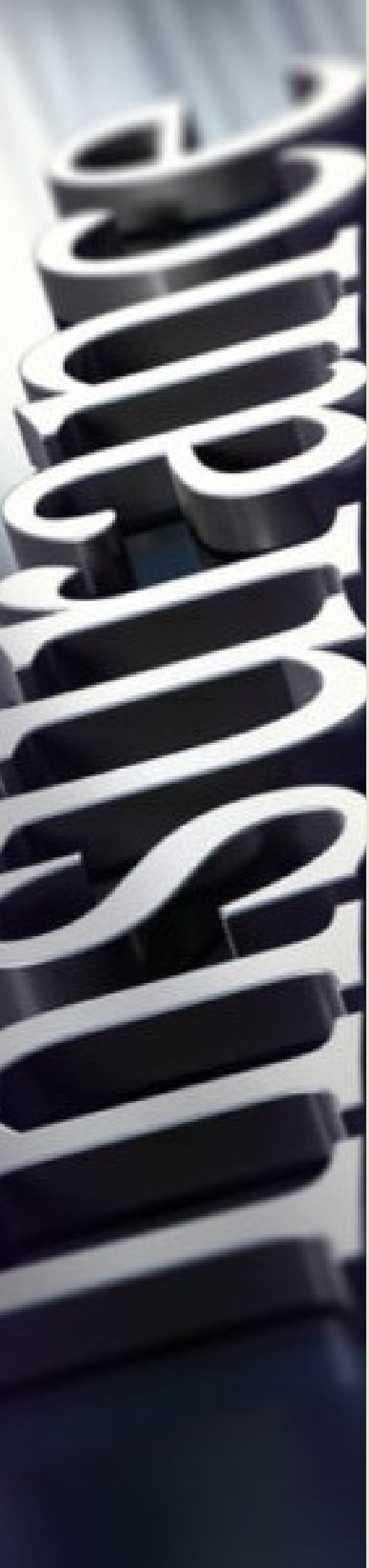
MACHINE LEARNING II
EXAM REPORT

INSURANCE FRAUD DETECTION

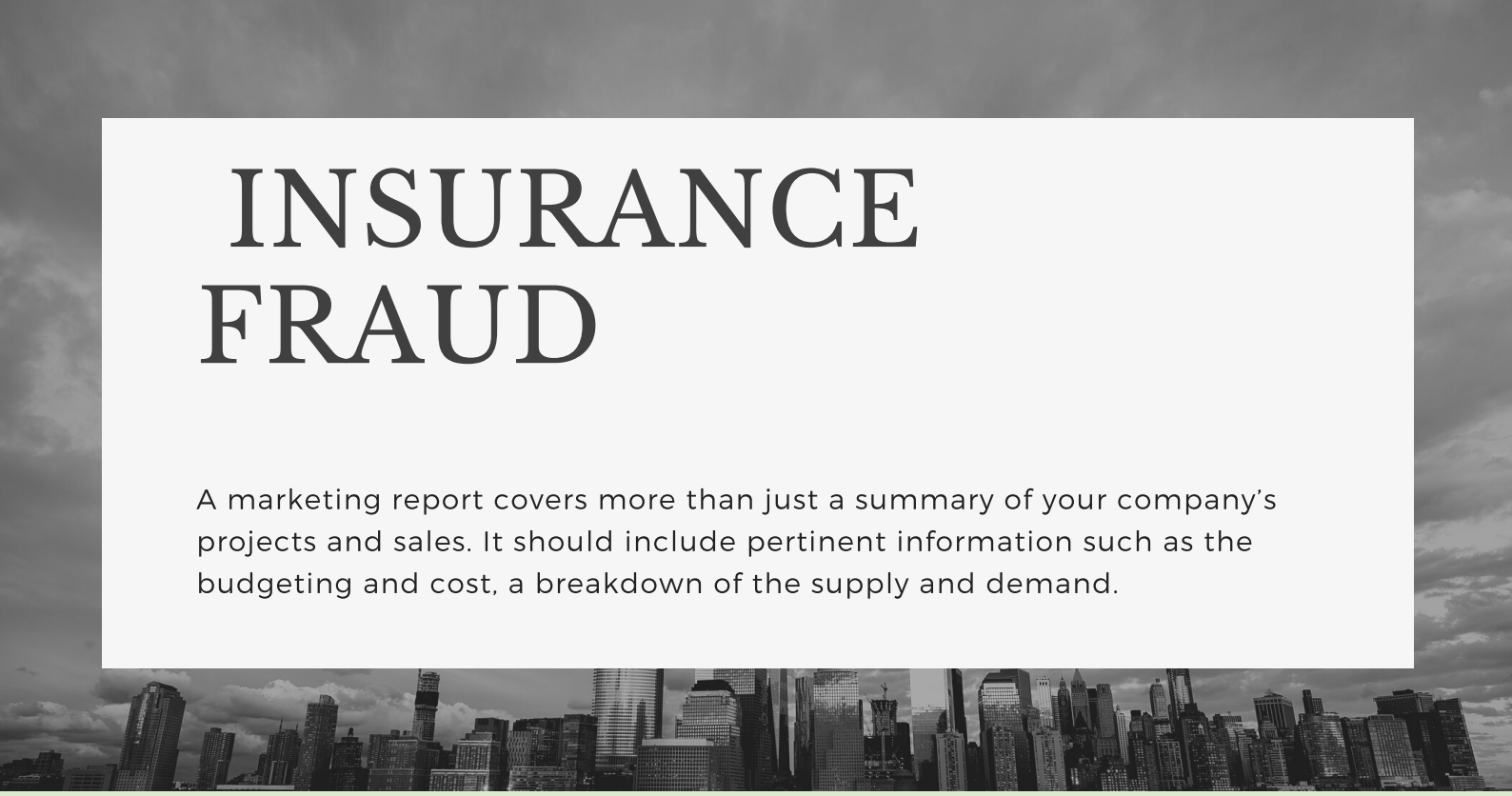
USING MACHINE LEARNING ALGORITHMS

Tashi Chotseo 20BDA01
Merin George 20BDA11
M J Aishwarya 20BDA42



CONTENTS:

- Insurance Fraud
- Problem Statement
  - Dataset
  - Libraries
- Dataset Features
- Dataset Preprocessing
- EDA
  - Univariate
  - Bivariate
  - Multivariate
- Feature Engineering
- Hypothesis Testing
- Modelling
- Conclusion
- Business Context
- References



INSURANCE FRAUD

A marketing report covers more than just a summary of your company's projects and sales. It should include pertinent information such as the budgeting and cost, a breakdown of the supply and demand.

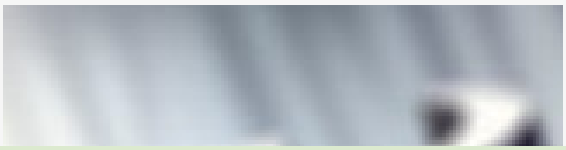
Fraud is one of the most significant and well-known issues that insurers confront. A deliberate deceit conducted against or by an insurance business or agent for the goal of financial benefit is referred to as insurance fraud. Applicants, policyholders, third-party claimants, and professionals that provide services to claims may all commit fraud at different points in the transaction.

Insurance fraud can also be committed by insurance brokers and corporate workers. Padding, or inflating claims, misrepresenting facts on an insurance application, submitting claims for injuries or damage that never occurred, and staging accidents are all examples of common frauds.

WHY INSURANCE FRAUD?

Insurance Fraud is one of the most serious concerns confronting insurance businesses today. According to industry statistics, one out of every ten claims is falsely filed. This is a concerning rate, particularly given the number of policyholders that an insurance firm may have. Some users who filed fraudulent claims did so carelessly, making it easier for the company to collect restitution and prosecute the criminals before they drive up premiums for future drivers.

Some may be done precisely in order for someone to get away with it. A significant volume of data may be checked in a short period of time using big data analytics. It encompasses a wide range of big data solutions, such as social network analysis and tele metrics. This is the most powerful tool available to insurers in the fight against insurance fraud. (https://www.smartdatacollective.com/why-data-analytics-insurance-industry-is-major-game-changer/)



Problem Statement

The objective is to create a Machine Learning model that is able to detect Car Insurance Fraud.

The Dataset

- The car insurance fraud data set we used comprises 1000 rows and 39 columns.
- *Source* :https://www.bing.com/search?q=kaggle+insurance+data&cvid=42a3af2acf6d42a8aebfceb4dbbcadec&aqs=edge.1.69i57j0l6j69i60l2.12651j0j1&pglt=43&FORM=ANNAB1&PC=U531

```
print("The number of rows  : ", data.shape[0])
print("The number of coloumns : ",data.shape[1])
print("The names of columns : ", data.columns)
data.head()
## Checking the rows and columns of the dataset
## The number of rows is 1000
## The number of columns are 39
```

```
The number of rows  :  1000
The number of coloumns :  39
The names of columns :  Index(['months_as_customer', 'age', 'policy_number', 'policy_bind_date',
       'policy_state', 'policy_csl', 'policy_deductable',
       'policy_annual_premium', 'umbrella_limit', 'insured_zip', 'insured_sex',
       'insured_education_level', 'insured_occupation', 'insured_hobbies',
       'insured_relationship', 'capital-gains', 'capital-loss',
       'incident_date', 'incident_type', 'collision_type', 'incident_severity',
       'authorities_contacted', 'incident_state', 'incident_city',
       'incident_location', 'incident_hour_of_the_day',
       'number_of_vehicles_involved', 'property_damage', 'bodily_injuries',
       'witnesses', 'police_report_available', 'total_claim_amount',
       'injury_claim', 'property_claim', 'vehicle_claim', 'auto_make',
       'auto_model', 'auto_year', 'fraud_reported'],
      dtype='object')
```

MACHINE LEARNING

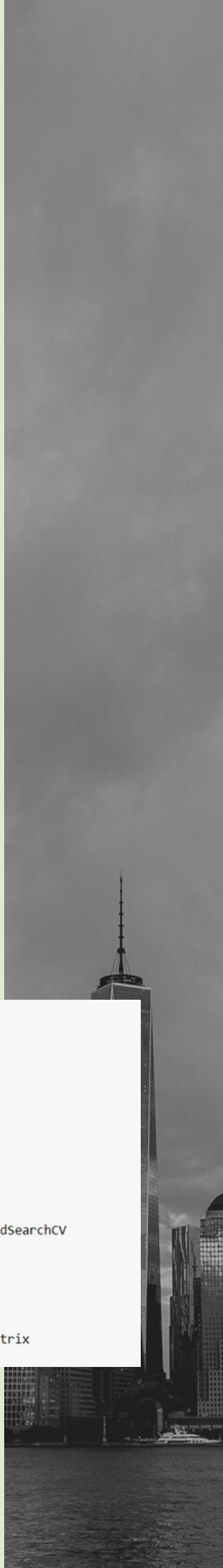
LIBRARIES

Typically, a ML library is a compilation of functions and routines readily available for use. A robust set of libraries are indispensable part of a developer's arsenal to research and write complex programs while saving themselves from writing a lot of code.

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

- Numpy : For carrying out efficient computations
- Pandas: For reading and writing on to the Csv/ Spreadsheets
- Matplotlib : For displaying plots
- Scikit-Learn: Machine Learning Library for several ML related task
- Seaborn : For Data Visualisation

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

# Importing sklearn libraries needed
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import  RandomForestClassifier
from xgboost import XGBClassifier
#model selection
from sklearn.model_selection import cross_val_score, train_test_split, GridSearchCV, KFold, StratifiedKFold, RandomizedSearchCV
from sklearn.preprocessing import  MinMaxScaler, LabelEncoder, OneHotEncoder
#model evaluation
from sklearn import metrics
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, log_loss, fbeta_score
from sklearn.metrics import auc, roc_curve, roc_auc_score, precision_recall_curve, classification_report, confusion_matrix
#oversampling
```



Dataset Features

1. months_as_customer: It denotes the number of months for which the customer is associated with the insurance company.
2. age: continuous. It denotes the age of the person.
3. policy_number: The policy number.
4. policy_bind_date: Start date of the policy.
5. policy_state: The state where the policy is registered.
6. policy_csl-combined single limits. How much of the bodily injury will be covered from the total damage.
7. policy_deductable: The amount paid out of pocket by the policy-holder before an insurance provider will pay any expenses.
8. policy_annual_premium: The yearly premium for the policy.
9. umbrella_limit: An umbrella insurance policy is extra liability insurance coverage that goes beyond the limits of the insured's homeowners, auto or watercraft insurance. It provides an additional layer of security to those who are at risk of being sued for damages to other people's property or injuries caused to others in an accident.
10. insured_zip: The zip code where the policy is registered.
11. insured_sex: It denotes the person's gender.
12. insured_education_level: The highest educational qualification of the policy-holder.
13. insured_occupation: The occupation of the policy-holder.
14. insured_hobbies: The hobbies of the policy-holder.
15. insured_relationship: Dependents on the policy-holder.
16. capital-gain: It denotes the monitory gains by the person.
17. capital-loss: It denotes the monitory loss by the person.
18. incident_date: The date when the incident happened.
20. 19. incident_type: The type of the incident.
21. collision_type: The type of collision that took place.
22. incident_severity: The severity of the incident.
23. authorities_contacted: Which authority was contacted.
24. incident_state: The state in which the incident took place.
25. incident_city: The city in which the incident took place.
26. incident_location: The street in which the incident took place.
27. incident_hour_of_the_day: The time of the day when the incident took place.
28. property_damage: If any property damage was done.
29. bodily_injuries: Number of bodily injuries.
30. Witnesses: Number of witnesses present.
31. police_report_available: Is the police report available.
32. total_claim_amount: Total amount claimed by the customer.
33. injury_claim: Amount claimed for injury
34. property_claim: Amount claimed for property damage.
35. vehicle_claim: Amount claimed for vehicle damage.
36. auto_make: The manufacturer of the vehicle
37. auto_model: The model of the vehicle.
38. auto_year: The year of manufacture of the vehicle.

Target Label: Whether the claim is fraudulent or not

- fraud_reported: Y or N

Data Preprocessing

Null Values and Missing Data.

- The Data set does not have any null values. or missing data.
- However, there are few columns in the data where there is a presence of "?" instead of values.

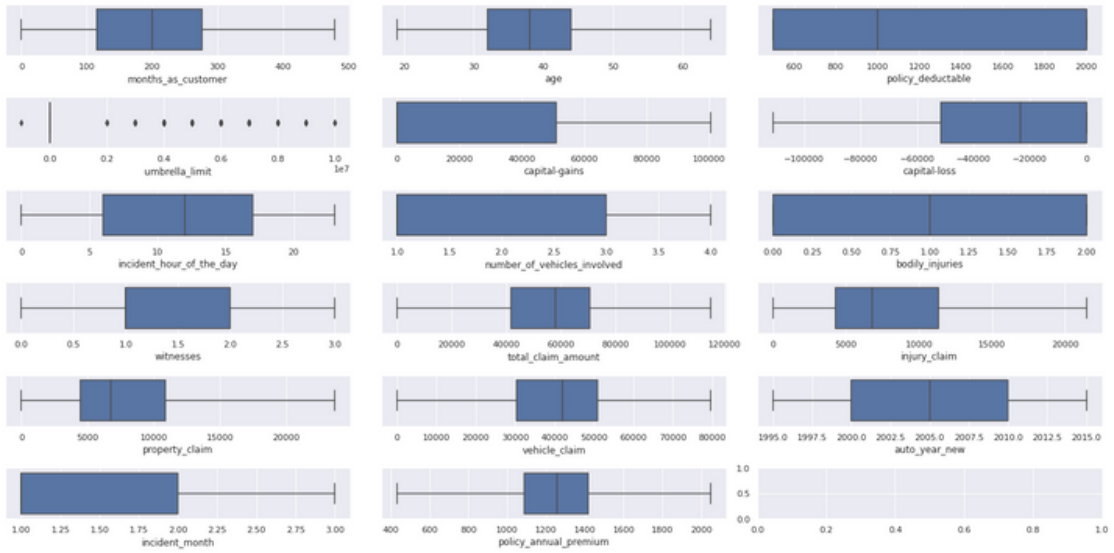
```
# There seem to be "?" in some of the features. So we need to extract t

