Insurance Claims – Fraud Detection

Using Machine Learning

* **Problem Definition:**

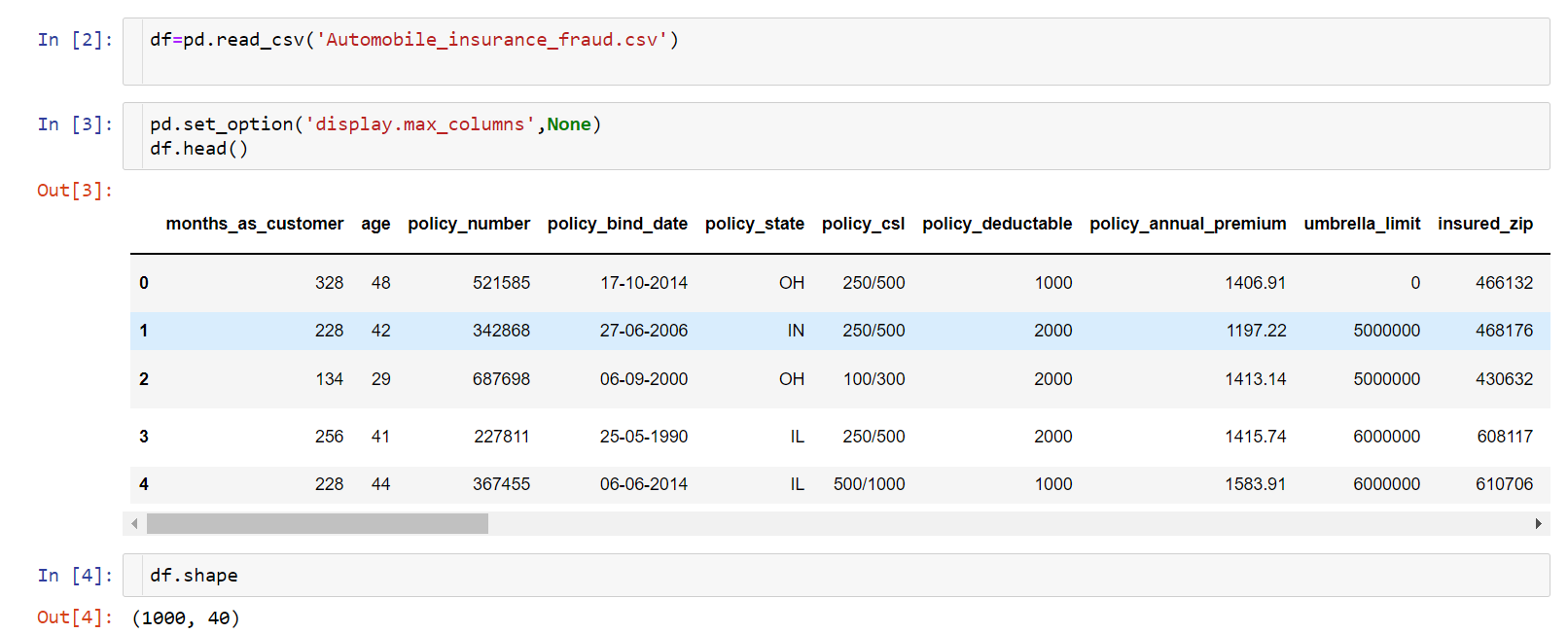
Business case:

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, you are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

* **Data Analysis:**



There are 1000 rows and 40 columns in the dataset.

The Independent Feature columns are:

Months as customer: Number of months for which the person has been a customer

age: Age of Customer

policy number: Identification number of policy

policy bind date: Time period between effective date of coverage and policy issuance.

Policy state: State where policy is active

Policy csl: Policy Combined single limit

Policy deductable: Amount paid before the insurance company starts paying up.

Policy annual premium: The total amount of premium paid annually

Umbrella limit: Provides excess limits and gives additional excess coverage

Insured zip: Zip Code of the Insured address

Insured sex : Gender

Insured education level: Education Background of Insured

Insured occupation: Occupation of Insured

Insured hobbies: Hobbies of the Insured

Insured relationship: Relationship of the Insured

Capital-gains: Capital Gains made from insurance

Capital-loss: Capital Loss incurred

Incident date: Date on which Incident Occured

Incident type: Type of Incident

Collision type: Type of collision

Incident severity: Severity of Incident

Authorities contacted: Whether authorities were contacted

Incident state: State where incident occurred

Incident city: City where incident occurred

Incident location: Location of incident

Incident hour of the day: Time of the day when incident occurred

Number of vehicles involved: Number of vehicles involved in incident.

Property damage: Whether there was property damage or not

Bodily injuries: Severity of bodily injuries

witnesses: Number of Witnesses

Police report available: Whether police reports are available

Total claim amount: Total amount of claim

Injury claim: Injury Claim amount

Property claim: Property Claim amount

Vehicle claim: Vehicle Claim amount

Auto make: Make of Vehicle

Auto model: Model of Vehicle

Auto year: Year of Vehicle Manufacture

Target Variable :

Fraud Detected : Whether fraud reported as Yes or No.

**Next step in data analysis is checking the unique values present in**

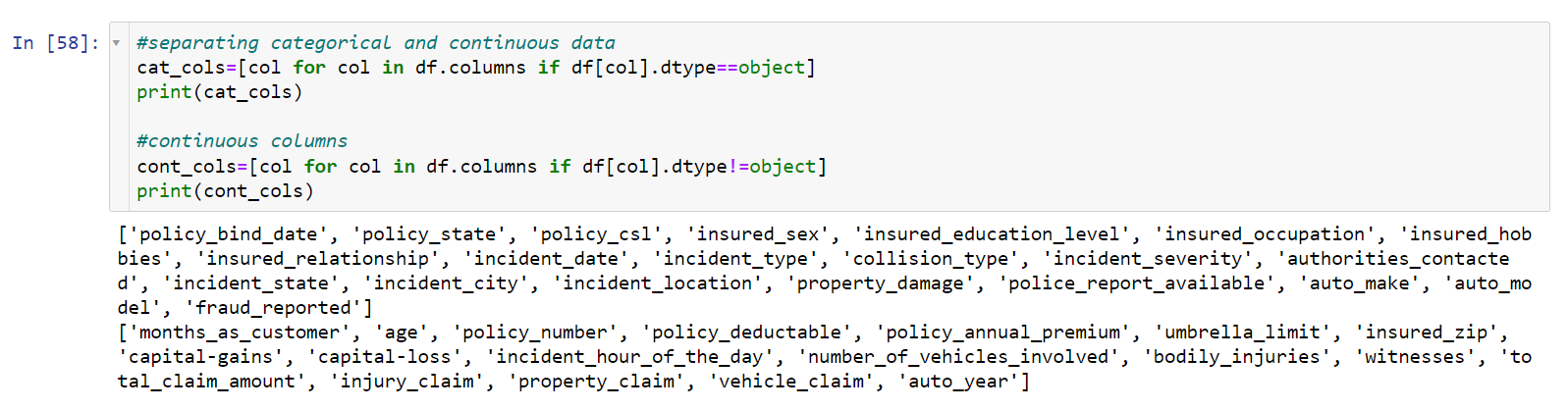
**data:**

****

By applying this method, we can find unique values in each column in dataset.

\_c39 has no usable data present. Other columns appear to have no null values. So we will drop it.

