**INSURANCE CLAIM – FRAUD DETECTION BLOG ARTICLE**



**Insrance Claim -Fraud Detection Status Prediction**

****

**Introduction:**

Insurance Claim is made by the requestor to the policy provider. Insurance can be made for Home, Property, Land, Accident, Car, Health, Auto etc. We can claim if our home is damage from Earth quake or got into an accident. At that moment we can do insurance claim by submitting the form. So, that we don’t give money from our pocket. For example, if we have Health Insurance, suddenly we get serious health issues, we no need to worry of money, whether we can afford the medical expenses or not. We can claim the health insurance by submitting the insurance claim form. All the medical expenses can be done in the health insurance claim. In this article we see the Auto insurance claim, to predict where the insurance claim is fraudulent or not.

**1)Problem Statement:**

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

**2)Data Analysis:**

Data Analysis is the process of cleaning, transforming, pre-processing, modelling data to get a useful information and make a prediction.

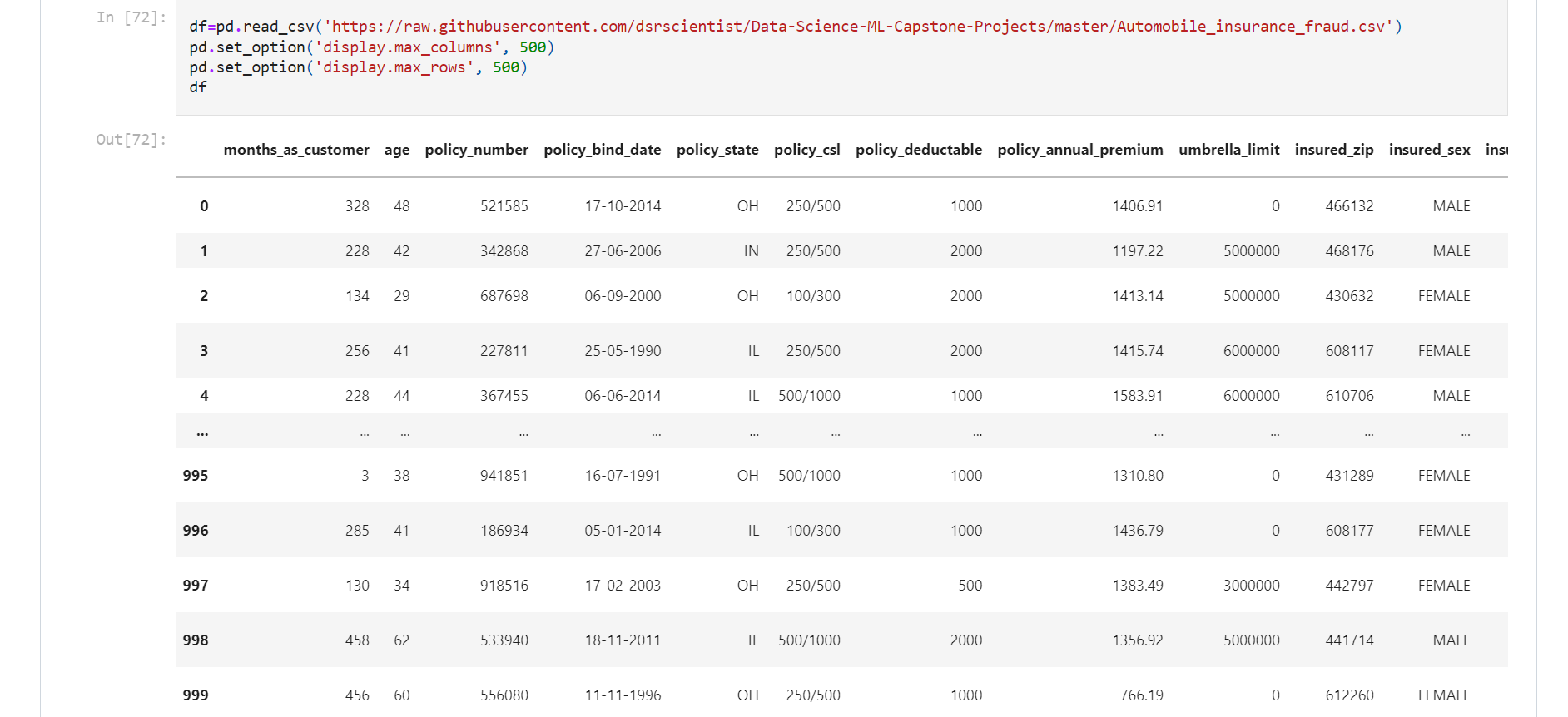
Data Analysis can consist of Text, Statistical analysis, Description, Diagnostic, Predictive Analysis.