features_with_missing_data = []
for feat in claims_data.columns:
    if '?' in claims_data[feat].values:
        features_with_missing_data.append(feat)

features_with_missing_data
```

```
['collision_type', 'property_damage', 'police_report_available']
```

- The columns "collision_type", "property_damage" and "police_report_available" contains "?".
- The "?" in each of these columns is replaced with "Undocumented" to make correct predictions
- The data is now fixed.

Outlier Analysis

- Checking for Outliers using Boxplots

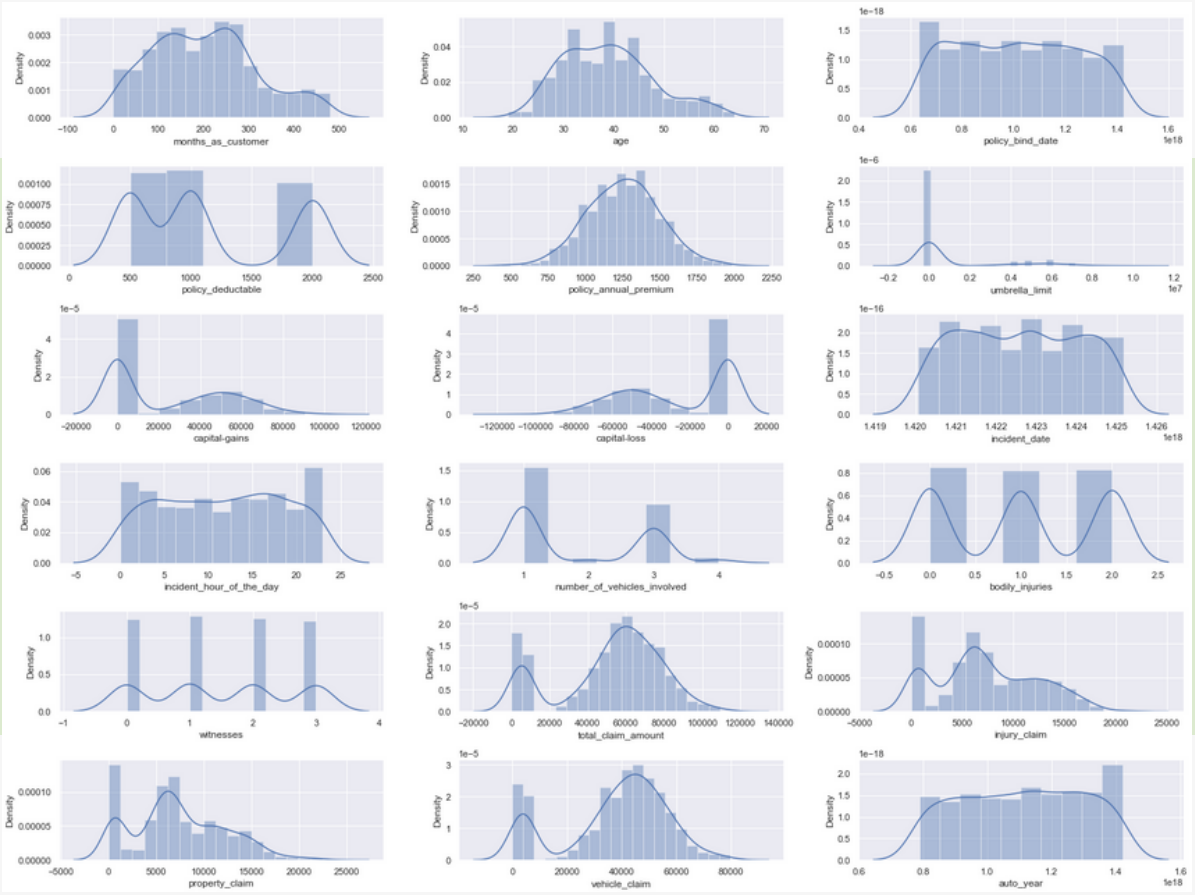


- There are no outliers in the data. But there are negative values in the Umbrella_limit because most of the Insurance Holders have an Umbrella_limit below Zero, this data is treated with 0 and 1 where 0 means no umbrella_limit and 1 otherwise.

Exploratory Data Analysis

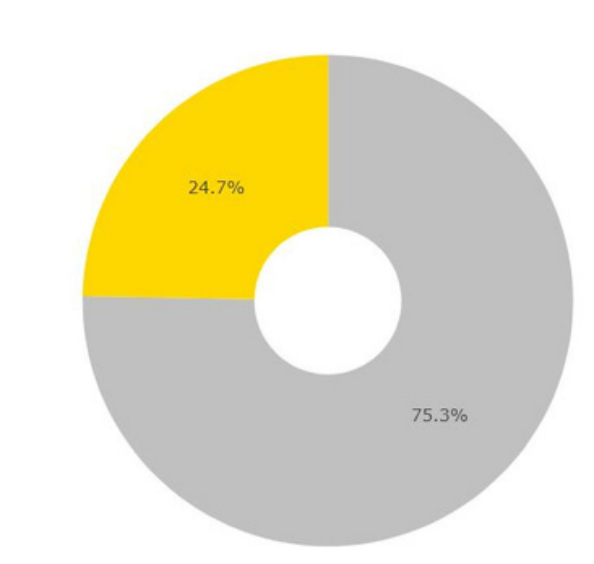
After Data pre-processing, performing Exploratory Data Analysis on the dataset will help us gain insights and pattern of the data, their distribution type, its contribution towards the target variable.

A distribution plot provides information about the kind of distribution the data follows. In the below image, the distributions of each numerical variable is graphically represented,



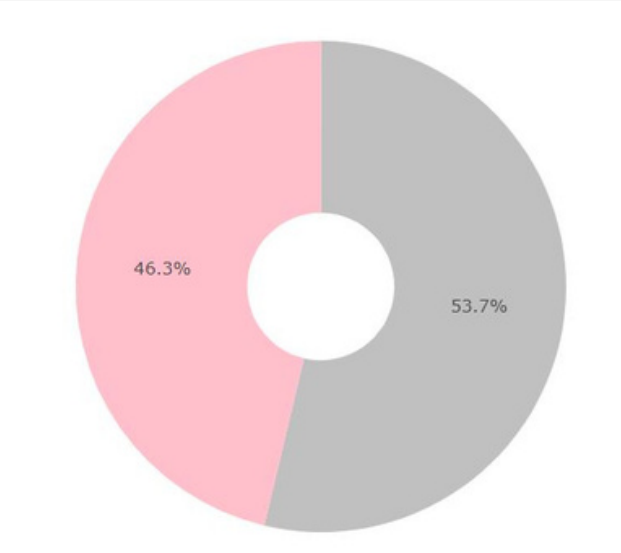
- "policy_annual_premium" follows a somewhat normal distribution.
- Whereas "insured_zip" and "total_claim_amount" seems to have two distributions.
- Policy number does not give us an insight because it contains unique id's for every individual with an insurance policy and hence we can discard this feature.
- The umbrella limit is a type of insurance that covers your responsibilities in the event that you are sued. As a result, it is impossible for this to be a negative number. Because it was assumed that the single negative value was a data input error, it was edited to become positive. It is highly skewed feature with most of the values to be 0.
- The insured_zip is the zip code where the policy is registered. This feature seems to have high number of unique value (995) i.e., unique id's which will not contribute much to our target variable.

Exploratory Data Analysis



Fraudulent Vs Non-Fraudulent Claims

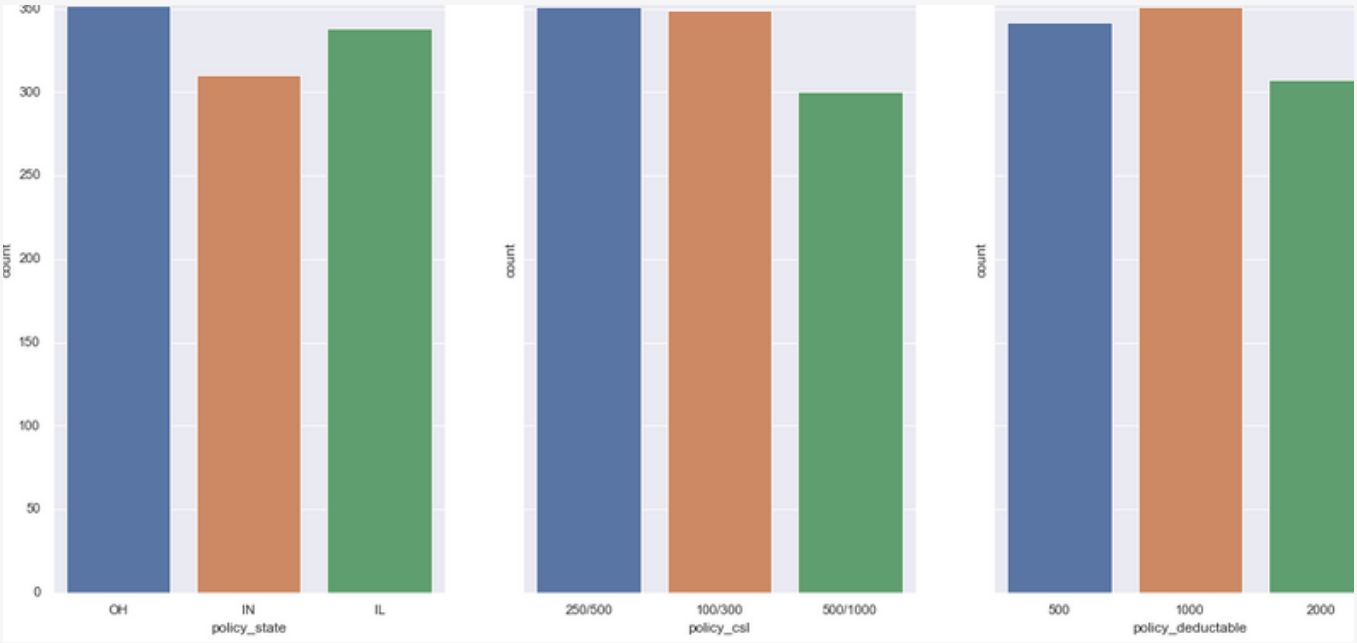
Out of 1000 records. 24.7% of the data was classified as Fraudulent and 75.3% as Non-Fraudulent.



Gender Distribution

The number of Male individuals(463 / 46.3%) is slightly lesser than the number of Female policy holders(537 / 53.7%)

Policy_State, Policy_csl and Policy_deductable

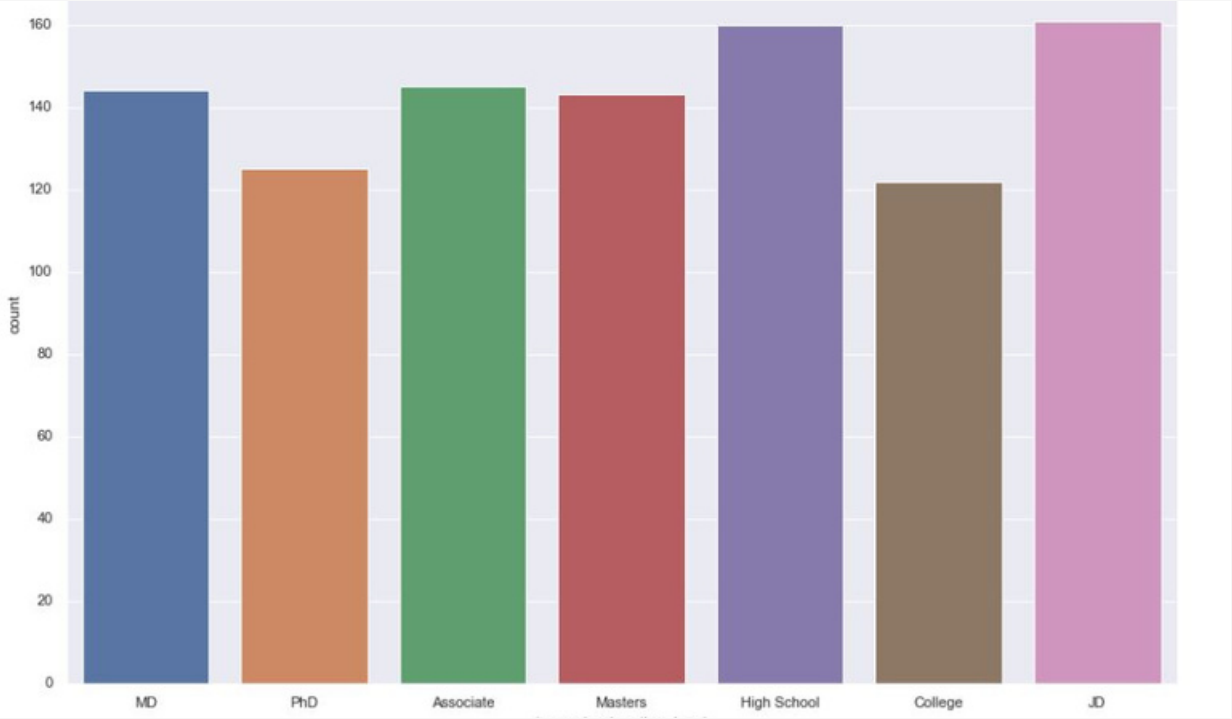


From the above graph we can infer that the features "policy_state",'policy_csl', and 'policy_deductable' have three categories in which:

- "policy_state" contains 3 states data, namely Ohio(OH), Inidana(IN), and Illinois(IL)
- "policy_csl" is subdivided into 3 sub units ranging from 250/500,100/300 and 500/100.
- "policy_deductable" is divided into categories 0,1,2

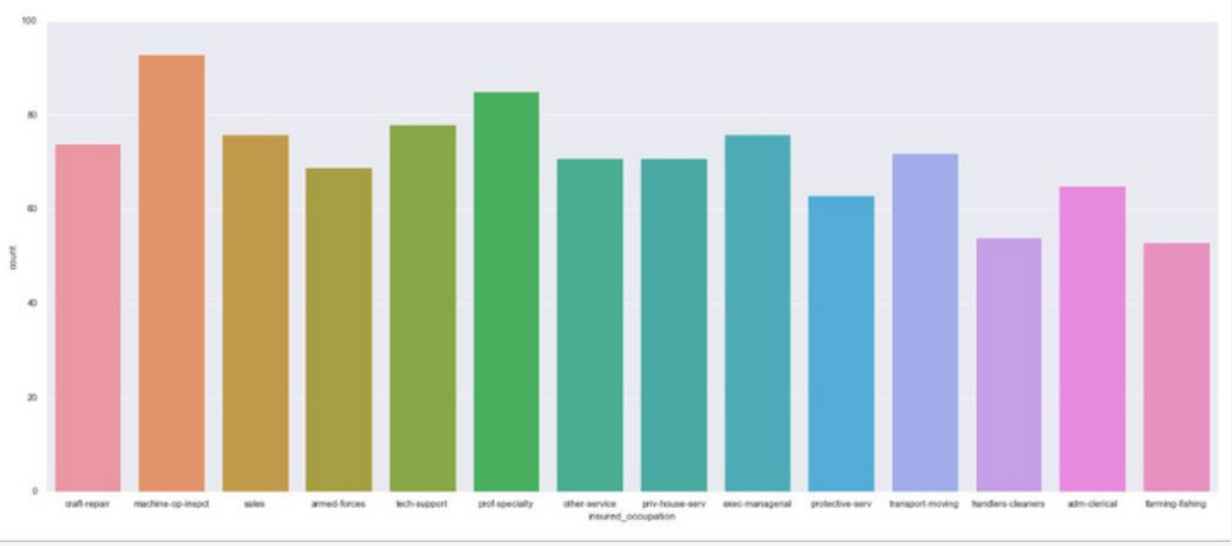