**Separating categorical and continuous data:**

****

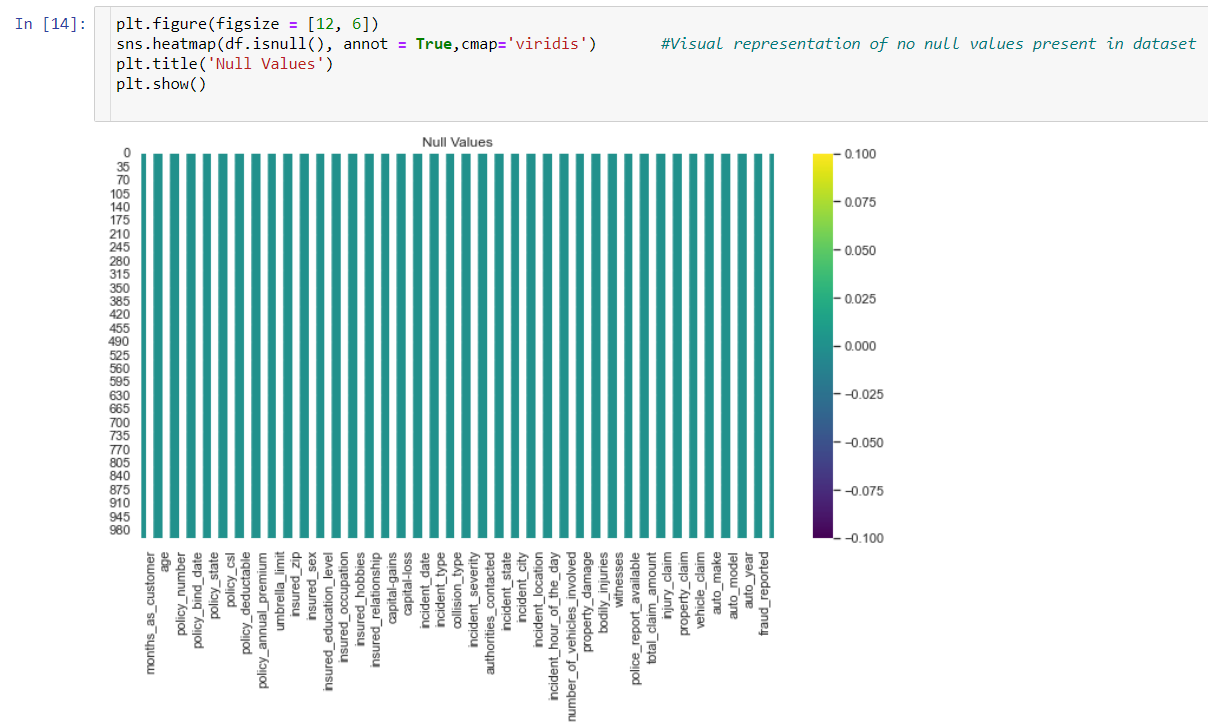
Separating categorical and continuous data will help in visualizing and

analyzing the data.

Continuous data : Data with int or float datatype.

Categorical data : Data with object datatype.

**Visualizing null values present in dataset:**

****

There are no null values present in dataset.

* **Exploratory Data Analysis(EDA):**

Exploratory Data Analysis (EDA) is an approach of analyzing

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 testing task. we can say that EDA

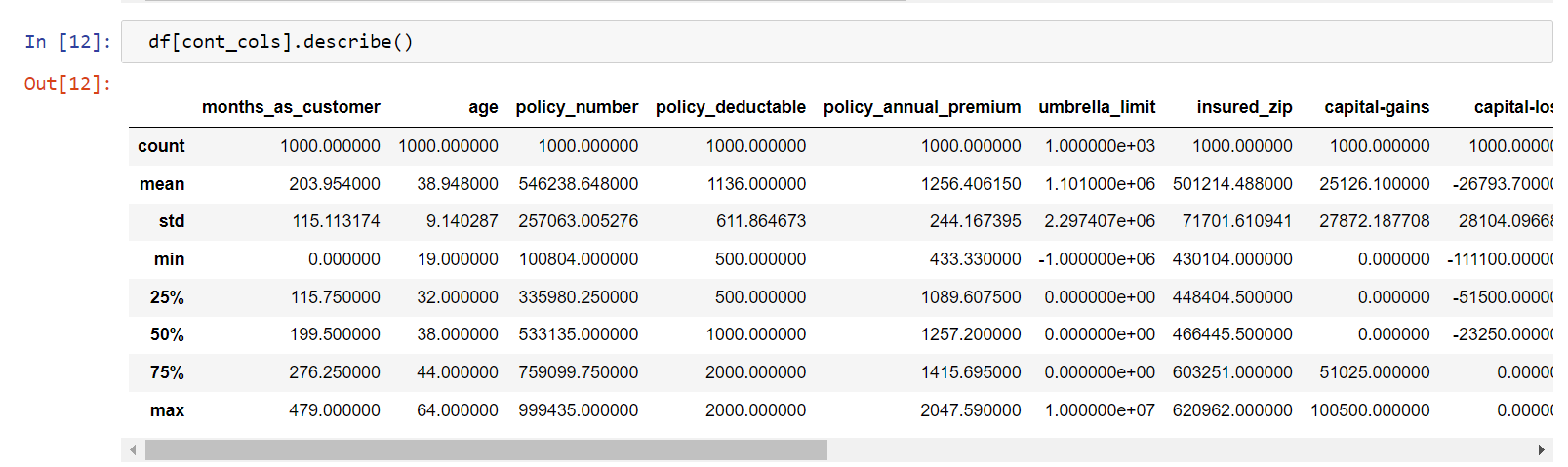
is statisticians’ way of storytelling where you explore data, find

patterns and tell insights. EDA is a phenomenon under data

analysis used for gaining a better understanding of data aspects

like: - main features of data variables and relationships that hold

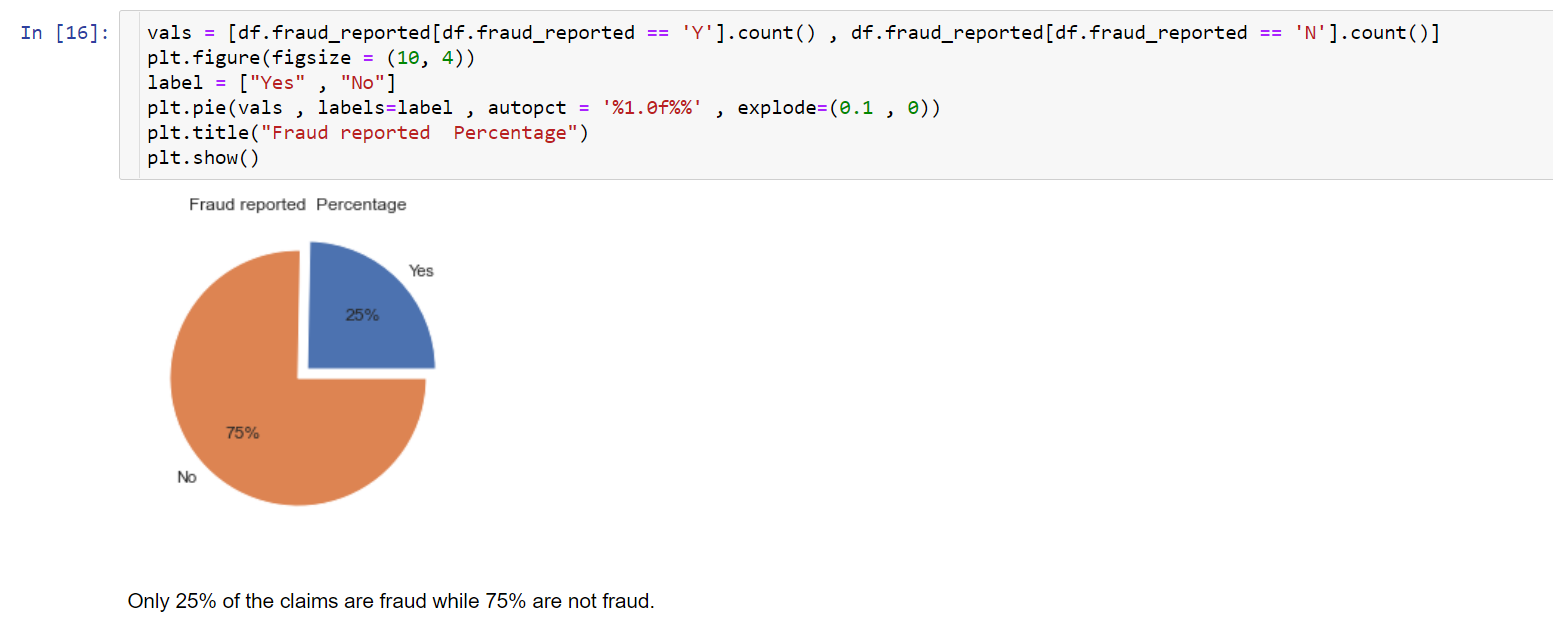
between them identifying which variables are important for our