Data Analysis plays a very important role, before building the model, we do the analysis of data, we can replace any missing value present in data, we can remove unnecessary column from dataset. We can convert categorical data to numerical data using Label Encoder/Original Encoder. Data Analysis play an important role in Decision making, improve accuracy and help in scientific approach to give a good insight using a visualization technique.

**Importing Necessary Libraries:**



**Loading the dataset:**



There are 1000 rows and 40 columns present in Insurance claim dataset. It contains both categorical and Numerical data.

Data set has missing values and values with "?" sign.

**Data Pre-processing and Data Cleaning:**

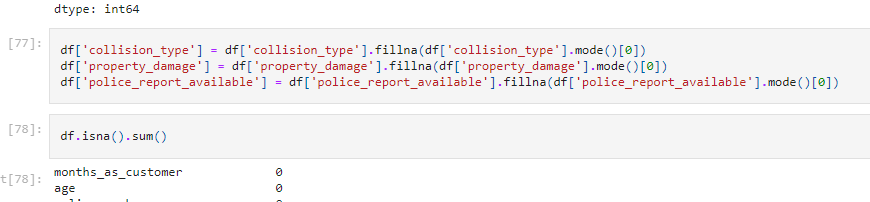
In data pre-processing we check the statistical summary, data info, Unique value, column name, shape of data, value count.

In Data cleaning, we drop the unnecessary column, fill the missing value with mean/median if it is numerical data. If it is categorical data, we can replace with mode.

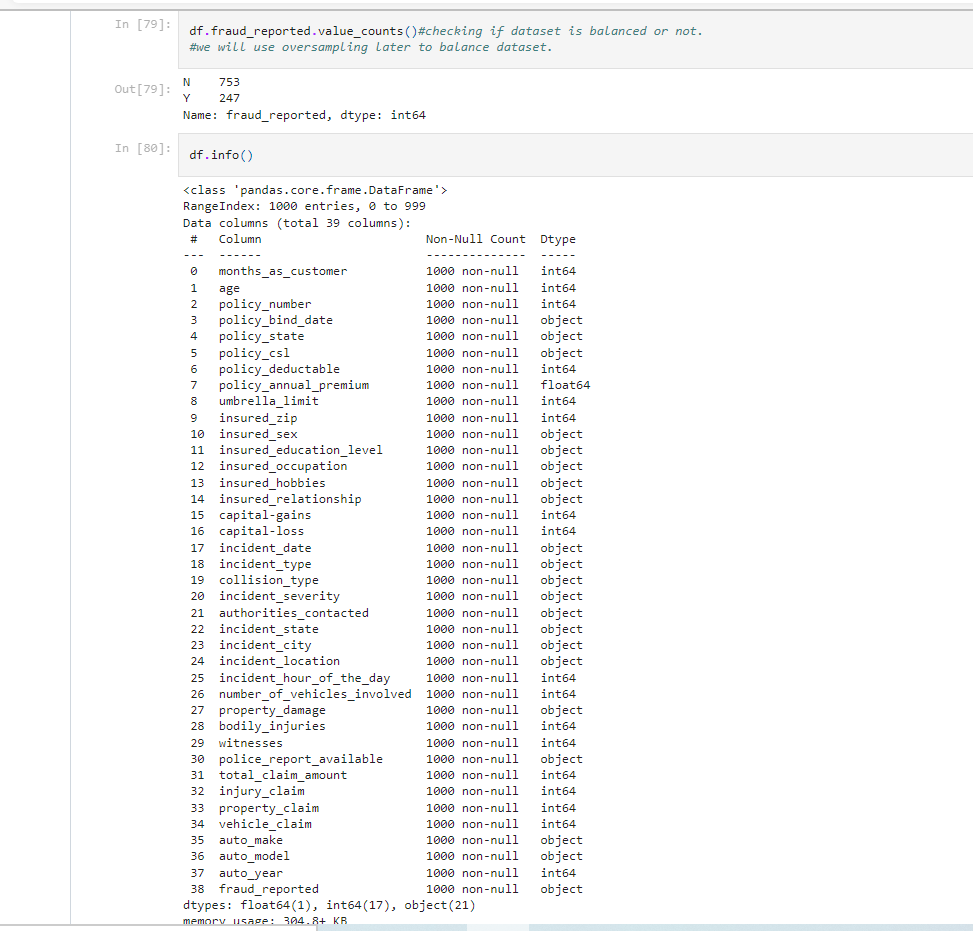
**Checking data for null entries and treating ‘?’ values:**