Exploratory Data Analysis

Education Level of the Insurance Holder



It is obvious from the graph above that the individuals who received insurance come from about seven different educational backgrounds. The highest number of people come from JD and high school education backgrounds, while MD, Associate, and Master's education levels are in the same range, and only college level education level persons are the least of them.

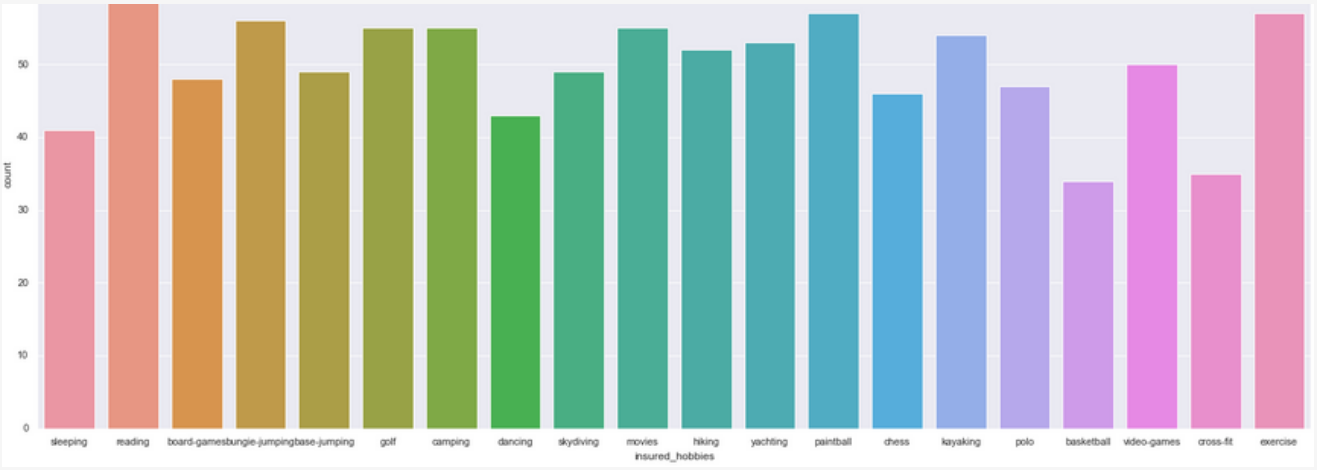
Occupations of the Insurance Holder



The characteristic "insured occupation" comprises roughly 14 categories, with the biggest number of persons belonging to machine-op-inspector, prof-speciality, and the lowest number belonging to handlers and cleaners.

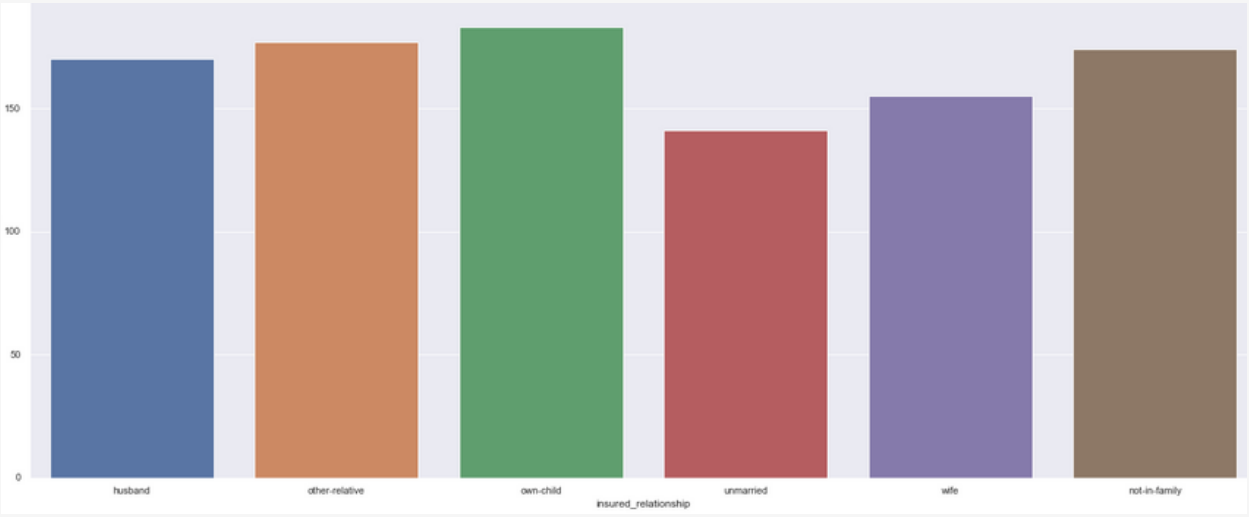
Exploratory Data Analysis

Hobbies of the Insurance Holder



The column insured hobbies denotes the hobbies of those who have purchased insurance. There are over 20 distinct hobbies ranging from sleeping to reading to dancing, among which the most popular are reading, paintballing, and exercising.

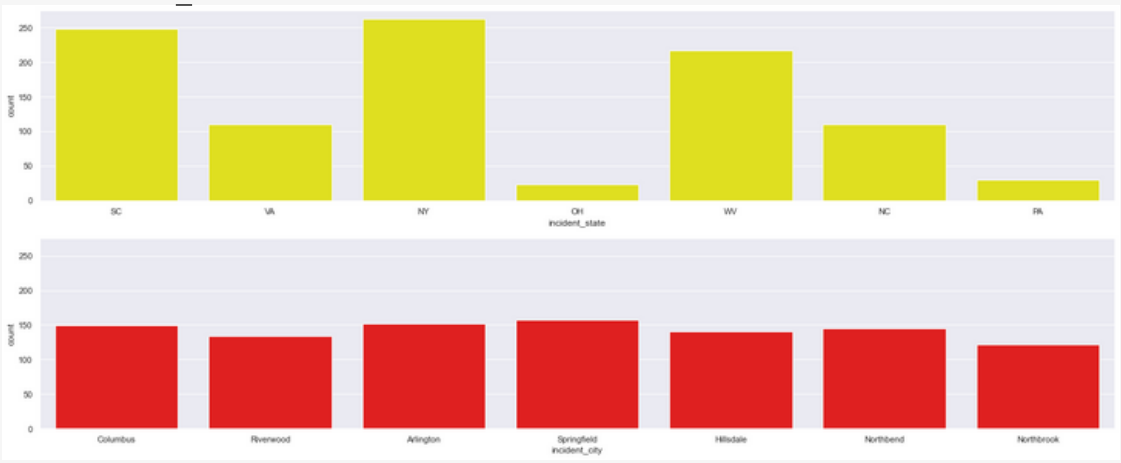
Dependents of the Insurance Holder



The feature "insured relationship" describes the policyholder's dependents. This feature has roughly 6 categories, with own child being the category with the most policyholders.

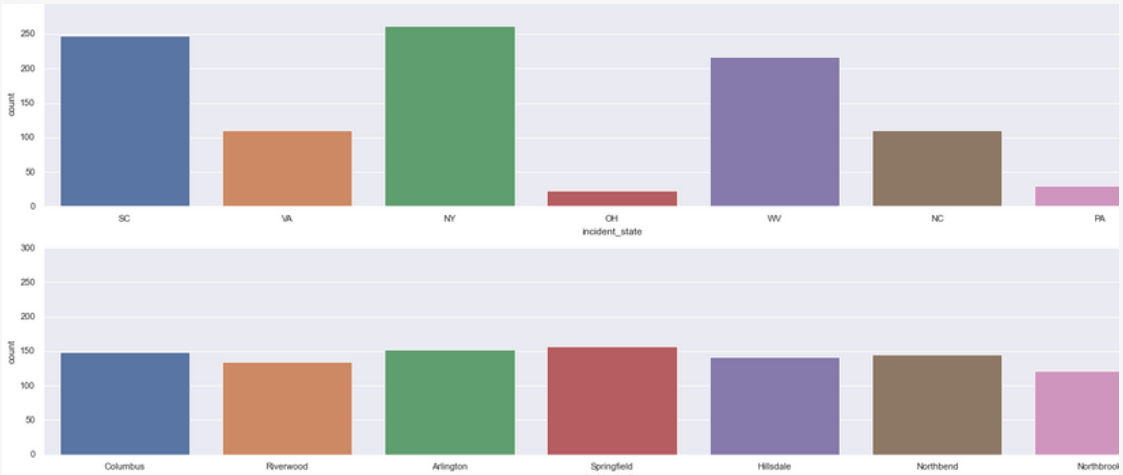
Exploratory Data Analysis

Count plots for Incident_type, Collision_type, Incident_Severity and Authorities_contacted



From the above count plot the following is inferred:
- The attribute "incident type" indicates that the most common types of incidents recorded are single-vehicle and multi-vehicle collisions.
- There are four categories in "collision type," one of which has been imputed as an undocumented or unknown case.
- The "incident severity" comprises four categories, which appear to be ordered by the severity of the incident. Indicating that type_1 (1) severity had more people.
- The "authorities contacted" field reveals who was contacted after the incident, Maximum number of policy holders reported to the Police, followed by the fire department.

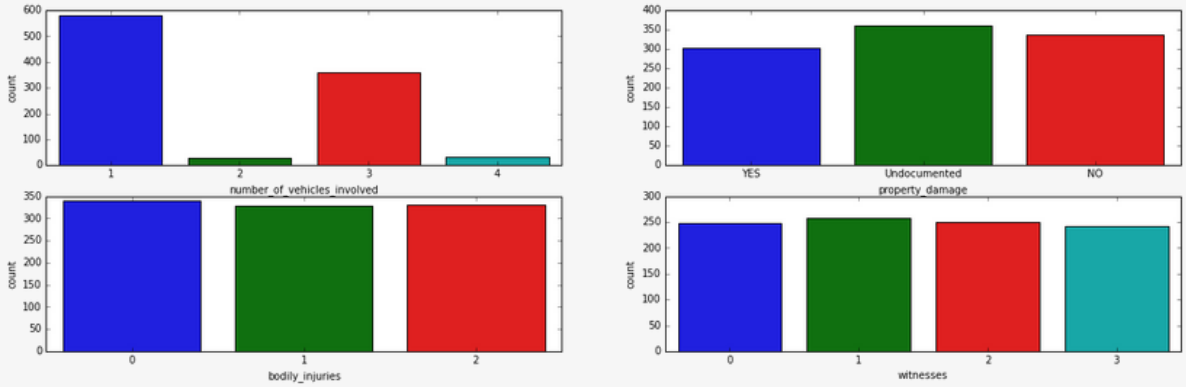
Dependents of the Insurance Holder



The feature "insured relationship" describes the policyholder's dependents. This feature has roughly 6 categories, with own child being the category with the most policyholders.

Exploratory Data Analysis

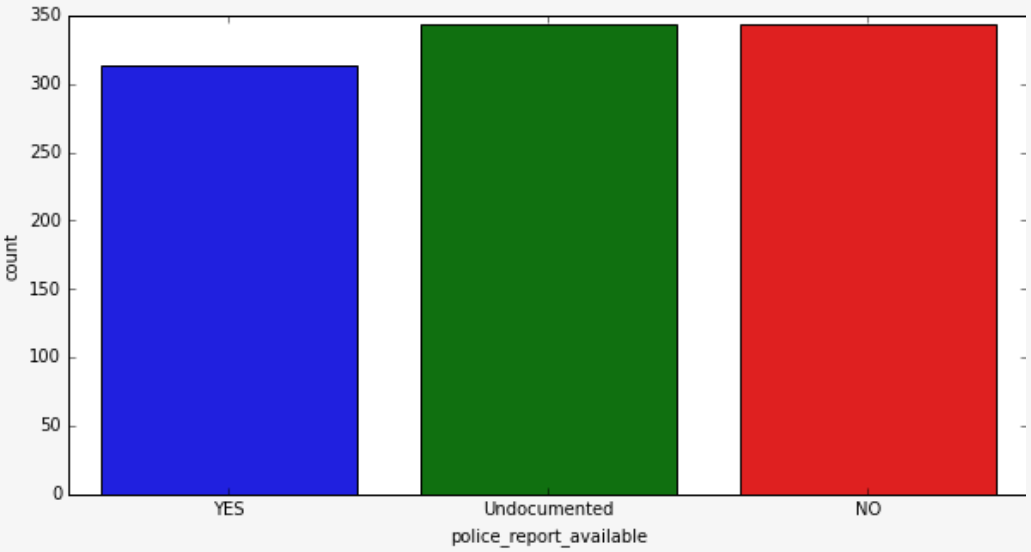
Count plots for Number_of_vehicles_involved, Property_damage, Bodily_injuries, and Witnesses



The following can be seen in the graph above:

- The "number of vehicle involved" is 1 in the vast majority of occurrences. However, there appears to be a considerable number of occurrences involving three automobiles as well.
- The "property damage" feature appears to be evenly distributed among the three categories, one of which was imputed due to missing values.
- With three values, the characteristic "bodily injuries" appears to be ordinal.