problem.****

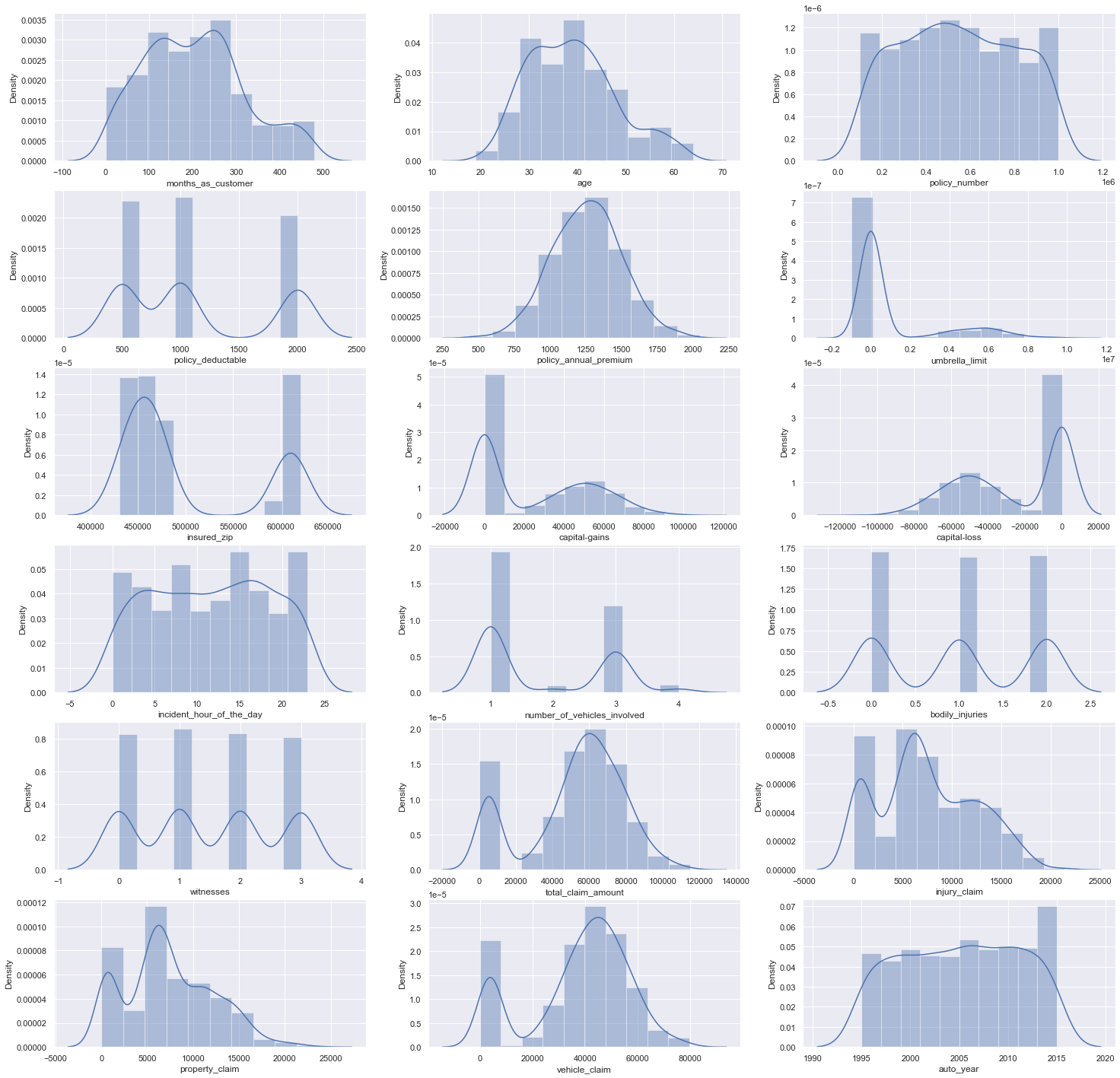
Difference in mean and 50% and considerable difference in 75% and max of columns months\_as\_customer,policy\_annual\_premium,capital-gains,total\_claim\_amount,injury\_claim and property\_claim suggests skewness in respective data distributions and presence of outliers.

**This is a Classification Problem since the Target variable / Label column ("fraud reported") has Categorical type of Data.**

**Analyzing Target Variable :**

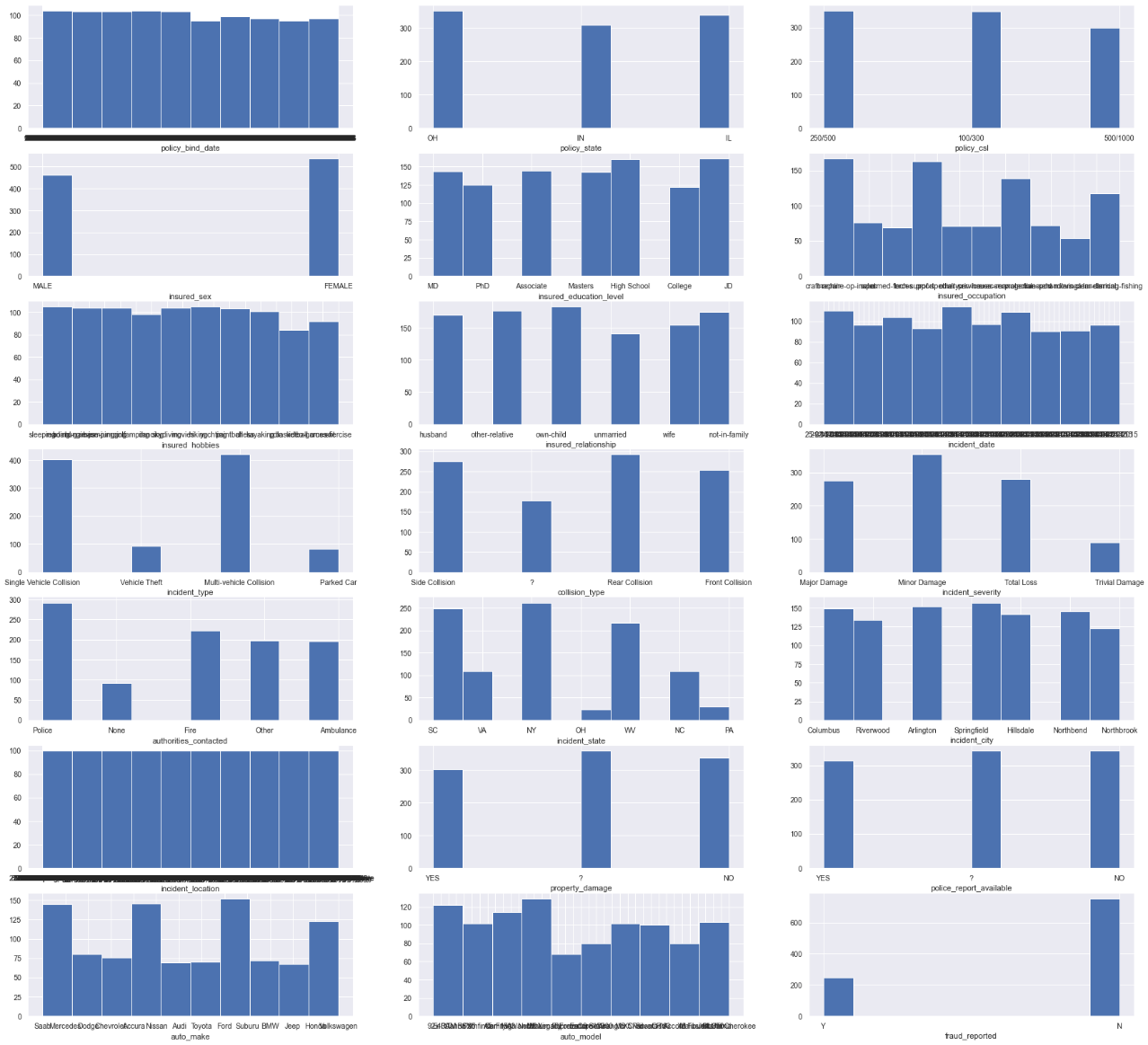


**Analyzing feature columns with continuous and categorical data:**

****

A lot of the graphs are almost perfectly bell shaped like policy\_annual\_premium, vehicle\_claim, total\_claim\_amount, etc.

