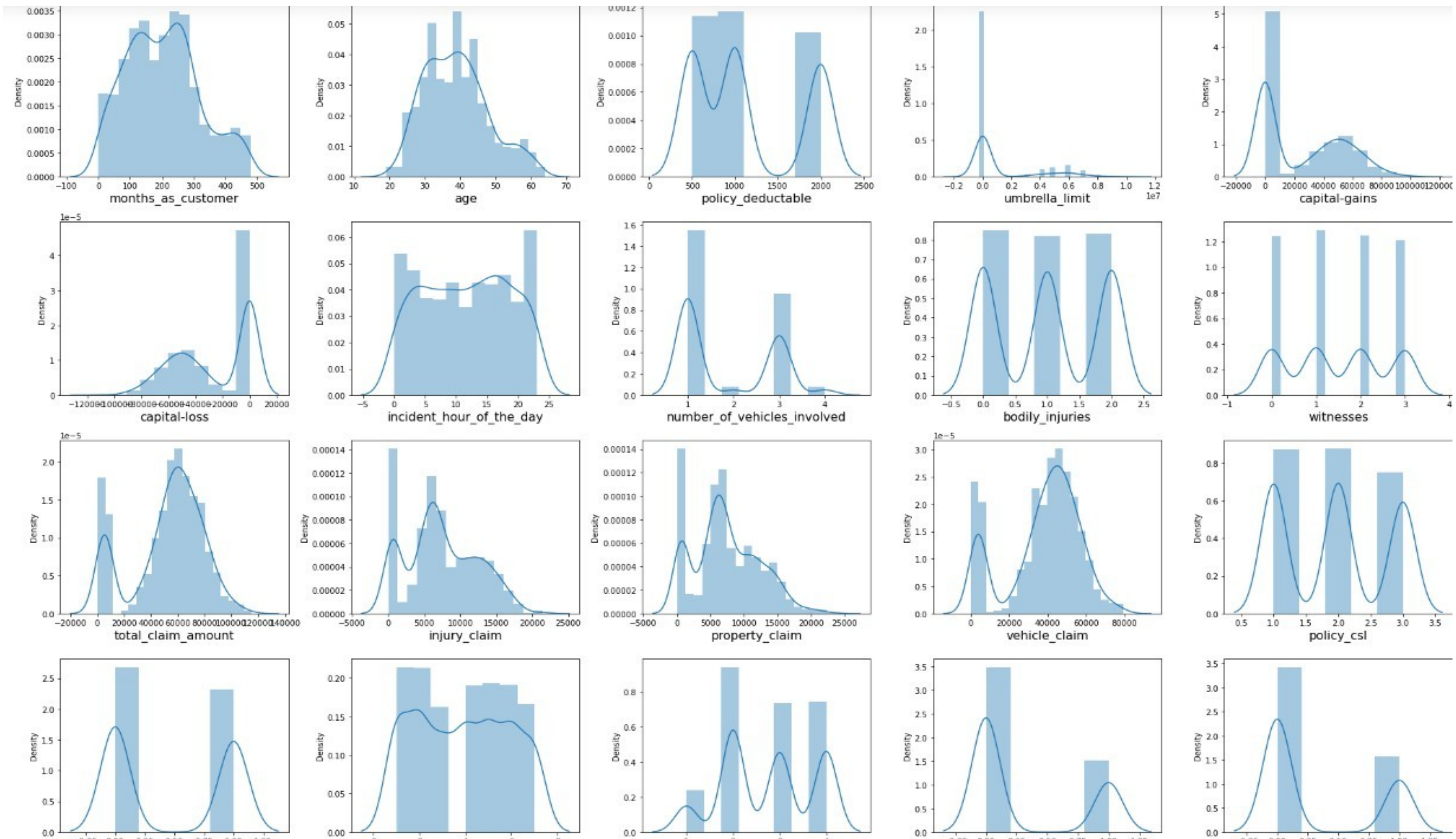
Removing highly correlated columns

- Here we remove age column and Total claim amount column

```
x.drop(columns=["age","total_claim_amount"],inplace = True)
```

Normalization

- Here our data is normally distributed



Standardisation

- Standardisation makes all variable to a common scale

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
```

Training & testing of model

- Here we split our dataset in train and test set. 75% of our dataset is for training purpose and 25 % is for testing .

```
from sklearn.model_selection import train_test_split
train_x,test_x,train_y,test_y = train_test_split(x,y,test_size=0.35)
```

Using Xgboost algorithm

- Here we use X gboost algorithm and train the model by 75% of the dataset

```
from xgboost import XGBClassifier
```

```
xgb=XGBClassifier()
```

```
y_pred = xgb.fit(train_x, train_y).predict(test_x)
```

Output

- Here we test the model by 25 % of the dataset .Where 0 represent fraud not happen and 1 represents fraud happen

```
y_pred
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0])
```

Conclusion

- Here we check all suitable algorithms for better accuracy and found that Xgboost has highest accuracy among all of them having 75% accuracy so we used xgboost algorithm for further predictions

```
ac2=accuracy_score(test_y,y_pred)
ac2
```

```
0.748
```

Future scope

- In this Project, we learned how machine learning can be applied to decide which claims are genuine and which claims are fraudulent . In future it saves time and money for dealing with fraudulent claims

Research paper

paper

- 1] Survey of Insurance Fraud Detection Using Data Mining Techniques H.Lookman Sithic, T.Balasubramanian
  - 2] Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection [Sharmila Subudhi, Suvasini Panigrahi]
- 3] Application of Clustering Methods to Health Insurance Fraud Detection Yi Peng1 , Gang Kou1, *, Alan Sabatka2 , Zhengxin Chen1 , Deepak Khazanchi1 ,Yong Shi1
- 4] CLAIMS AUDITING IN AUTOMOBILE INSURANCE: FRAUD DETECTION AND DETERRENCE OBJECTIVES Sharon Tennyson Pau Salsas-Forn
- 5] Big Data and Specific Analysis Methods for Insurance Fraud Detection Ana- Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA
- 6] Analytics for Insurance Fraud Detection: An Empirical Study Carol Anne Hargreaves* ,Vidyut Singhania* (Business Analytics)Institute of Systems Science, National University of Singapore, Singapore, Singapor

THANK YOU!