- With up to four witnesses for incidents, the function "witness" appears to be standard.

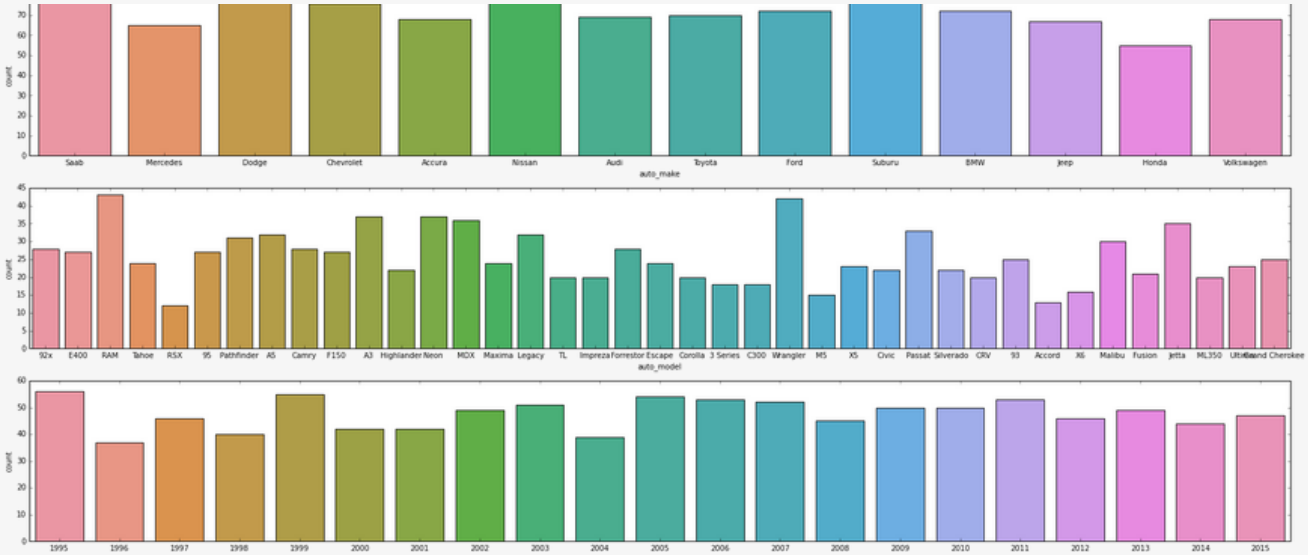
Police_report Availability of the Insurance Holder



The above graph represents the availability of police_reports for the incidents. It is divided into three categories i.e., "Yes", "Undocumented" and "No". Reports are not available or undocumented for most cases.

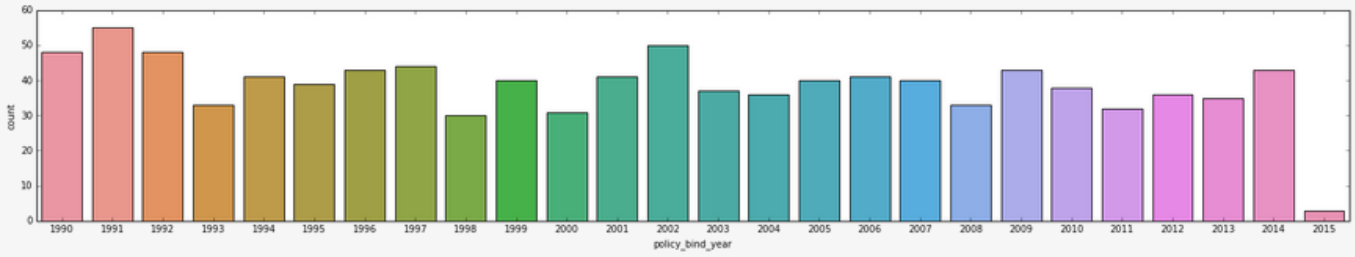
Exploratory Data Analysis

Count plots for Auto_make, Auto_model and Auto_year



- The feature "auto make" displays all of the vehicle's different automakers i.e., 14 in total.
- The feature "auto model" displays a car model with a high cardinality due to several car types.
- The feature "auto year" shows the manufacturing year for the car which could affect the claim amount.

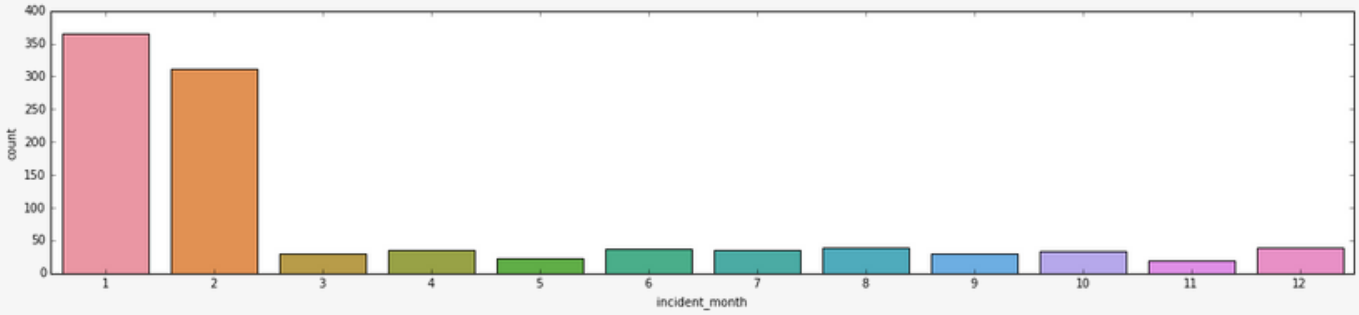
Policy Bind year of the Insurance Holder



The Policy_Bind year represents the starting year of the insurance. The policy bind year records the data from 1990-2015. For the years 1991,1992,2002 and 2014 the numbers are relatively higher compared to the other years.

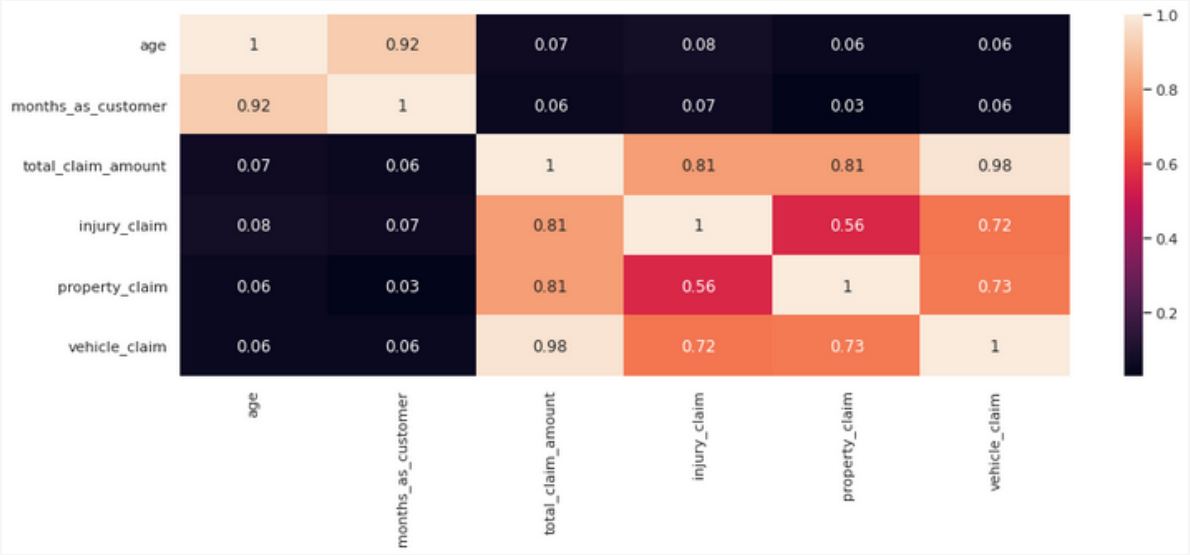
Exploratory Data Analysis

Count plot for Incident_Month



- Incident month is the month in which the accident occurred. Surprisingly, the majority of the events occur during first two months of 2015.

Correlation Matrix

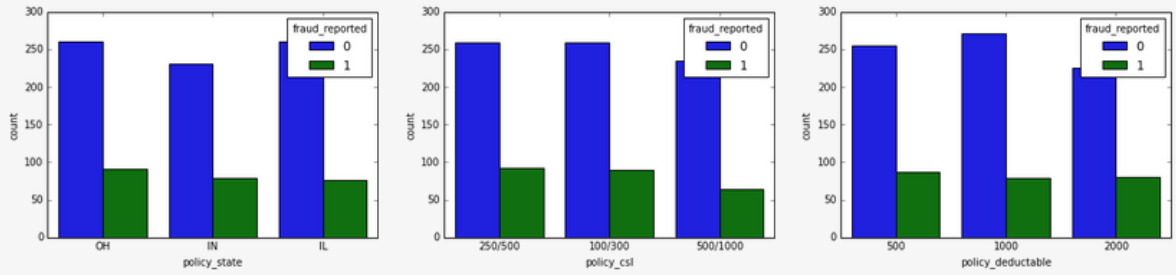


From the correlation diagram the relationship between two variables can be understood. In the above diagram, age and month_as_customer are showing a high positive relationship and vehicle_claim and total_claim_amount also strongly positively correlated.

Exploratory Data Analysis

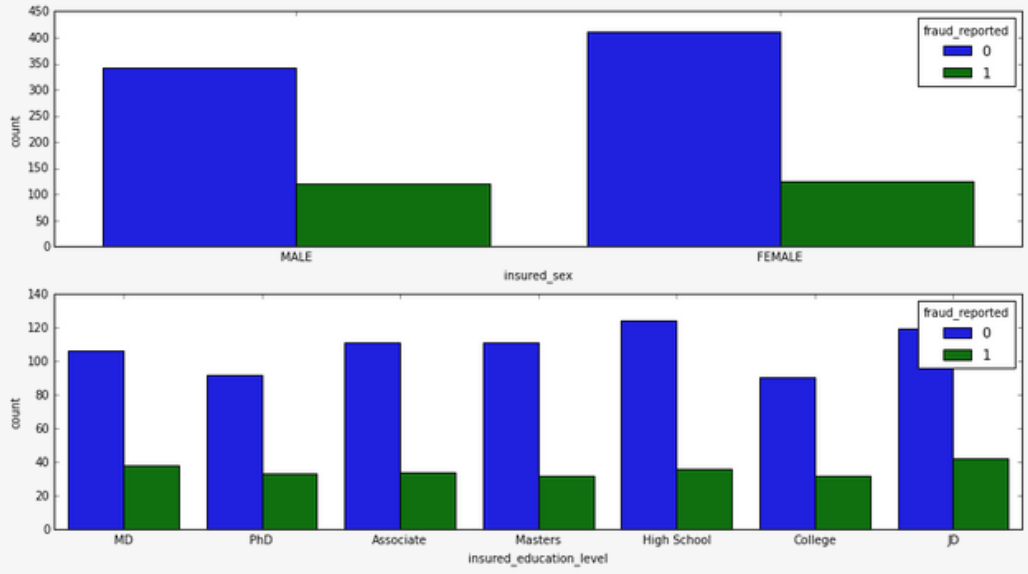
MULTIVARIATE ANALYSIS

Representing Fraud Reported (Target variable) with Policy_state, Policy_csl and Policy Deductable.



- In all three policy states, the amount of fraud reported is identical.
-For 500/1000 policy csl, the amount of fraud reported is the smallest.
-For fraudulent consumers and policy decuctable, there is no discernible trend.
-When compared to non-fraud consumers, the variation in policy annual premium is lower for fraud customers.

Fraud Reported with respect to Insured_sex and Insured_education_level.

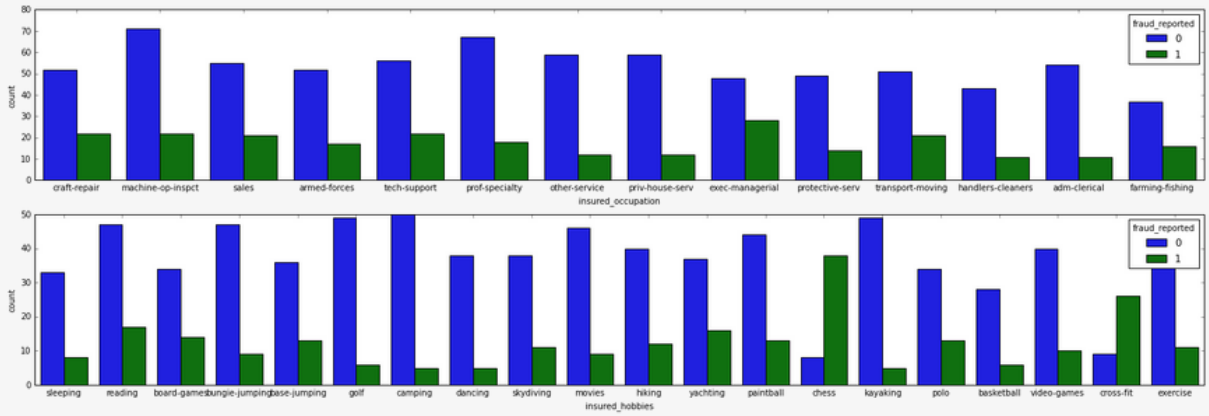


From the graph above, it can be seen that females had slightly more fraud detected than males. This could be due to the fact that females make up a larger proportion of the population than males.
Except for JD, which has a somewhat larger representation, the proportion of fraud recorded for various degree levels is similar.

Exploratory Data Analysis

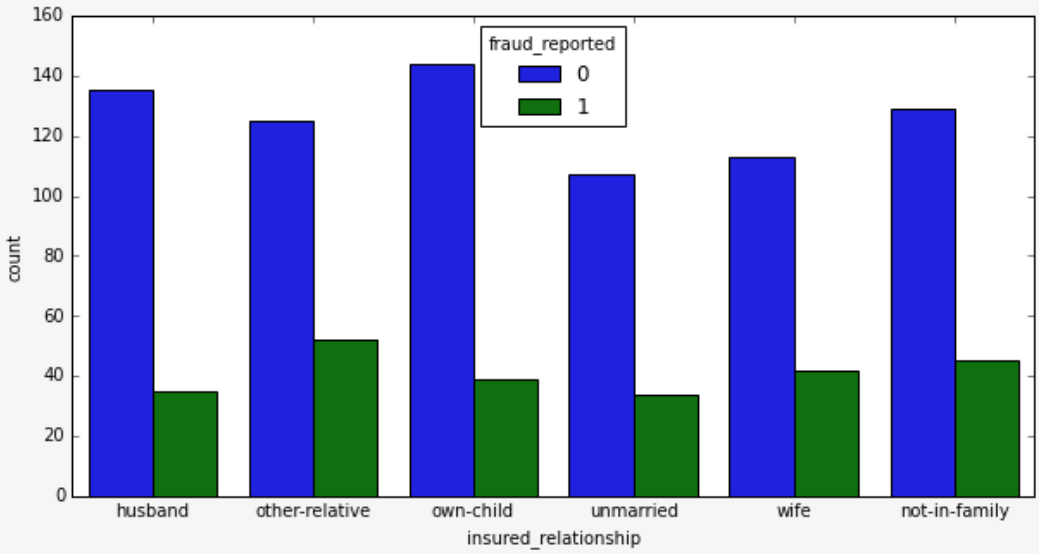
MULTIVARIATE ANALYSIS

Fraud Reported with respect to Insured_Occupation and Insured_Hobbies



-The occupation "executive-managerial" has the most fraud reports, followed by "craft-repair," "machine-op-inspct," "sales," "tech-support," and "transport-moving."
