ANALYSIS OF MOTOR INSURANCE FRAUD

Capstone Project

<u>Group 16;</u>

a. Aarvinda Kumari Manur
b. Haritha Purushotham
c. Parag Mulik

Introduction

Problem Statement

- Insurance is the biggest domain in world and insurance fraud or (fraudulent claims) is second biggest crime in the World.
- Insurance industry counts 7,000 organizations in the US alone.
- Insurance scams cause $29 billion of damage to auto insurers annually.
- Roughly 85% of insurers have dedicated investigation teams.
- Once an insurance fraud happens, it doesn’t hurt just the insurance company, but it can affect ordinary consumers, too.
- An average American family spends $1,548.28 annually on car insurance.

Goal

- Increase ROI and reduce risk.
- To Predict fraud accuracy frequency.
- Root cause of the fraud.

Project Path

- Data Collection
- Data Cleaning and Exploratory Data Analysis
- Modelling
- Model Validation

Scope

- Random sampling data
- Sample size constraint

Current As is State

- All the claims submitted are checked manually.
- Dependency and non-visibility of the surveyor.
- Many handshakes

Future to be State

- Claims with tagged Fraud as No can be processed without Manual intervention.
- Increase in ROI and decrease in cost.
- Extensive process.

Project Roadmap

Project Methodology

01

Data Collection

02

Data Cleansing & Data exploration

03

Modelling

04

Model Validation

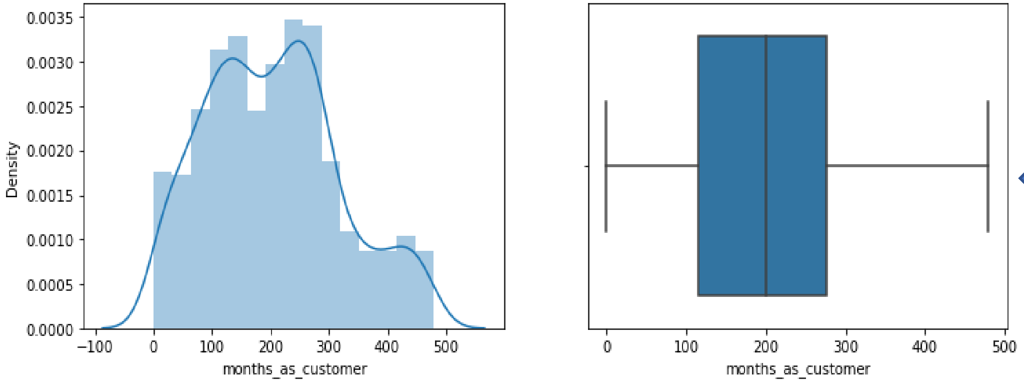
05

Model can be used

01& 02 : Data Collection, Data Cleaning and Exploratory Data Analysis

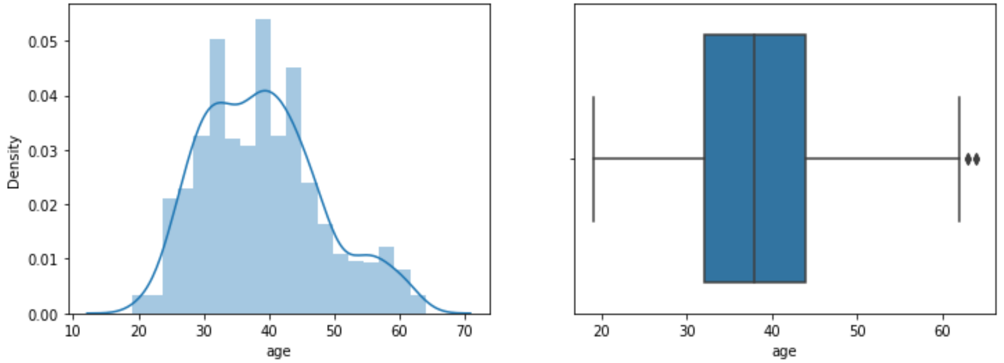
Process	What we Did ?	What we Learnt
Data Collection	Data Collection - Fetched the Data	From Kagal
Data Cleansing	Central Tendencies	We identified that the data needed cleaning
	Identified the Null Values	Replaced the Null Values by NaN
Exploratory Data Analysis	To identify the outliers	We did the skewness and kurtosis on Numerical Variables and identified that we don't have substantial outliers which can affect the analysis.
	Correlation between numerical variables	Identified the variables which are not highly correlated and removed the same.
	Identified the Categorical variables.	Converted the Categorical into Numerical Values
	Significant variables impacting my response variable	Selected the variables having high impact.