There are some categorical data too in these columns, like number of vehicles involved, bodily injuries, witnesses, etc

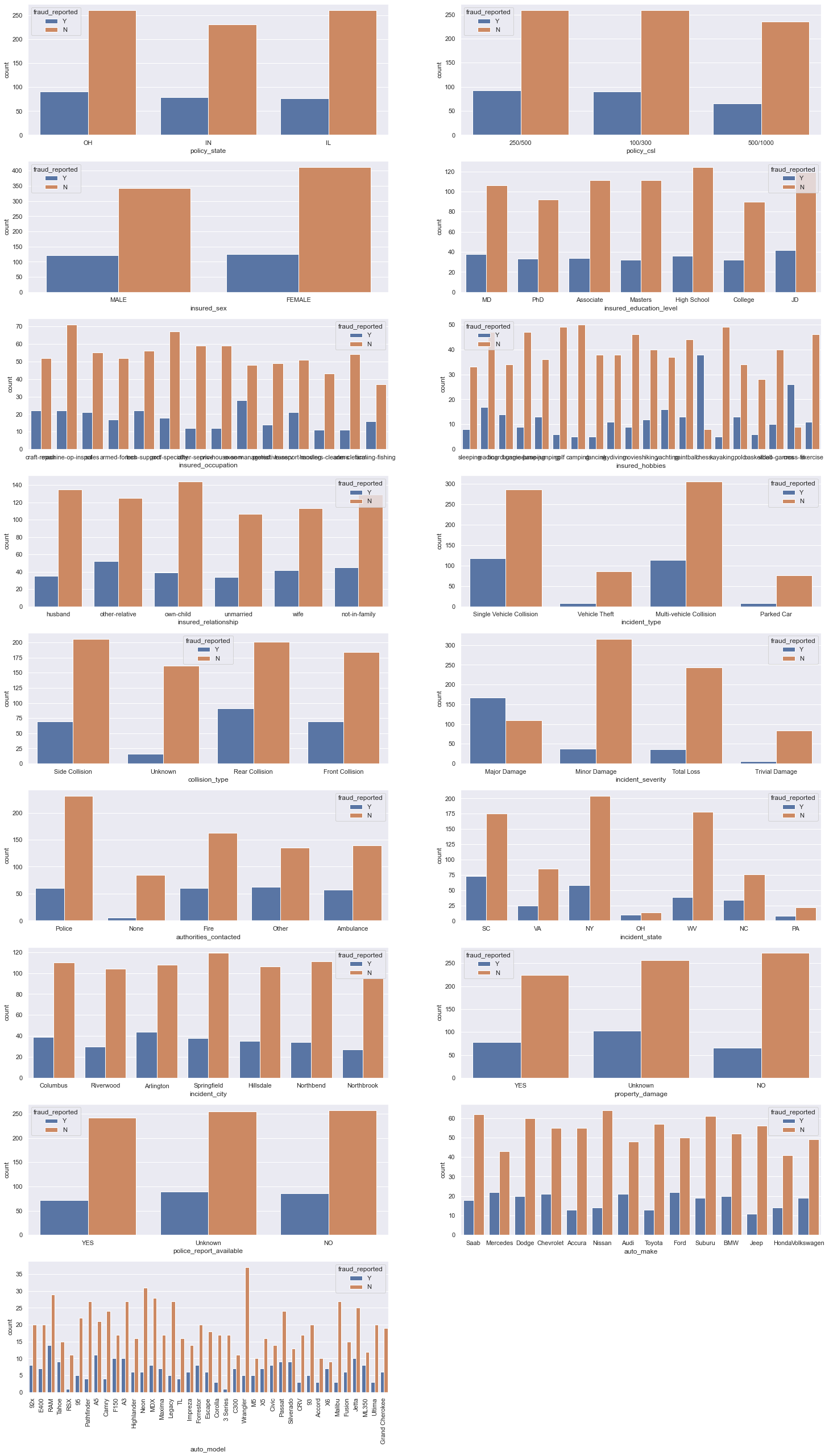


From above histogram of categorical data, we get that the columns policy\_bind\_date, incident\_date, incident\_location does not give much information.

So we will drop these columns.