Customers who enjoy "chess" and "cross-fit" had even more frauds reported than non-frauds.

Fraud Reported with respect to Insured_relationship .

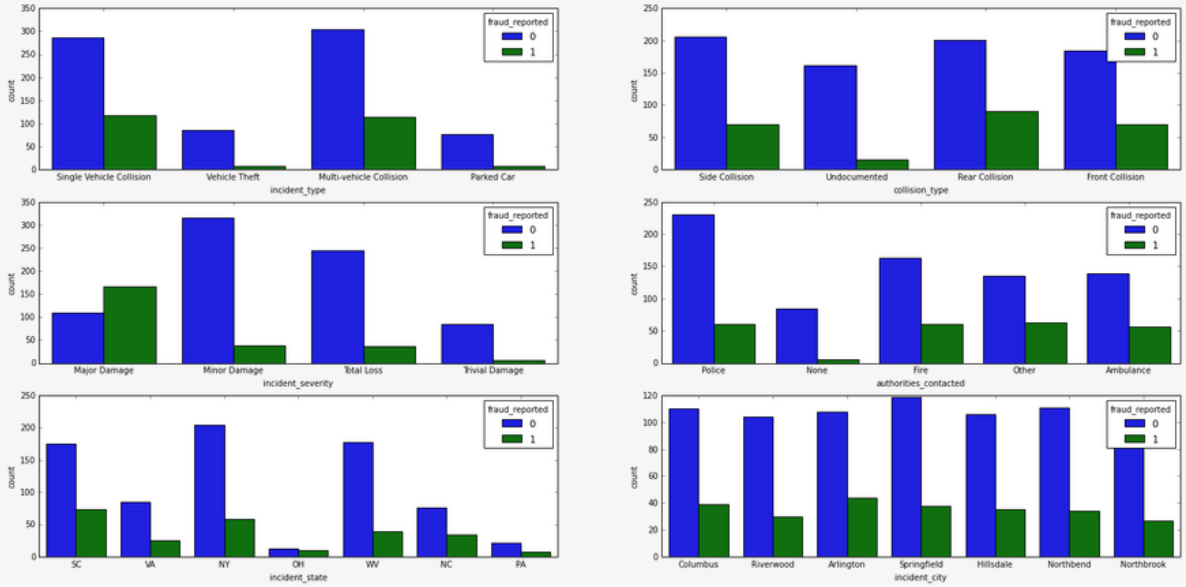


FThe percentage of different "insured relationship" appears to be similar. However, "other-relative" has the most reported frauds, followed by "not-in-family" and "wife."

Exploratory Data Analysis

MULTIVARIATE ANALYSIS

Fraud Reported with respect to Incident_type, Collision_Type, Incident_Severity, Authorities_contacted , Incident_state and Incident_city



-When compared to other incident types, "Single vehicle Collision" and "Multi vehicle Collision" appear to have more fraud cases.

The "undocumented" collision category has the fewest fraud instances, indicating that minor collisions have a lower probability of fraud cases than other collision types.

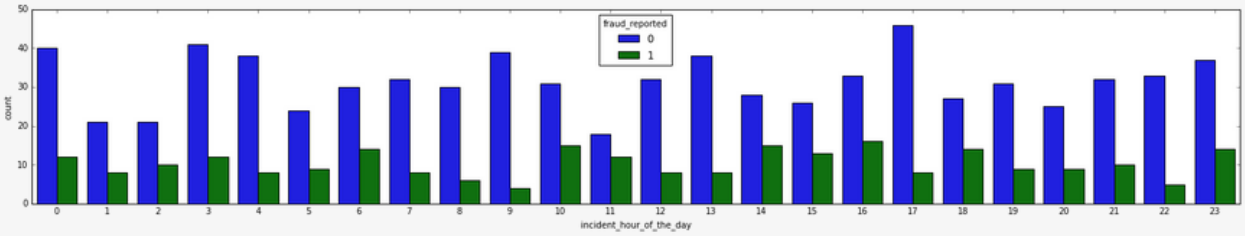
"Major Damage" has the most fraud cases based on incident severity.

Except when no one is contacted, the fraud cases appear to have a similar proportion of "authorities contacted."

The states with the highest proportion of fraud cases are South Carolina and New York, with the rest of the states having fraud cases as well.

For "incident city," the fraction of fraud cases is comparable.

Fraud Reported with respect to Incident_hour

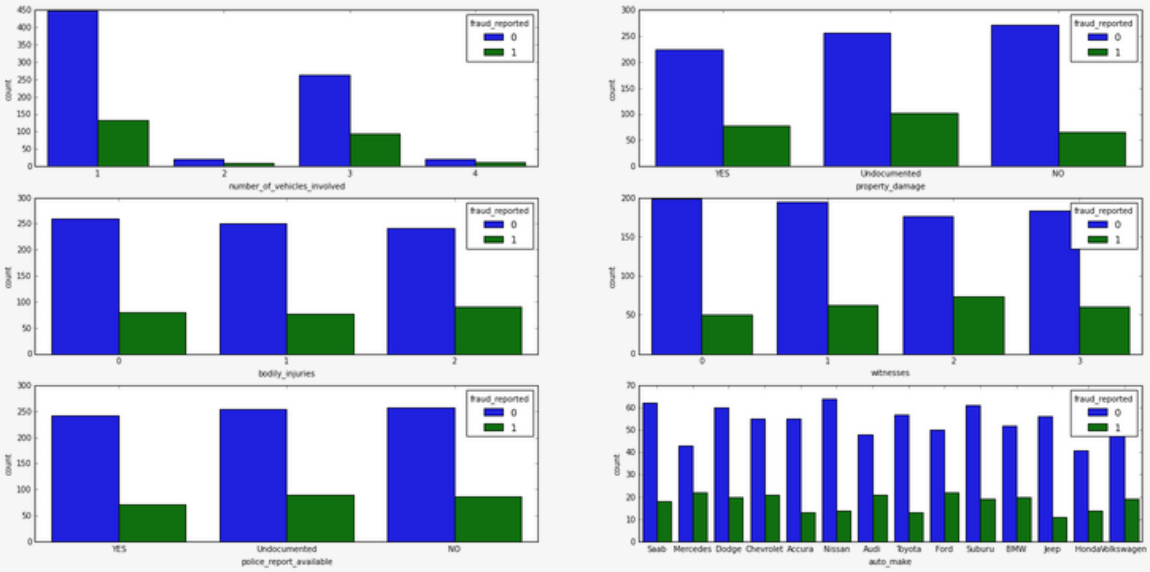


-Most number of Fraud's seems to have been reported during the 10:00-18:00 time frame.

Exploratory Data Analysis

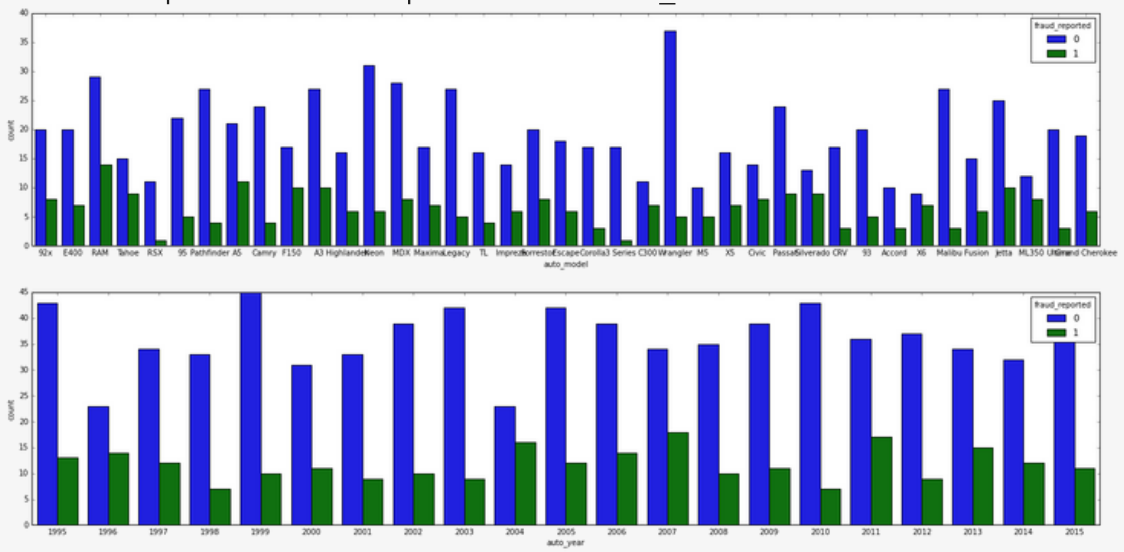
MULTIVARIATE ANALYSIS

Fraud Reported with respect to Number_of_vehicles_involved, property_damage, Bodily_injuries,Witnesses, Police_report_Available,
, Collision_Type, Incident_Severity, Authorities_contacted , Incident_state



-Fraud cases are more common when one or three automobiles are involved, but they are rare when two or four vehicles are involved.
For fraud instances, the proportions of property damage are identical.
Injuries to the body have similar dimensions to fraud instances.
For fraud cases, Witness has similar proportions.
For fraud cases, the access of police reports is distributed in a similar way.
In comparison to the rest of the automakers, Mercedes, Audi, Ford, Subaru, and BMW have more fraud instances.

Fraud Reported with respect to Incident_hour

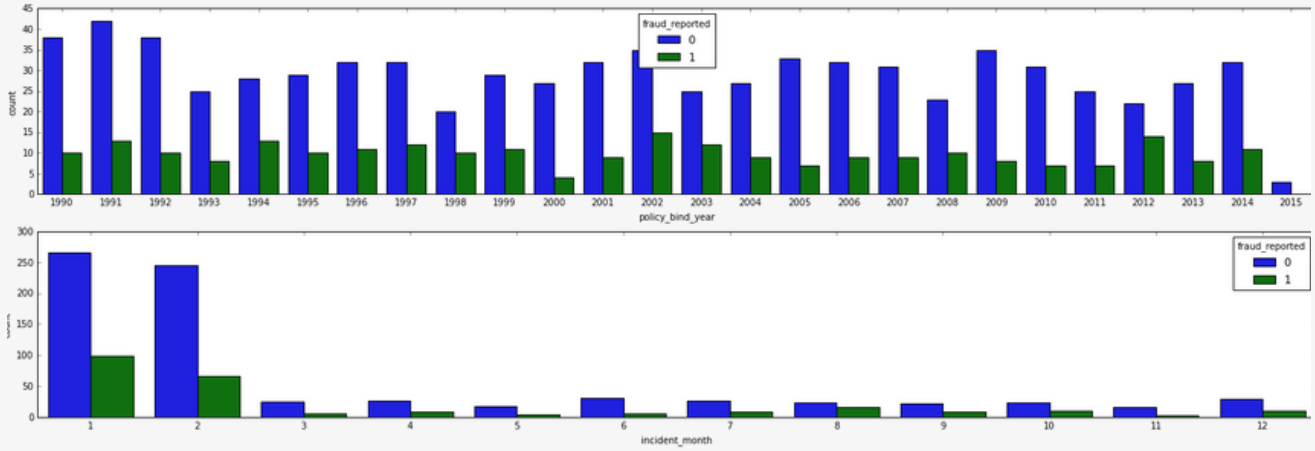


Fraud charges are quite common in RAM, A5, and Jetta.
In comparison to early year vehicles, modern year vehicles from 2007 have a higher rate of fraud.

Exploratory Data Analysis

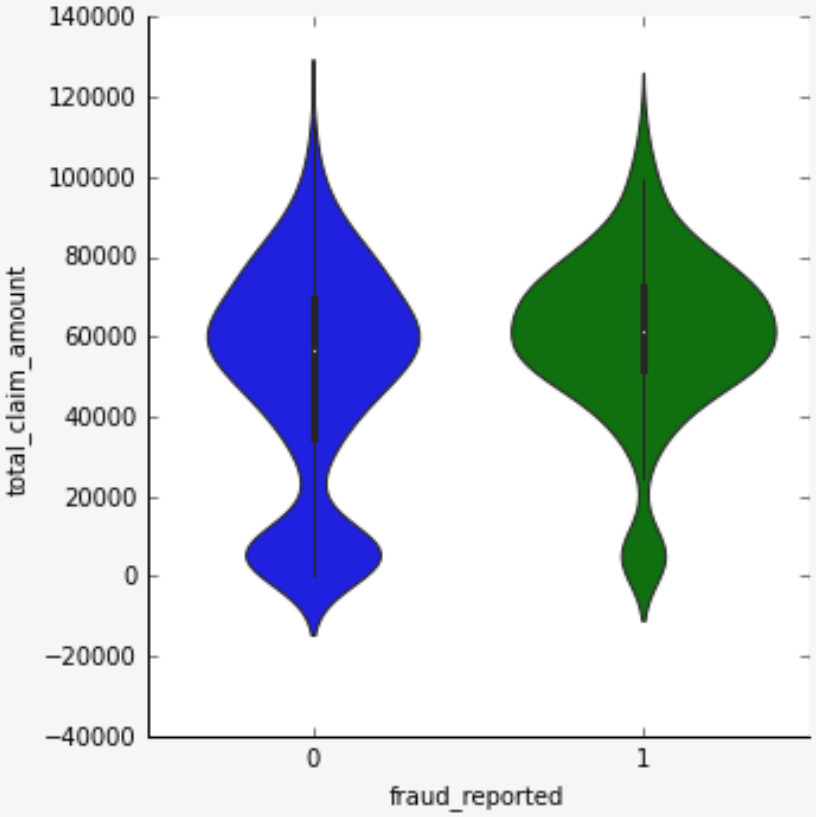
MULTIVARIATE ANALYSIS

Fraud Reported with respect to Policy_bind_year and Incident_month



-For different policy bind years, the proportion of fraud cases is similar.
The number of fraud instances is higher in the first two months of 2015, but this is largely due to the fact that the data has a large sample size in the first two months.

Fraud Reported with respect to Total_claim_amount

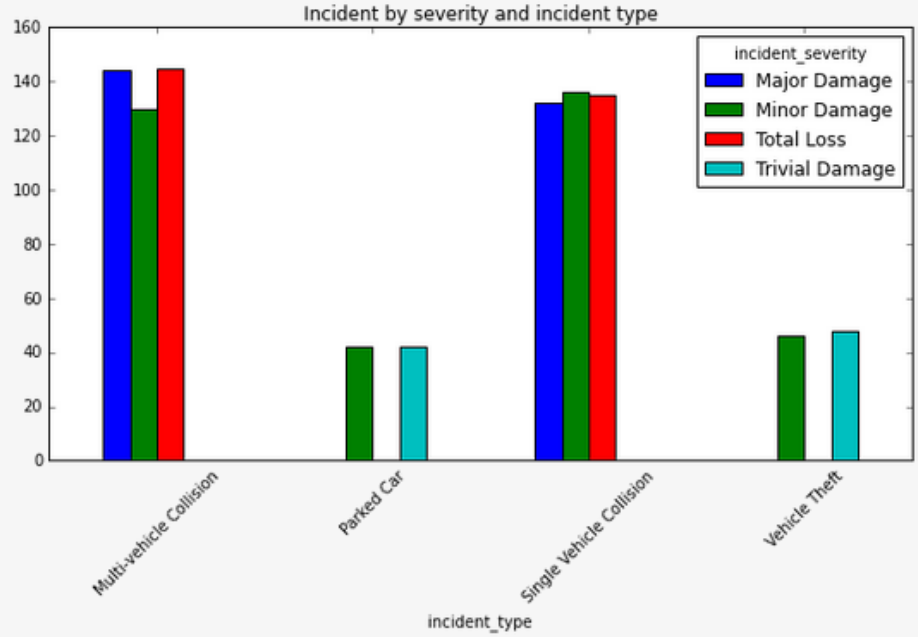