****

Using df.drop() method to drop null entries and df.replace() method to replace ‘?’ entries with np.nan.

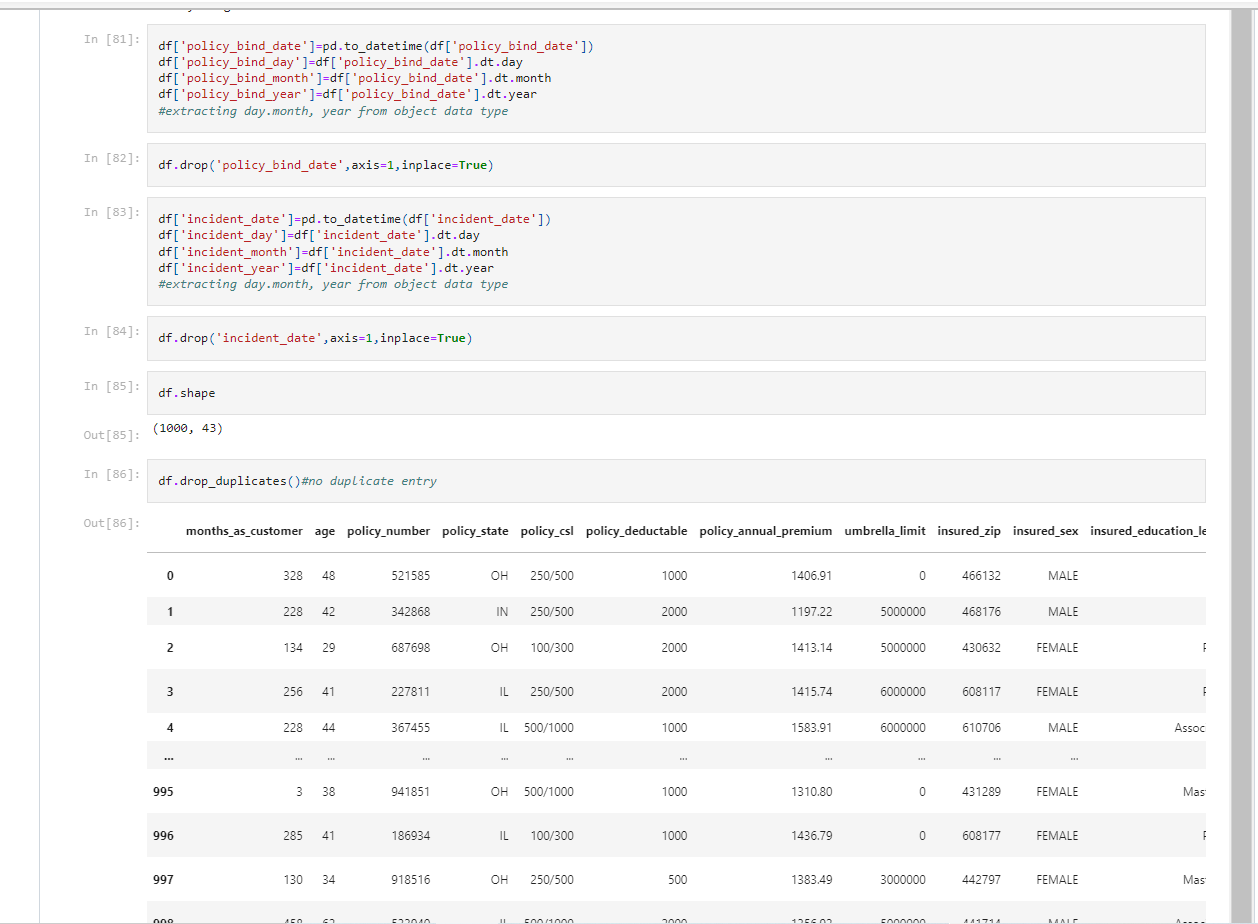


Replacing categorical columns with mode.



Checking target column for imbalanced set and then checking datatype and count of each column using df.info().

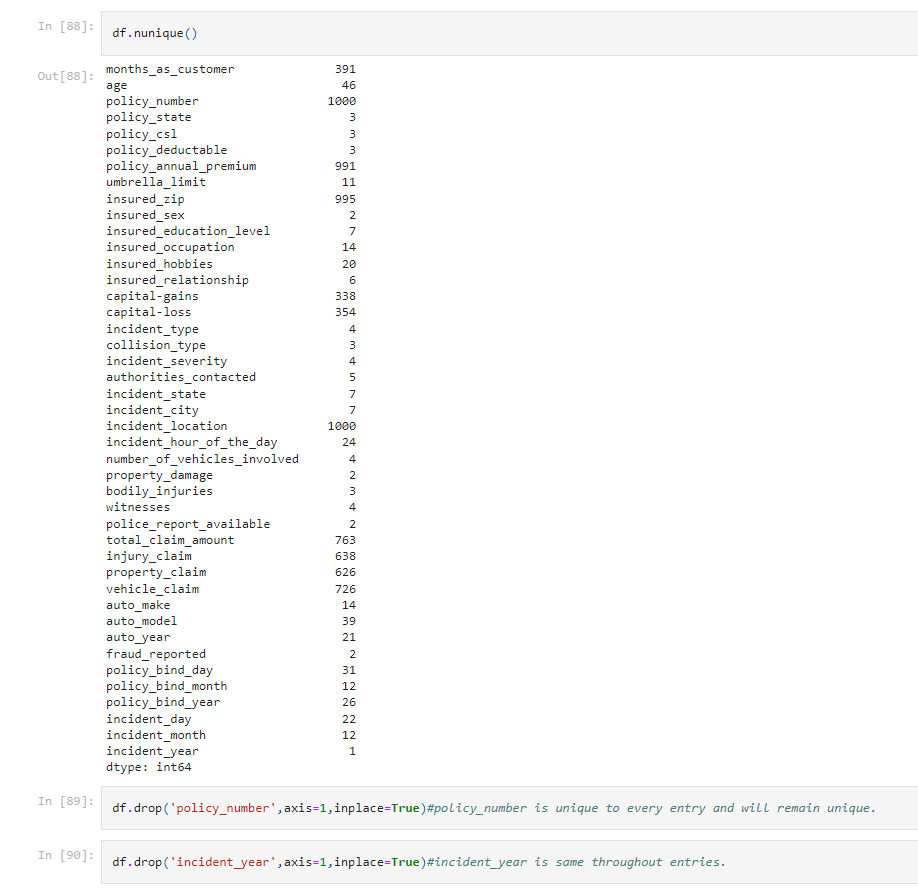
We can see that a greater number of people reported No for Fraud Claim. No-752. Yes-247. Still few people reported there a fraud claim.



In the Incident Date column, it is considered as Object data type. Which is not correct. So, we have converted date time data type and we have separated into Day, Month and Year.

In Policy Bind date, the same way we have converted into datetime data type and separated into Day, Month and Year. Then we can drop the Policy bind date column.