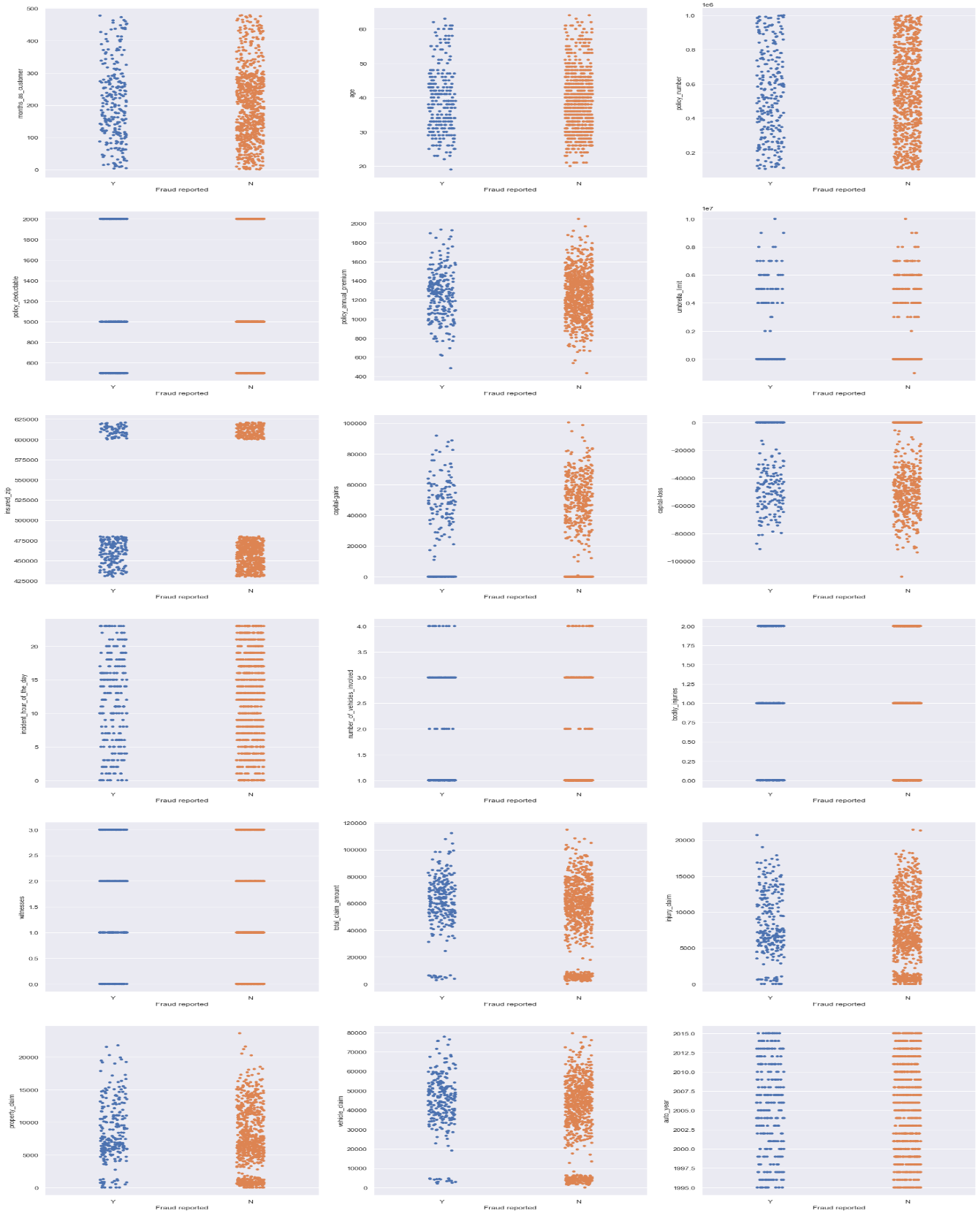


**Interpreting relationship between features and target variable:**



In incident severity we can see that fraud is reported in alot of major damage cases. Apart from that other columns and bars have low fraud reported rate.

Target Variable vs Continuous columns:



Following observations can be made from above graphs:

'age','months\_as\_customer','policy\_deductable','policy\_annual\_premium','capital-gains','capital-loss', don't seem to contribute to fraud probability.

Higher the umbrella limit, more the fraud claims are filed.

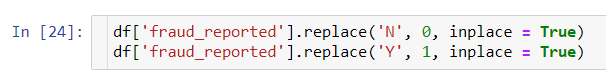
Higher the total claim amount, more the fraud claims are filed.

Higher the injury claim amount, more the fraud claims are filed.

Higher the property claim amount, more the fraud claims are filed.

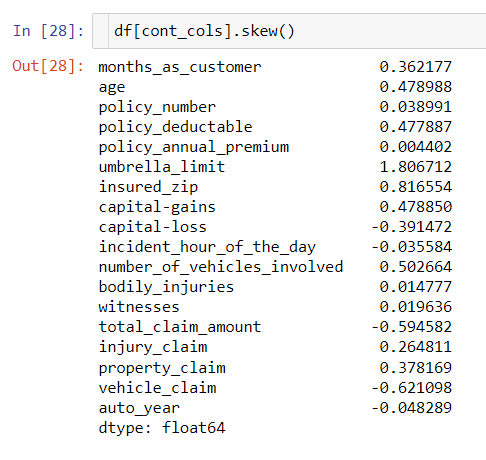
Higher the vehicle claim amount, more the fraud claims are filed.

Changing Target variable from categorical to numerical using replace function.

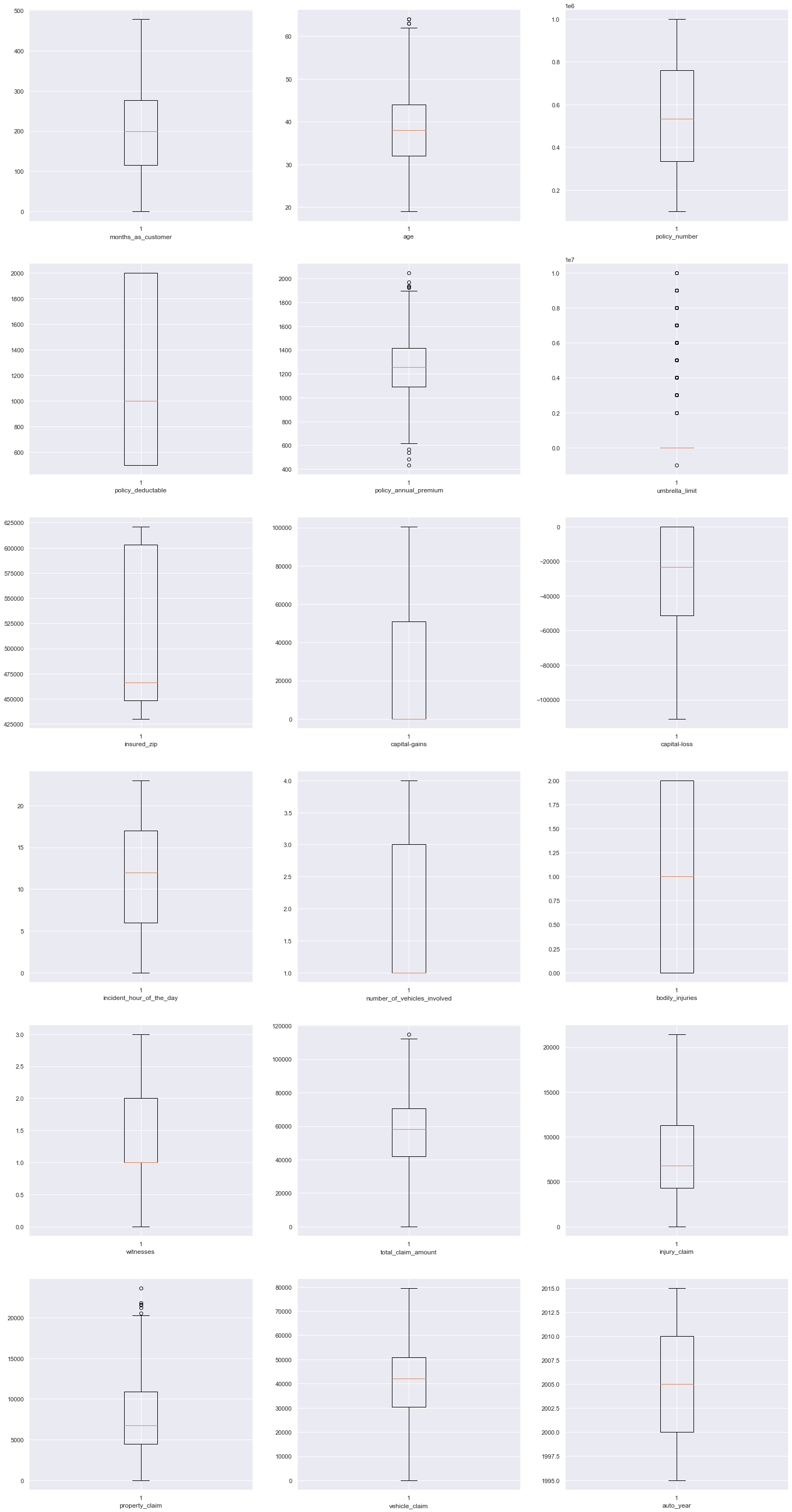


In fraud\_reported column, mapping ‘N’ as 0 and ‘Y’ as 1

**Checking for outliers and skewness and its treatment:**



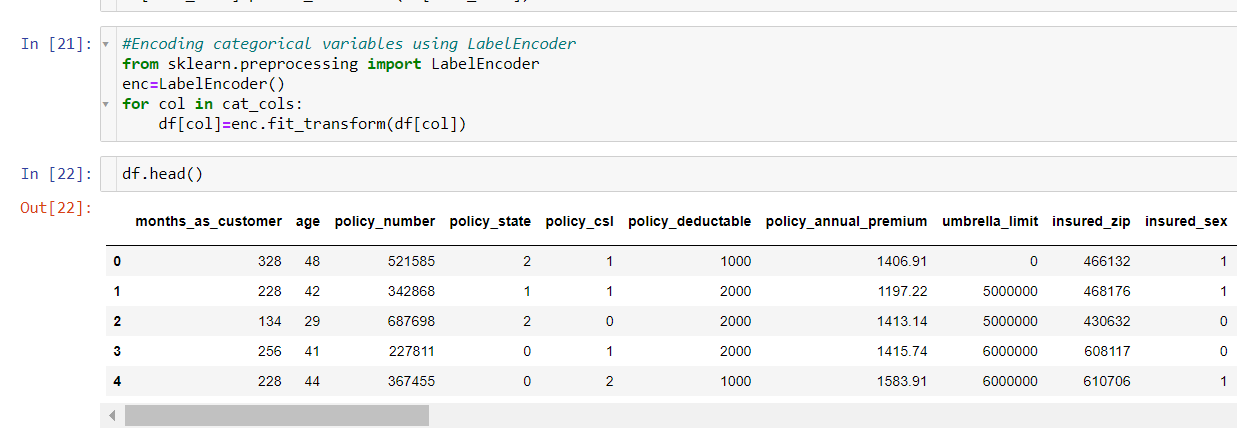
There no need to treat skewness as values for each column are within normal values.