The total claim amount for fraud cases is more than the total claim amount for non-fraud instances.

Exploratory Data Analysis

MULTIVARIATE ANALYSIS

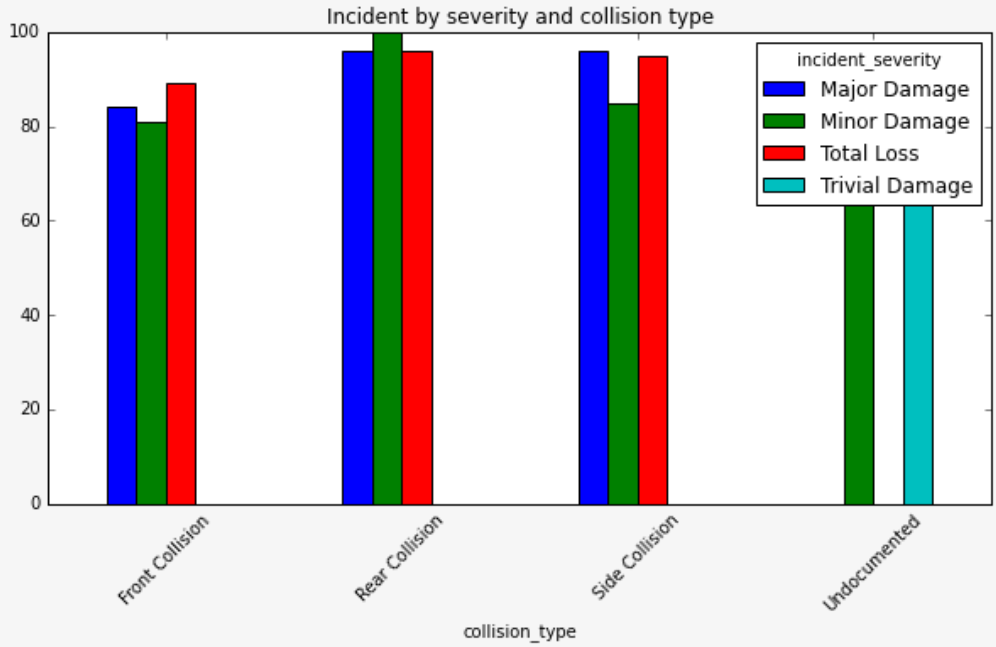
Incident By Severity and Incident_type



-When a multi-vehicle or single-vehicle accident occurs, the severity is usually major or total.

However, if the incidence is a parked automobile or a vehicle theft, the seriousness appears to be minor..

Incident by Severity and Collison_type



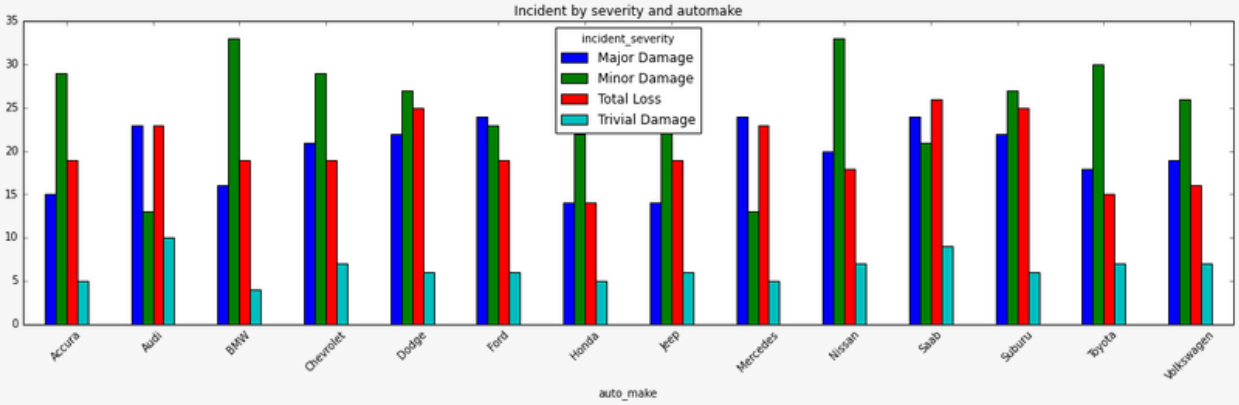
The severity of the occurrence appears to be trivial or minor damage if the collision type was Undocumented.

Front, rear, and side impacts are expected to cause severe damage.

Exploratory Data Analysis

MULTIVARIATE ANALYSIS

Incident By Severity and Auto_make

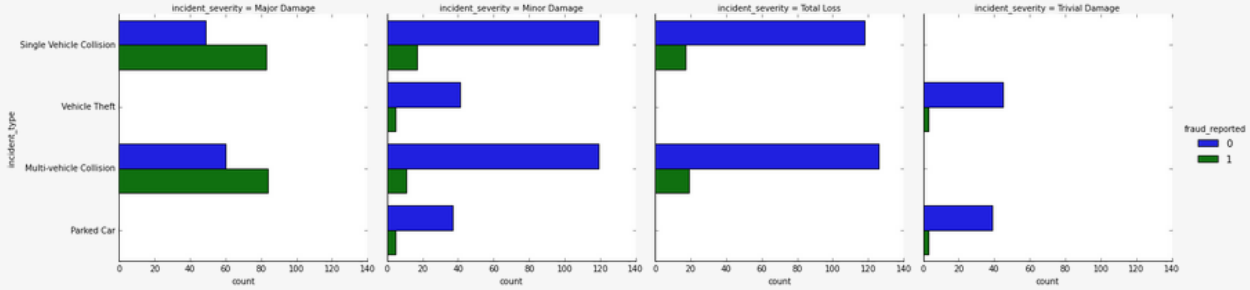


Automobile manufacturers such as Audi, Saab, Subaru, Dodge, and Ford have a greater rate of severe damage instances.

Most major portion of minor to serious damage cases are associated with the automakers BMW, Chevrolet, and Toyota.

The rest feature a moderate mix of different types of situations.

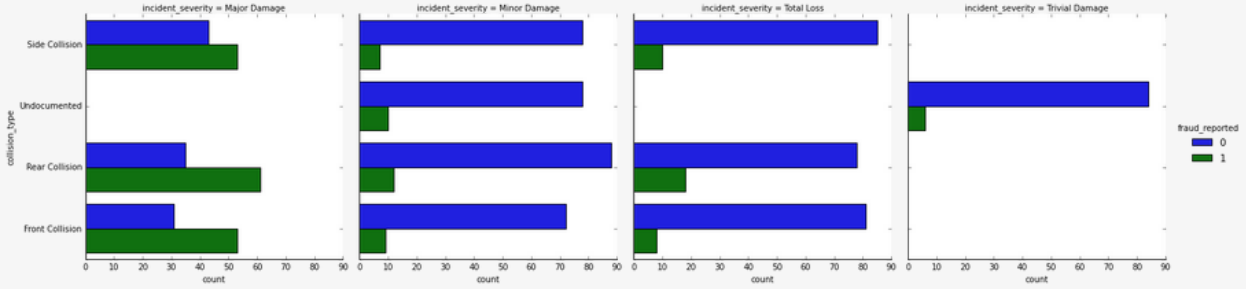
Fraud claims with respect to Incident_Severity and Incident_Type



When the severity of the incident is "Major Collision," there are far more fraudulent instances than non-fraudulent cases for single-vehicle and multi-vehicle incident types.

There is an imbalance for fraud vs. non-fraud cases per incident category throughout the rest of the incident severity.

Fraud claims with respect to Incident_severity and Collison_type



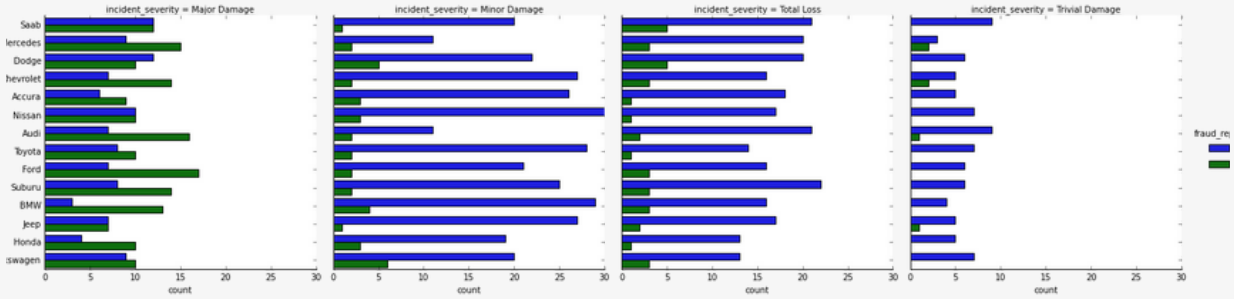
When the incident severity is "Major Damage," there are a lot of fraud instances for all sorts of collisions except "undocumented," which could be because the collision was minimal.

There is no similar trend based on collision type in the rest of the incident severity.

Exploratory Data Analysis

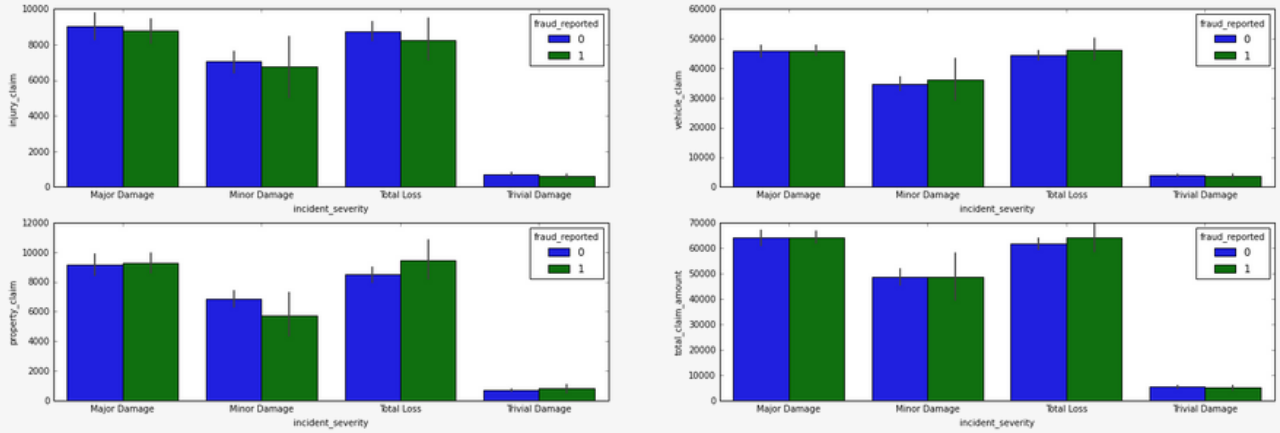
MULTIVARIATE ANALYSIS

Fraud Claims with respect to Incident_severity and Auto_make



When the severity of the occurrence is "Major Damage," automakers including Audi, Ford, BMW, and Subaru have a higher proportion of fraud cases than non-fraud instances.
For the rest of the incident severity types, there is no discernible pattern.

Fraud claims with respect to Incident_Severity and Injury_claim



There are more frauds in vehicle and property claims than in injury claims. Damage to vehicles and property may contribute to the perception of unfairness.
Because the differences balance out after summing and aggregating across different categories of claims, the fraud trend is less visible in total claims versus incident severity.

Feature Engineering

1. To improve the model accuracy, we used various forms of encoding for the1. given features. The model's accuracy was lower before the features were encoded.

Five Types of Encoding is done on various features:

a. Hot Encoding

b. Mean Encoding

c. Target Guided Ordinal Encoding

d. Frequency Encoding

e. Ordinal Encoding