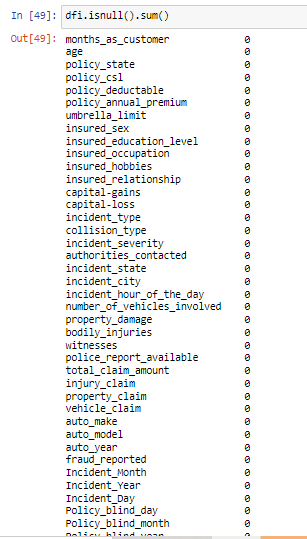
Checked and dropped for duplicate entries using df.drop\_duplicates().



We can drop the unnecessary column, policy number, incident year as , they consist unique values all throughout the entries.



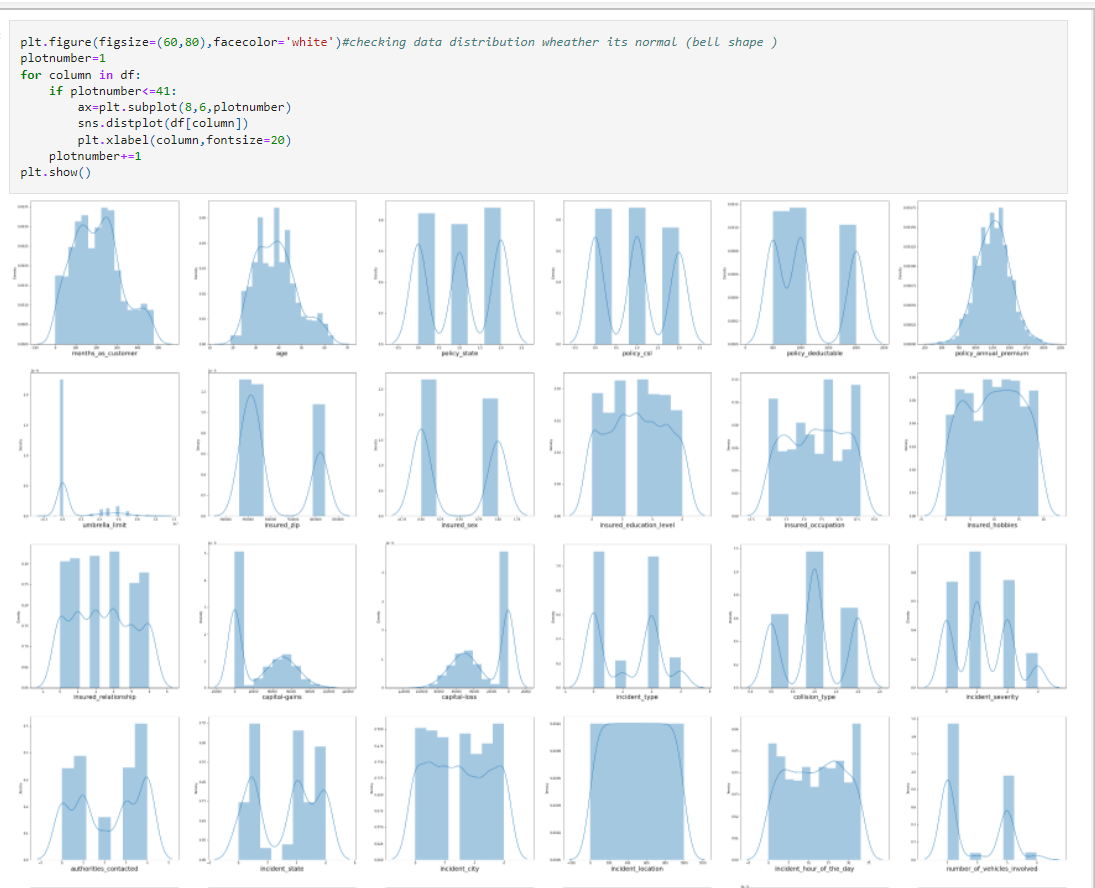
Then we encoded all the categorical features using label encoder.