There are some outliers present, but are acceptable. No need for outlier treatment.

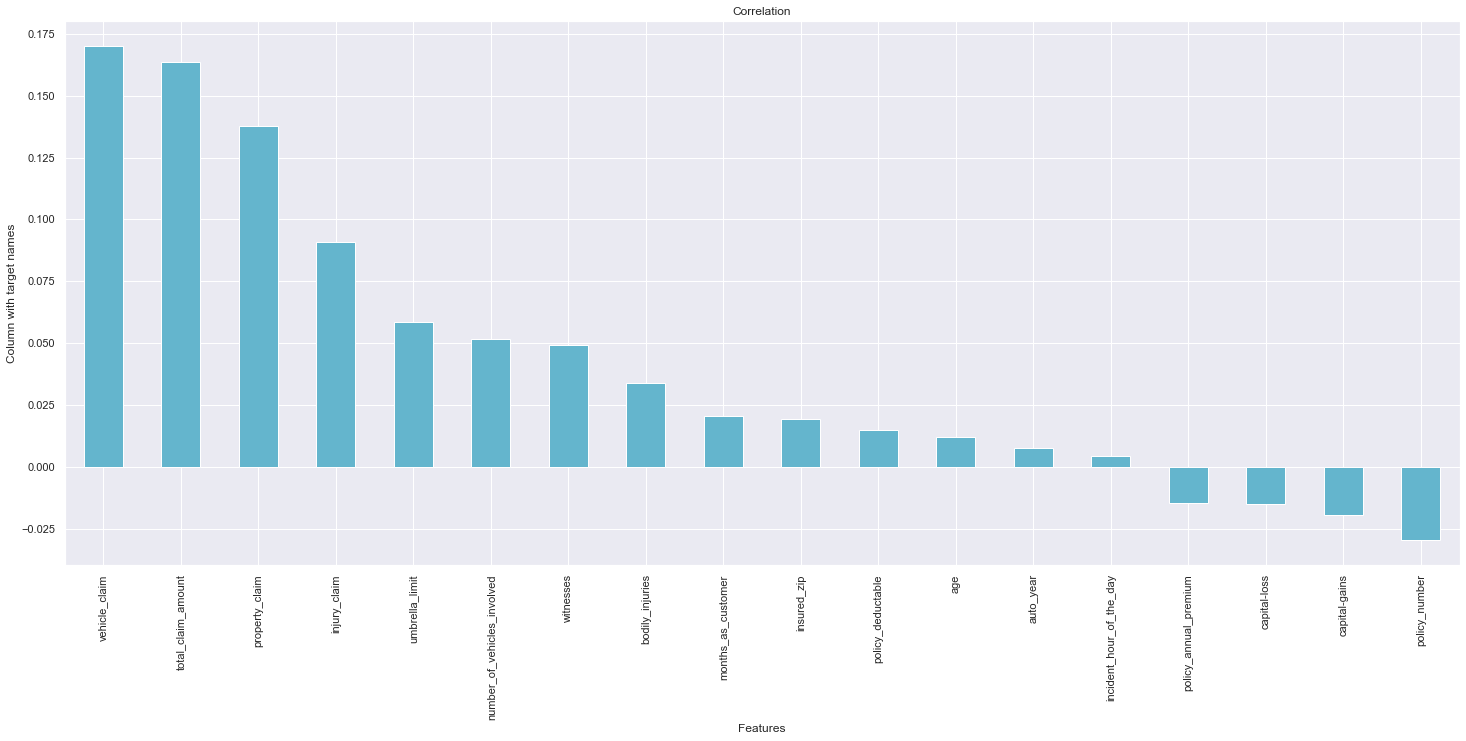
Encoding categorical variables using LabelEncoder:

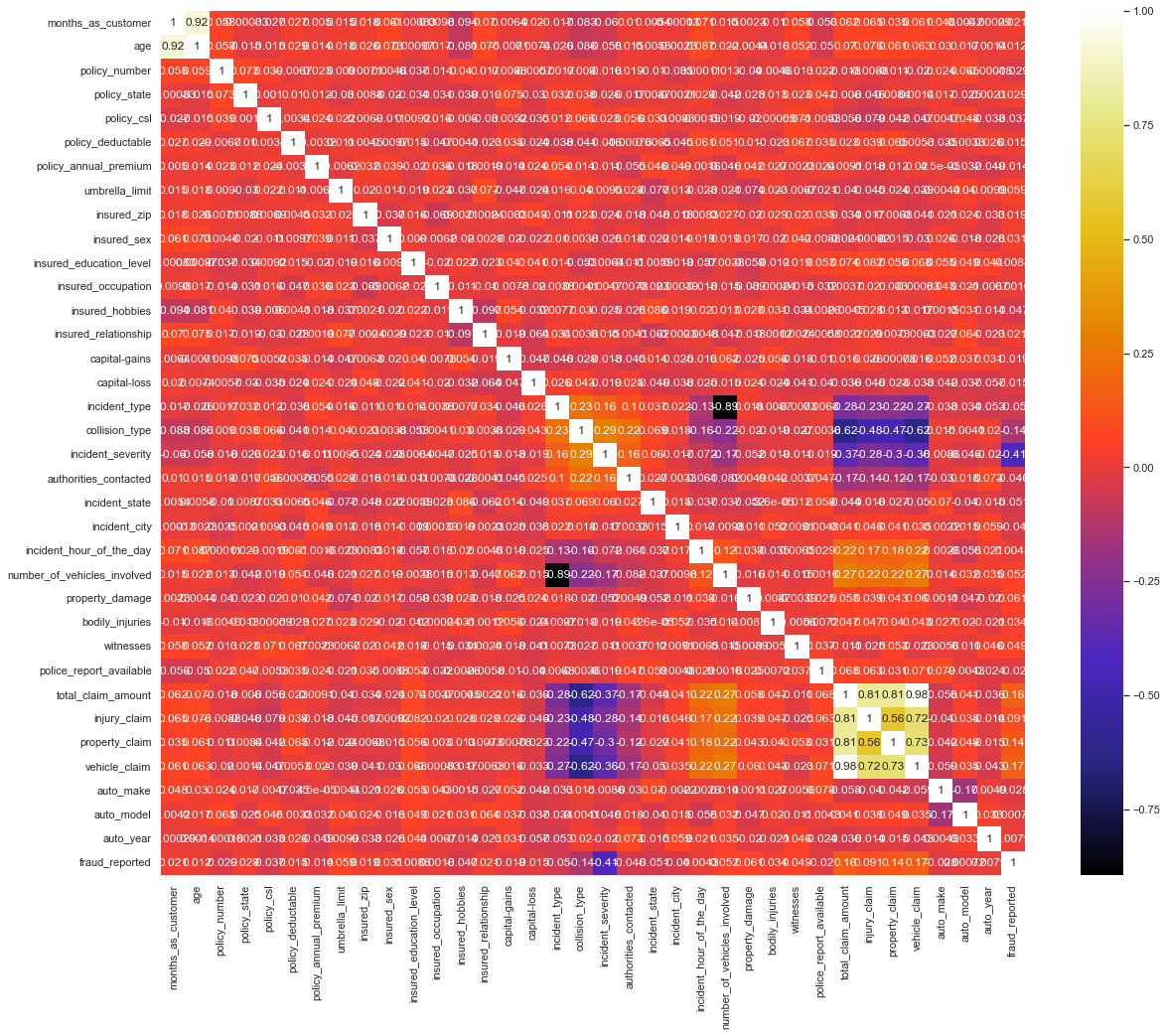
Most of the statistical models cannot take Objects / Strings as input they only takes numbers as inputs, with LabelEncoder () it is possible to categorize the string into Numbers as 1,2,3 and so on.