```
# One Hot Encoding
sex_map={'MALE':1,'FEMALE':0}
data.insured_sex=data.insured_sex.map(sex_map)


# Mean Encoding for normal data
policy_state_map=data.groupby(['policy_state'])['fraud_reported'].mean().to_dict()
# policy_state_map = { 'IL': 0.22781065088757396,'IN': 0.25483870967741934,'OH': 0.2585227272727273 }
data.policy_state=data.policy_state.map(policy_state_map)

policy_csl_map=data.groupby(['policy_csl'])['fraud_reported'].mean().to_dict()
# policy_csl_map={'100/300': 0.25787965616045844,'250/500': 0.2621082621082621,'500/1000': 0.21666666666666667}
data.policy_csl=data.policy_csl.map(policy_csl_map)

insured_hobby_map=data.groupby(['insured_hobbies'])['fraud_reported'].mean().to_dict()
#  insured_hobby_map={'base-jumping': 0.2653061224489796,'basketball': 0.17647058823529413,'board-games': 0.2916666666666667,'bungie-jumping': 0.16071428571428573,'camping': 0.0909096
#                     'cross-fit': 0.7428571428571429, 'dancing': 0.11627906976744186,'exercise': 0.19298245614035087,'golf': 0.10909090909090909,'hiking': 0.23076923076923078,'kayc
#                     'movies': 0.16363636363636364,'paintball': 0.22807017543859648,'polo': 0.2765957446808511,'reading': 0.265625,'skydiving': 0.22448979591836735,'sleeping': 0.15
data.insured_hobbies=data.insured_hobbies.map(insured_hobby_map)

insured_relation_map=data.groupby(['insured_relationship'])['fraud_reported'].mean().to_dict()
# insured_relation_map={'husband': 0.20588235294117646,'not-in-family': 0.25862068965517243,'other-relative': 0.2937853107344633,'own-child': 0.21311475409836064,'unmarried': 0.241134
data.insured_relationship=data.insured_relationship.map(insured_relation_map)

collision_map=data.groupby(['collision_type'])['fraud_reported'].mean().to_dict()
# collision_map={'Front Collision': 0.24605678233438485,'Rear Collision': 0.2774566473988439,'Side Collision': 0.2166172106824926}
data.collision_type=data.collision_type.map(collision_map)

incident_state_map=data.groupby(['incident_state'])['fraud_reported'].mean().to_dict()
# incident_state_map = {'NC': 0.3090909090909091,'NY': 0.22137404580152673,'OH': 0.43478260869565216,'PA': 0.26666666666666666,'SC': 0.29435483870967744,'VA': 0.22727272727272727,'WV'
data.incident_state=data.incident_state.map(incident_state_map)

incident_city_map=data.groupby(["incident_city"])['fraud_reported'].mean().to_dict()
# incident_city_map={'Arlington': 0.2894736842105263,'Columbus': 0.26174496644295303 ,'Hillsdale': 0.24822695035460993 ,'Northbend': 0.23448275862068965 ,'Northbrook': 0.2213114754098
data.incident_city=data.incident_city.map(incident_city_map)

auto_make_map=data.groupby(["auto_make"])['fraud_reported'].mean().to_dict()
# auto_make_map={'Accura': 0.19117647058823528,'Audi': 0.3043478260869565,'BMW': 0.2777777777777778,'Chevrolet': 0.27631578947368424,'Dodge': 0.25,'Ford': 0.3055555555555556,'Honda':
data.auto_make=data.auto_make.map(auto_make_map)

# we'll remove this column as it contains too many unique categ values---any manufactureer can have several models
#auto_model_map=data.groupby(["auto_model"])['fraud_reported'].mean().to_dict()
#data.auto_model=data.auto_model.map(auto_model_map)

##
property_damage_map=data.groupby(["property_damage"])['fraud_reported'].mean().to_dict()
data.property_damage=data.property_damage.map(property_damage_map)


# Target Guided Ordinal Encoding

occupation_map={j:i+5 for i,j in enumerate(data.groupby(["insured_occupation"])['fraud_reported'].mean().sort_values().index)}
# occupation_map={'other-service': 5,'priv-house-serv': 6,'adm-clerical': 7,'handlers-cleaners': 8,'prof-specialty': 9,'protective-serv': 10,'machine-op-inspct': 11,'armed-forces': 12,'sal
#                 'tech-support': 14,'transport-moving': 15,'craft-repair': 16,'farming-fishing': 17,'exec-managerial': 18}
data.insured_occupation=data.insured_occupation.map(occupation_map)

severity_map={j:i+5 for i,j in enumerate(data.groupby(["incident_severity"])['fraud_reported'].mean().sort_values().index)}
# severity_map={'Trivial Damage': 5, 'Minor Damage': 6, 'Total Loss': 7, 'Major Damage': 8}
data.incident_severity=data.incident_severity.map(severity_map)

authorities_map={j:i+5 for i,j in enumerate(data.groupby(["authorities_contacted"])['fraud_reported'].mean().sort_values().index)}
# authorities_map={'None': 5, 'Police': 6, 'Fire': 7, 'Ambulance': 8, 'Other': 9}
data.authorities_contacted=data.authorities_contacted.map(authorities_map)
```

```
# Freq encoding

incident_type_map=data.incident_type.value_counts().to_dict()
data.incident_type=data.incident_type.map(incident_type_map)

# Ordinal encoding

education_map={'High School':1,'College':2,'Associate':3,'Masters':4,'JD':5,'MD':6,'PhD':7}
data.insured_education_level=data.insured_education_level.map(education_map)
```

Feature Engineering

2. We apply Min_Max Scaling On the continuos Features

```
data["injury_claim"] = scaler.fit_transform(data[["injury_claim"]])
data["property_claim"] = scaler.fit_transform(data[["property_claim"]])
data["vehicle_claim"] = scaler.fit_transform(data[["vehicle_claim"]])
data['total_claim_amount']= scaler.fit_transform(data[['total_claim_amount']])
```

```
data["policy_bind_year"] = scaler.fit_transform(data[["policy_bind_year"]])
```

```
data["auto_year_new"] = scaler.fit_transform(data[["auto_year_new"]])
```

```
data["capital-gains"] = scaler.fit_transform(data[["capital-gains"]])
data["capital-loss"] = scaler.fit_transform(data[["capital-loss"]])
```

```
data['incident_hour_of_the_day']= scaler.fit_transform(data[['incident_hour_of_the_day']])
data['number_of_vehicles_involved']= scaler.fit_transform(data[['number_of_vehicles_involved']])
data['bodily_injuries']=scaler.fit_transform(data[['bodily_injuries']])
data['witnesses']=scaler.fit_transform(data[['witnesses']])
data['incident_month']=scaler.fit_transform(data[['incident_month']])
data['policy_bind_month']=scaler.fit_transform(data[['policy_bind_month']])
```

```
data['incident_type']=scaler.fit_transform(data[['incident_type']])
data['umbrella_limit']=scaler.fit_transform(data[['umbrella_limit']])
data['incident_severity']=scaler.fit_transform(data[['incident_severity']])
data['authorities_contacted']=scaler.fit_transform(data[['authorities_contacted']])
data['insured_education_level']=scaler.fit_transform(data[['insured_education_level']])
data['insured_occupation']=scaler.fit_transform(data[['insured_occupation']])
data['policy_deductable']=scaler.fit_transform(data[['policy_deductable']])
```

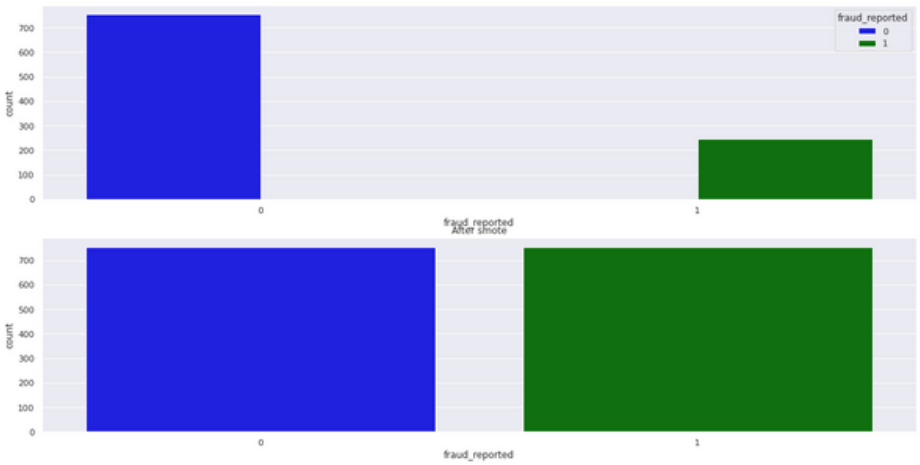
3. Drop the columns with more than 80% correlation from the corplot

```
#drop features those has more than 80% correlation with independent
data=data.drop(['age','collision_type','injury_claim'],axis=1)
```

Through Exploratory Analysis and Correlation plots, some features have been dropped to give better predictions. (Feature Drop)

['insured_zip', 'auto_year', 'policy_number' , 'policy_bind_date', 'incident_location', 'incident_date', 'auto_model', 'age', 'collision_type', 'injury_claim']

4, The dataset is highly imbalanced and hence we try to balance it with SMOTE analysis.



Hypothesis Testing

In Hypothesis Testing we try to evaluate two mutually exclusive statement ie., Null Hypothesis (H0) and Alternate Hypothesis (H1) on a population data using a sample size of the same.

Steps:
1, Make an initial assumption {H0}.
2. Collect the data
3. Gather evidence too either reject or accept the Null Hypothesis {H0}.

For modelling, one of the columns that was dropped is "auto_model" and hence by using Hypothesis testing, we are checking if it is a significant contributor to the Target variable or Not.

H0: There is no relation between the column auto_model and target variable "fraud_reported"

H1: There is a relation between the column auto_model and target variable "fraud_reported"

We run Chi-square test to check the p-value and infer accordingly.
- If the p-value is <0.05 then we reject the Null Hypothesis.
- If the p-value is >0.05 then we fail to reject the Null Hypothesis.