<AxesSubplot:xlabel='months_as_customer'>



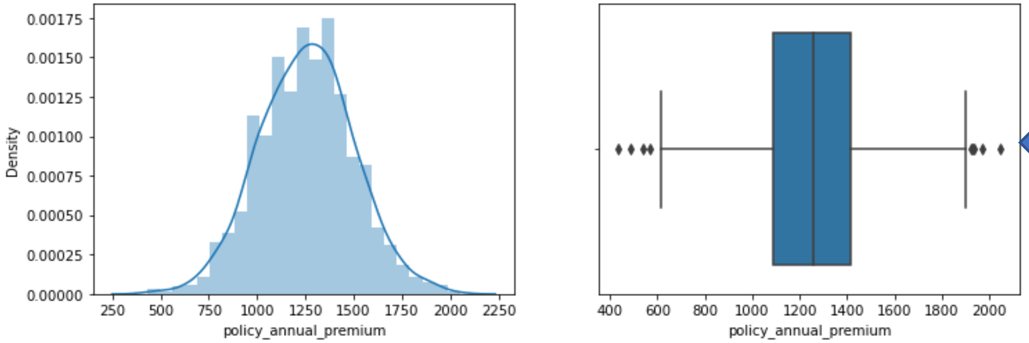
0.47898804709224163 -0.26025501504003934