**Finding Correlation:**

Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr() function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure.

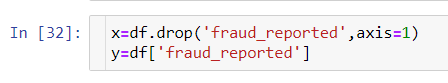


Here our EDA is complete. We now move to next step that is model building.

* **Building Machine Learning Models**

**Separating Feature and Target column :**

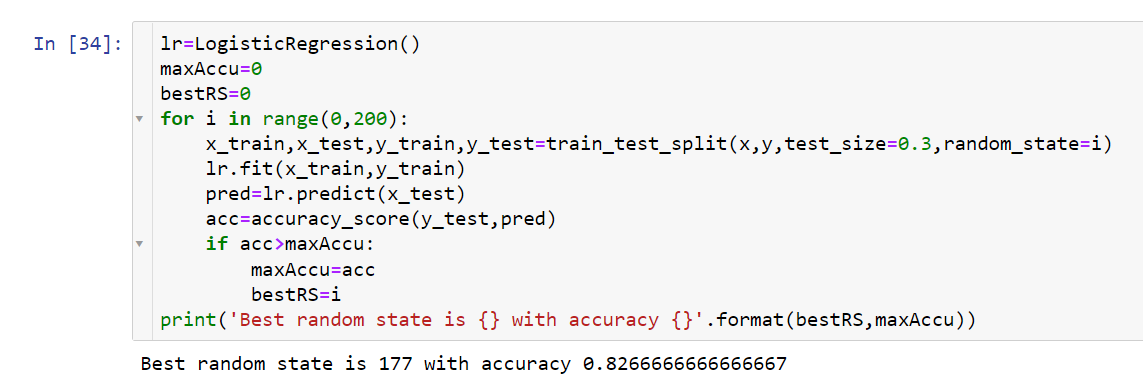
It is necessary to separate the independent/Features column into a variable (x) and target column into a variable (y). here we have to separate all columns in x Data Frame (variable) and target variable in y Data Frame (variable).



**Finding best random state:**

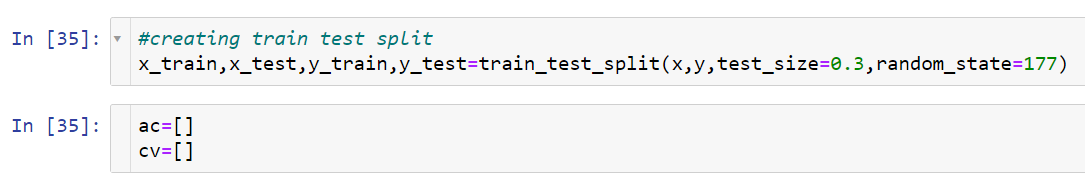
The random state parameter is used for initializing the internal random number generator, which will decide the splitting of data into train and test indices in your case.

We will use logistic regression for finding best accuracy by iterating it over random state values between 0 to 200.



**Splitting the data for training and testing:**

In ML the separated data is split into 4 parts for Training and Testing of features (x) and for Training and Testing of Target (y) like x\_train, x\_test, y\_train, y\_test. It is possible through a inbuilt library of sklearn’s train\_test\_model.



I will store the accuracy score values and cross validation scores in separate lists for cross validation of models in later part.

**Training the Models:**

To find the best model it is necessary to train 3-4 models, In the same way I have trained. I will use following models for training and checking the accuracy of each model.

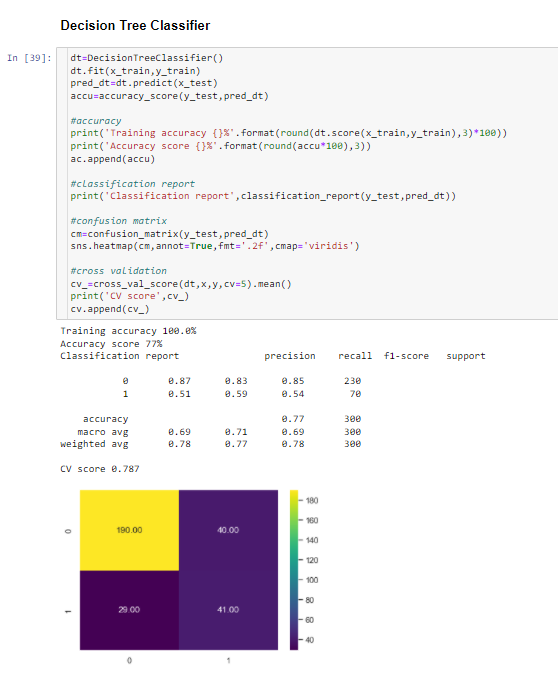
1. lr = LogisticRegresssion()
2. dt = DecisionTreeClassifier()
3. rf = RandomForestClassifier()
4. knn = KNeighborsClassifier()
5. svc = SVC()

**Model Building:**

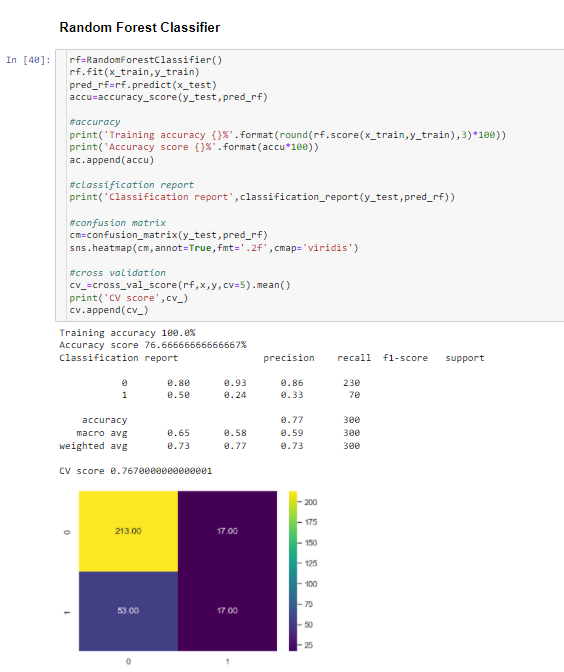
Logistic Regression Training and Accuracy-

****

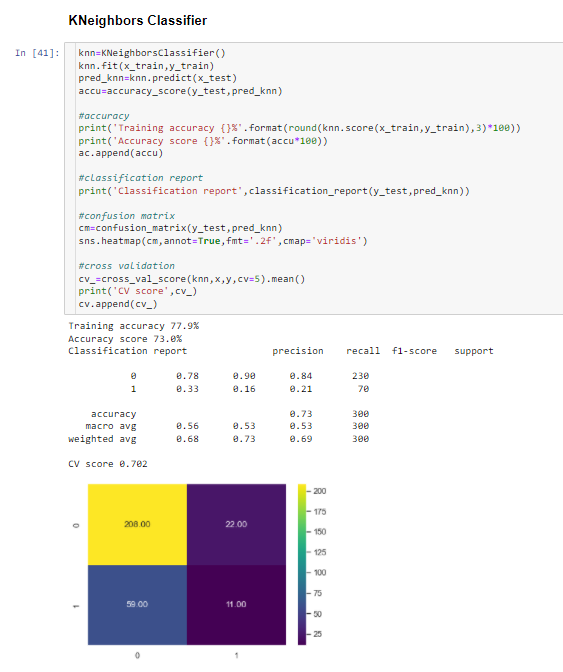
Decision Tree Classifier Training and Accuracy-