```
[83] print("chi2 :",chi2)
     print("p_value :",p)
     print("Degree of freedom :",dof)

     chi2 : 46.65817014569841
     p_value : 0.15826457876312205
     Degree of freedom : 38
```

Inference:

Since the p_value is 0.15 which is seemingly greater than 0.05, hence we fail to reject the Null Hypothesis, that "There is no relation between the column auto_model and target variable "fraud_reported"

Modelling.

The major aspect of a Machine Learning project comes down to the accuracy of predictions/classification that is obtained by fitting various Machine Learning Algorithms for the pre-processed data.

For the Car Insurance Fraud Detection, Three models have been tested,

1. Logistic Regression
2. RandomForestClassifier
3. XgBoostClassifier

Logistic Regression

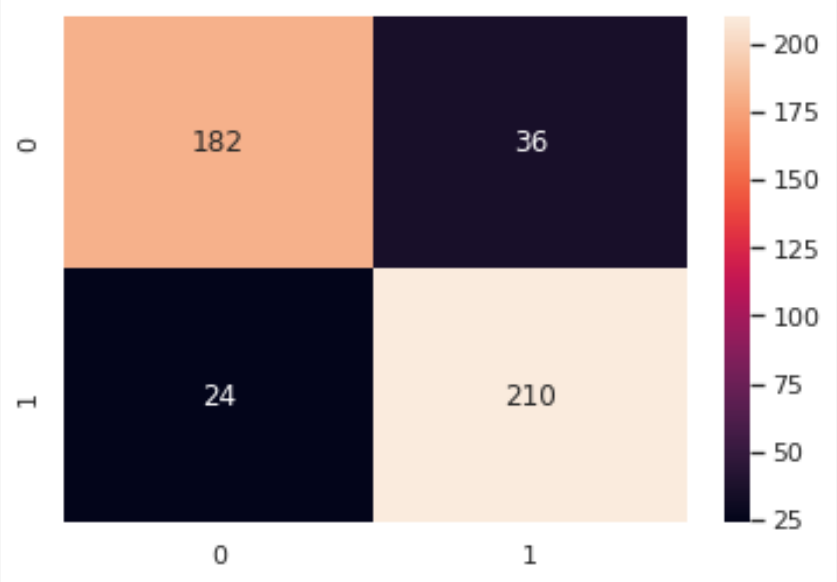
- Logistic Regression is the go to model for Insurance fraud prediction in several use cases, It is also applicable for the dataset given as the task at hand is Classification and prediction .

The result obtained from Logistic Regression are :

```
LogisticRegression accuracy score is
Training: 86.62%
Test set: 86.73%
```

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	0.88	0.83	0.86	218
1	0.85	0.90	0.88	234
accuracy			0.87	452



Modelling.

RandomForestClassifier

-A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

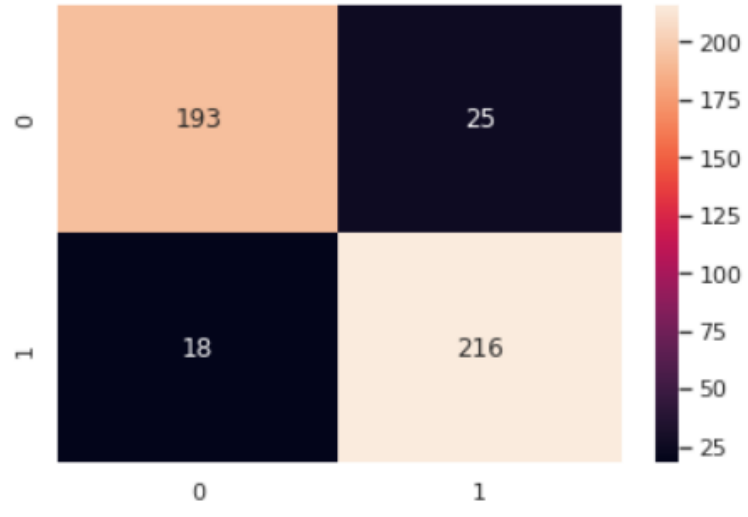
```
RandomForestClassifier accuracy score is
Training: 100.00%
Test set: 90.49%
```

However, the model Overfits.

Classification Report for RandomForestClassifier:

	precision	recall	f1-score	support
0	0.91	0.89	0.90	218
1	0.90	0.92	0.91	234
accuracy			0.90	452
macro avg	0.91	0.90	0.90	452
weighted avg	0.91	0.90	0.90	452

Accuracy Score : 0.9048672566371682



Modelling.

XgBoostClassifier

XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. XgBoost is also widely used to detect Insurance Fraud Detection.

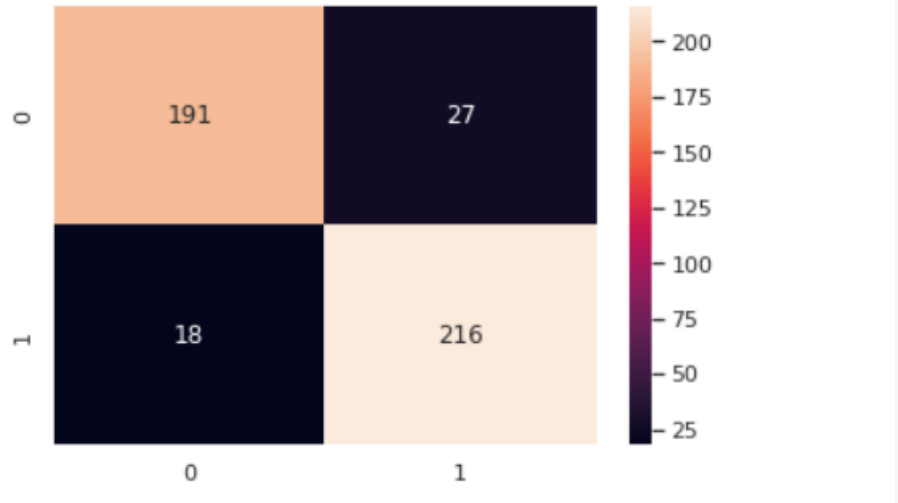
```
XgboostClassifier accuracy score is
Training: 100.00%
Test set: 90.04%
```

However, the model Overfits

Classification Report for XgBoostClassifier:

	precision	recall	f1-score	support
0	0.91	0.88	0.89	218
1	0.89	0.92	0.91	234
accuracy			0.90	452
macro avg	0.90	0.90	0.90	452
weighted avg	0.90	0.90	0.90	452

Accuracy Score : 0.9004424778761062



Hyperparameter Tuning

There are concerns with Model Overfiltting for RandomForestClassifier and XgBoostClassifier . Hyperparameter Tuning is the next step performed to fine-tune the models and improve prediction.

Mertics and Evaluation:
Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes.
We have done SMOTE analysis to balance the dataset and hence, F1 score is the Metric we are looking at.

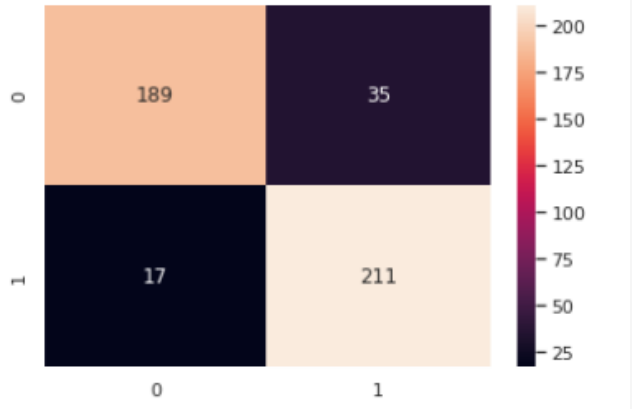
After Hyperparameter Tuning of RandomForestClassifer and XgBoostClassifier, the following results are obtained.

RandomForestClassifer

```
RandomForestClassifier accuracy score is
Training: 93.83%
Test set: 88.50%
```

	precision	recall	f1-score	support
0	0.92	0.84	0.88	224
1	0.86	0.93	0.89	228
accuracy			0.88	452
macro avg	0.89	0.88	0.88	452
weighted avg	0.89	0.88	0.88	452

```
Accuracy Score :  0.8849557522123894
```

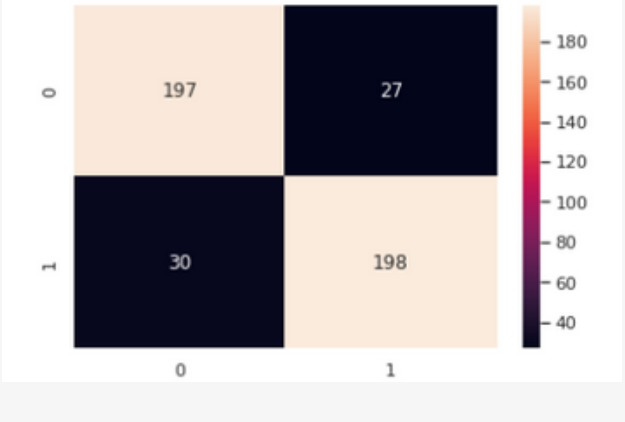


XgBoostClassifier

```
XgboostClassifier accuracy score is
Training: 94.69%
Test set: 87.39%
```

	precision	recall	f1-score	support
0	0.87	0.88	0.87	224
1	0.88	0.87	0.87	228
accuracy			0.87	452
macro avg	0.87	0.87	0.87	452
weighted avg	0.87	0.87	0.87	452

```
Accuracy Score :  0.8738938053097345
```



Conclusion

With respect to the given Problem Statement and after employing multiple ML models to identify frauds, the following is observed

Some of the most essential numerical and categorical variables were examined in relation to the objective variable "Fraud reported" in order to achieve the categorization aim.
We tested and compared a variety of classifiers included in the scikit-learn machine learning models. To categorise the data, we utilised Logistic Regression, RandomForestClassifier, and XgBoostClassifier.

The performance metrics of several models are shown in the table below. The accuracy of many models was chosen, and the highest test F1 was determined based on their performance which is RandomForestClassifier after hyper parameter tuning with 88% accuracy, and F1 score of 88% for 0(Fraud-not reported) and 89% for (Fraud-reported).

Because the dataset only comprises 1000 rows and 39 features, it is recommended that this model be tested and run in real time on a larger dataset.

Classification Report Before Parameter Tuning

Model	Accuracy	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score
Logistic regression	83%	86.24%	83.195	0-86% 1-81%	0-79% 1-88%	0-82% 1-84%
RandomForestClassifier	90%	100%	90.04%	0-93% 1-88%	0-87% 1-93%	0-90% 1-90%
XGboostClassifier	90.92%	100%	90.93%	0-92% 1-90%	0-89% 1-93%	0-91% 1-91%

Classification Report After Parameter Tuning

Model	Accuracy	Training Accuracy	Testing Accuracy	Precision	Recall	F1 score
RandomForestClassifier	88%	93.83%	88.50%	0-92% 1-86%	0-84% 1-93%	0-88% 1-89%
XGboostClassifier	87%	94.69%	87.39%	0-87% 1-88%	0-88% 1-87%	0-87% 0-87%

Business Context

Insurance fraud is a massive concern for insurers around the world, and it is increasing year after year. The most popular topic of conversation in the Property and Casualty Insurance Industry is claims fraud, with the auto segment accounting for the majority of it. Vehicle Insurance Fraud occurs when someone deceives an auto insurance provider for financial gain. While some examples of car insurance fraud are more serious than others, fraud is far from a victimless crime. Insurance fraud in the United States (excluding health care) is estimated to be worth more than $40 billion each year. Insurance companies lose at least $29 billion a year as a result of vehicle insurance fraud.

Around 90% of respondents to a 2019 insurance fraud investigation survey claimed they utilise technology to detect claim frauds as a remedy to such situations. Insurers, on the other hand, face a significant challenge in fighting insurance fraud due to data integration. There's no denying that technology like machine learning and artificial intelligence aid in finding solutions to similar problems.

Fraud in this context is not limited to criminal activity, but is expanded to include misrepresentations and omissions, whether malicious or less than complete disclosure.

Fraud Prevention and Detection
Insurers have practices in place to detect and limit the impact of fraud, often with dedicated teams in place that take advantage of fraud prevention tools and technological advances.Technology

Technology
Technology has enabled companies to leverage data and build predictive models. Companies generally do not have an algorithm in place to identify and flag questionable claims; however, many have plans to implement a program.

Business Context

What innovations in fraud detection insurance should be seen/included in the next 5 years?

- More modeling, data use.
- Advancements in using analytics, in particular for new business fraud detection.
- Needs to be easy and quick.
- Big data analytics.
- There are areas of fraud that will occur that we haven't and can't even think of. By the time we figure it out it's potentially too late.

There are many challenges to creating a solution to reduce the incidence of insurance fraud. Most solutions will require data usage, but insurers may have privacy and security concerns regarding the use of this data outside of its traditional uses. Additionally, insurance data is often technically difficult to obtain from fragmented internal administration systems.

Fraud, including fraud that is unknown to and undetected by insurers, is a costly and challenging problem facing the industry. Insurance innovations that focus on data solutions may indirectly result in a lower incidence of fraud. The motivation for these advancements is more likely to stem from the development of fluid less underwriting solutions, improved tools to serve customers or improved online distribution capabilities. The solution to the problem of fraud lies in creating tools, using data, and developing products that enhance trust and transparency among all parties involved in the insurance transaction.

References

1. Insurance Claim | Kaggle
2. Insurance Fraud Definition (investopedia.com)
3. Car Insurance Frauds | Motor Insurance Frauds in India (policybazaar.com)
4. 21 Huge Insurance Fraud Statistics [The View from 2021] (thehighcourt.co)
5. The Role of Data and Analytics in Insurance Fraud Detection | Reuters Events | Insurance
6. Handling imbalanced data [Analytics Vidhya]
7. Concept of mean encoding and scaling [Coursera -HSE UNIVERSITY]
8. Threshold-Moving for imbalanced classification [Machine Learning Mastery]
9. XGBoost model for car insurance prediction | Kaggle
10. Robust logistic regression for insurance risk classification (repec.org)
11. Vehicle insurance — Random forest classifier | Aviral Bhardwaj | Medium | Medium
12. Accuracy vs. F1-Score. A comparison between Accuracy and… | by Purva Huilgol | Analytics Vidhya | Medium
13. Statistics - Hypothesis testing (tutorialspoint.com)