<AxesSubplot:xlabel='age'>



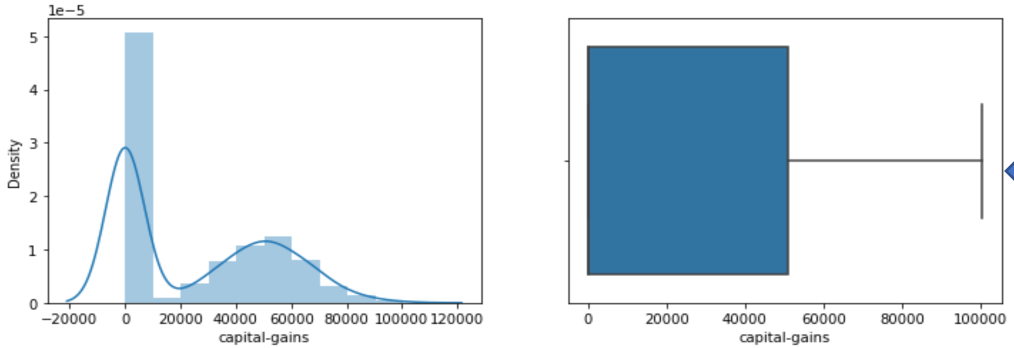
0.004401994526610823 0.0738894402077599

<AxesSubplot:xlabel='policy_annual_premium'>



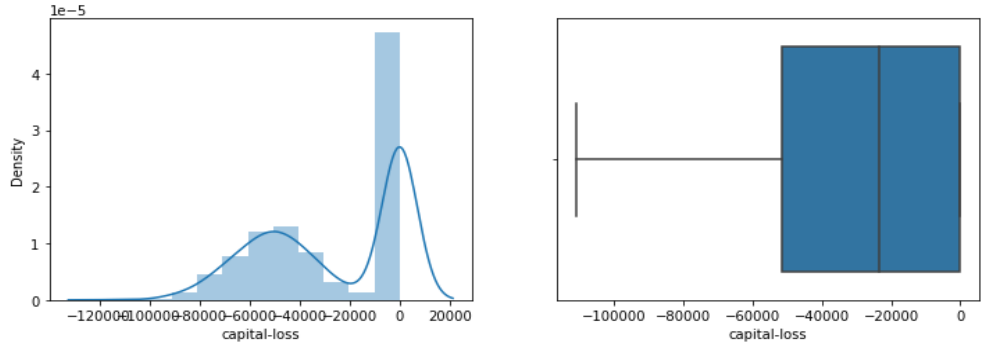
0.4788502295807907 -1.276703510816485

<AxesSubplot:xlabel='capital-gains'>



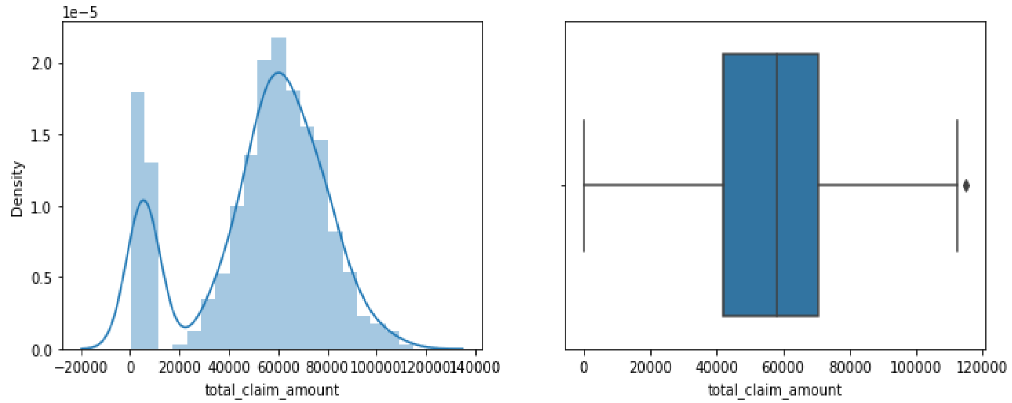
-0.39147194299389343 -1.3138745001493803