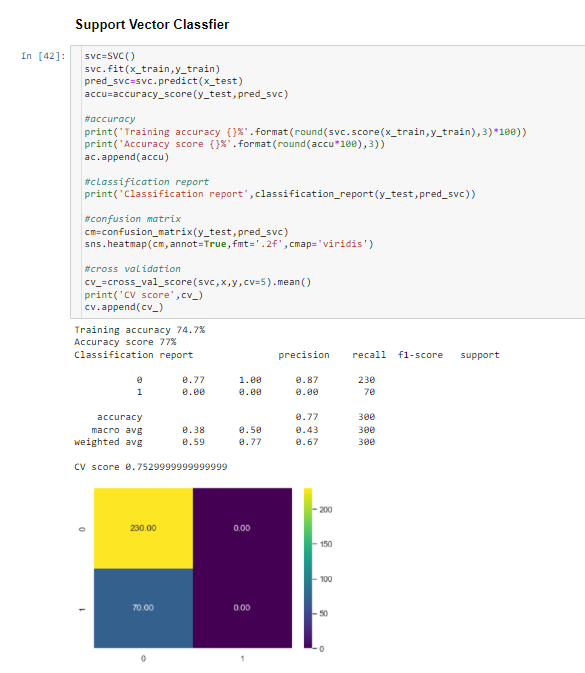
Random Forest Classifier Training and Accuracy-

****

KNeighbors Classifier Training and Accuracy-

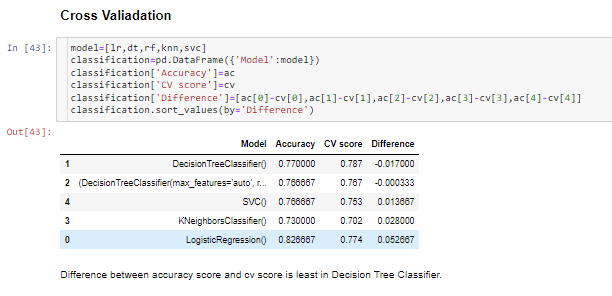


Support Vector Classifier Training and Accuracy-



**Cross Validation of models:**

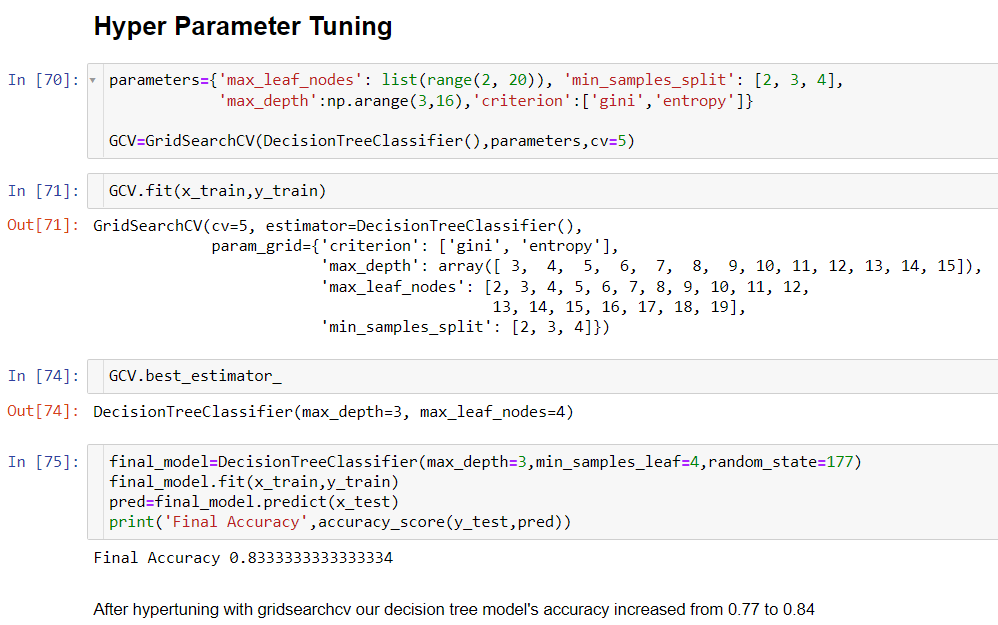
Cross validation is used to check the best model for prediction. Lower the difference between the accuracy score and cv score better the model.



From model training and cross validation, Decision Tree Classifier is best model for prediction.

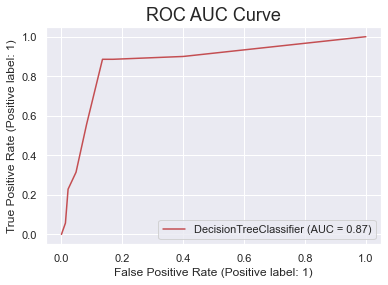
**Hyper Parameter Tuning:**

Tuning of parameters of model will increase the accuracy of model.

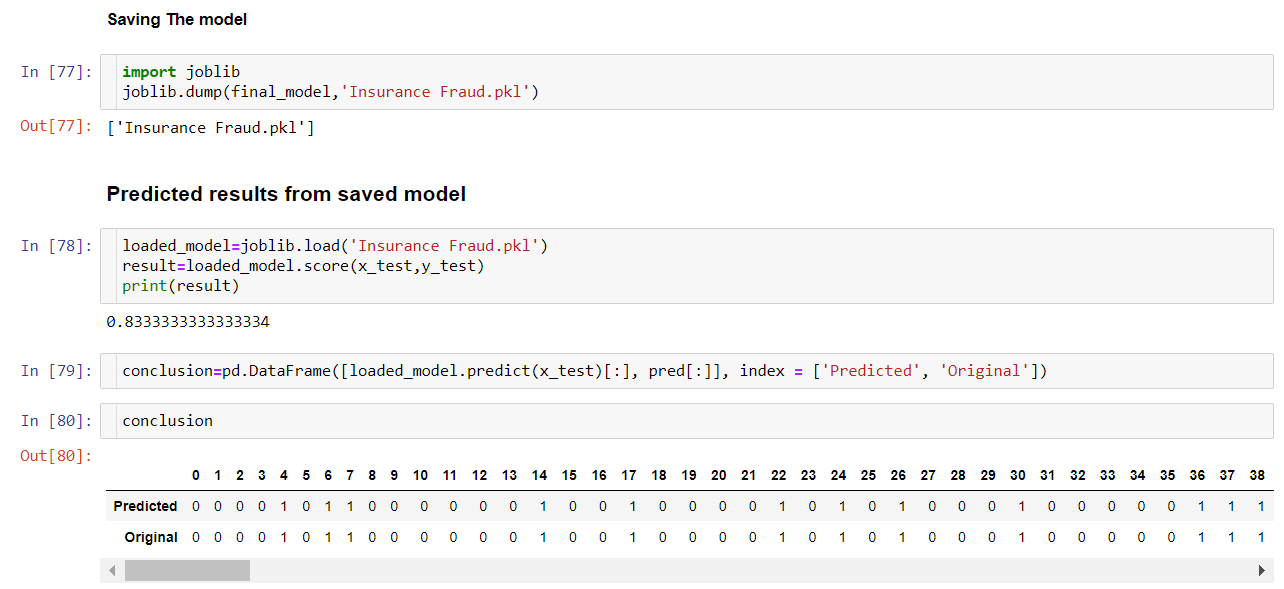


After tuning parameters accuracy of model is increased from 0.77 to 0.84.

**ROC AUC Curve:**

****

**Saving and Loading Model:**

****

* **Conclusion:**

In conclusion, Decision Tree Classifier Model is able to correctly distinguish between Fraud claims and legitimate claims with high accuracy.

The dataset had very limited data which is problematic as models show greater stability when the dataset is of a good size. However, a large set of feature columns enabled selecting a smaller feature size that provides the best accuracy for the model and obtaining results in optimal time.