<AxesSubplot:xlabel='capital-loss'>



-0.594581988510234 -0.45408142669809326

<AxesSubplot:xlabel='total_claim_amount'>

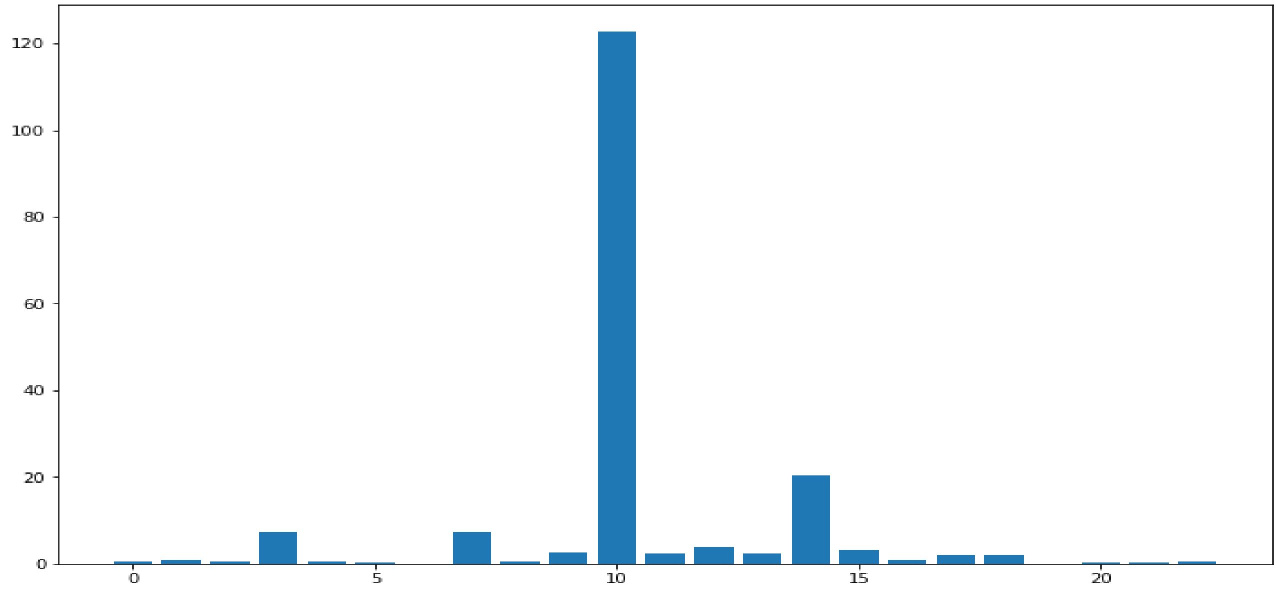


0.26481087847181833 -0.7630870610902973

Numerical Correlation

	months_as_customer	age	policy_annual_premium	capital-gains	capital-loss	total_claim_amount	injury_claim	property_claim	vehicle_claim	bind_to_incident
months_as_customer	1	0.922098	0.005018	0.006399	0.020209	0.062108	0.065329	0.03494	0.061013	0.047937
age	0.922098	1	0.014404	-0.007075	0.007368	0.069863	0.075522	0.060898	0.062588	0.036753
policy_annual_premium	0.005018	0.014404	1	-0.013738	0.023547	0.009094	-0.017633	-0.011654	0.020246	-0.001205
capital-gains	0.006399	-0.00708	-0.013738	1	-0.046904	0.01598	0.025934	-0.000779	0.015836	-0.042206
capital-loss	0.020209	0.007368	0.023547	-0.046904	1	-0.03606	-0.04606	-0.022863	-0.032665	0.027628
total_claim_amount	0.062108	0.069863	0.009094	0.01598	-0.03606	1	0.805025	0.810686	0.982773	-0.000765
injury_claim	0.065329	0.075522	-0.017633	0.025934	-0.04606	0.805025	1	0.563866	0.722878	-0.002476
property_claim	0.03494	0.060898	-0.011654	-0.000779	-0.022863	0.810686	0.563866	1	0.73209	-0.000447
vehicle_claim	0.061013	0.062588	0.020246	0.015836	-0.032665	0.982773	0.722878	0.73209	1	-0.000315
bind_to_incident	0.047937	0.036753	-0.001205	-0.042206	0.027628	-0.000765	-0.002476	-0.000447	-0.000315	1

Categorical Feature selection




```
Feature 3: 7.387715 (0.006567)
Feature 7: 7.250932 (0.007086)
Feature 10: 122.804296 (0.000000)
Feature 14: 20.361605 (0.000006)
```

'umbrella_limit','insured_hobbies', 'incident_severity',
'incident_hour_of_the_day'

P Value less than 0.05

Modelling

Methods of modelling

Model 1	Model 2	Model 3	Model 4
Logistic Regression	K Nearest Neighbors (KNN)	Naïve Bayes	Support Vector Machines

Model Validation

Training and Test Data split. 80:20

Accuracy score comparison to determine the best model.

```
## seperate the data to train and validate the models
array = data.values
X = array[:,:-1]
y = array[:,-1]
validation_size = 0.20
seed = 7
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=validation_size, random_state=seed)
```

Model Validation- Output

Model 1	Model 2	Model 3	Model 4
Logistic Regression	K Nearest Neighbors (KNN)	Naïve Bayes	Support Vector Machines

accuracy	roc_auc	precision
ScaledLR: 0.780000 (0.055396)	ScaledLR: 0.751576 (0.084693)	ScaledLR: 0.589583 (0.070158)
ScaledKNN: 0.735000 (0.060673)	ScaledKNN: 0.582181 (0.076194)	ScaledKNN: 0.389462 (0.212009)
ScaledNB: 0.708750 (0.055916)	ScaledNB: 0.717048 (0.062346)	ScaledNB: 0.431714 (0.051892)
ScaledSVM: 0.756250 (0.047516)	ScaledSVM: 0.717799 (0.082011)	

From the above results, the best performing algorithm on this training data is Logistic Regression, based on roc_auc score.

Logistic Regression on Test Data

```
Accuracy: 0.765
ROC_AUC: 0.623
[[136  13]
 [ 34  17]]
              precision    recall  f1-score   support

         0.0       0.80      0.91      0.85       149
         1.0       0.57      0.33      0.42        51

    accuracy                           0.77       200
   macro avg       0.68      0.62      0.64       200
weighted avg       0.74      0.77      0.74       200
```

Recommendation

- It will help predict fraud frequency in upcoming future claims.
- It will help identify root cause for Fraud to take necessary options for future corrections.
- Fraud impacting claims data provided from the model can be identified.
- Necessary automated steps can be taken for extensive processes to reduce manual efforts and Cost.

Total Claim Amount	Total % of Fraud rightly identified by model	Total Saving on the basis of Predicted model	Overall saving on insurance claim
5,27,61,940	80%	23,63,734	5%


Thank
You