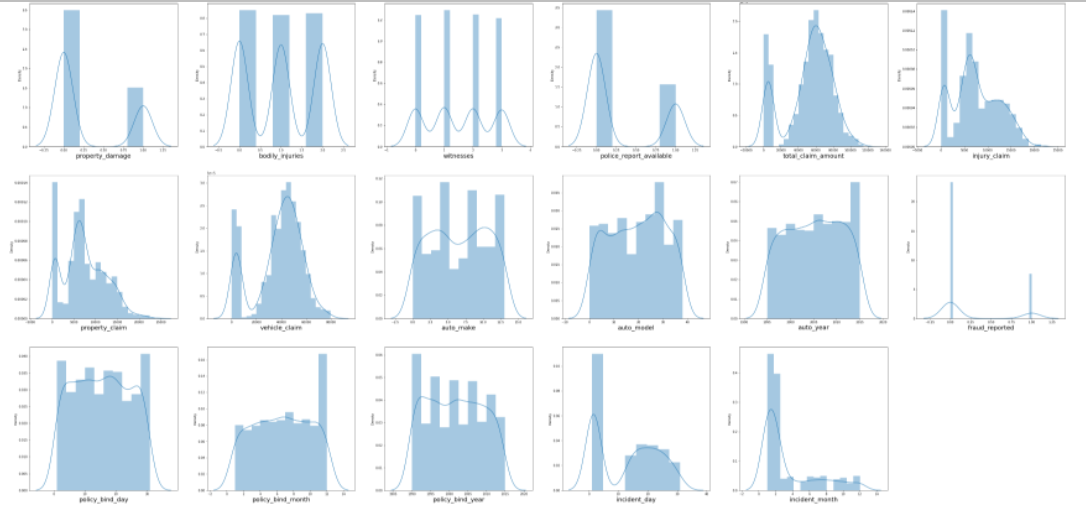


We can see there is no null value present in data.



There is no null value present in the data. We have visualized using the heatmap.

****

****

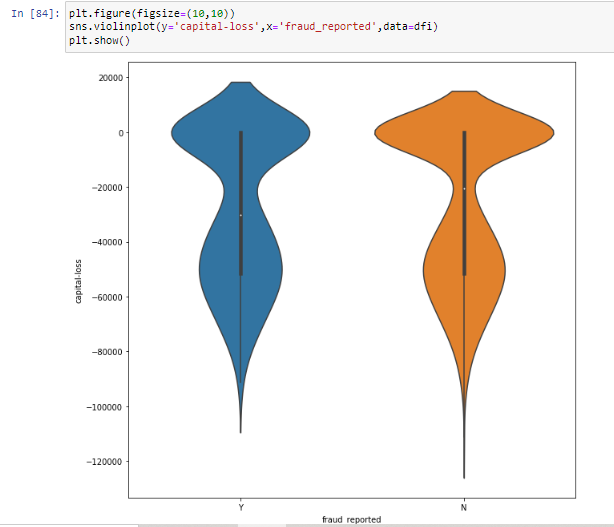
All numerical features seems to be in normal distribution except capital\_gain,capital\_loss,total\_claim\_amount,injury\_claim,property\_claim,vehicle\_claim.

In the distribution plot we see that data are normally distributed and skewness also present. In insurance claim, property claim, injury claim, that data is slightly goes high and the diagram looks like a wave. We can remove the skewness to make the good accuracy.

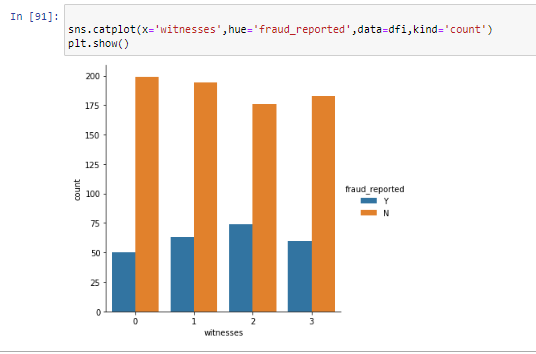
**EDA:**



Using the count plot, we have checked the number of vehicles involved with fraud\_reported. Those who have only 1 vehicle, the count is high that there is a no fraud claim. Those who are having more than 1 vehicle, there is high change of fraud claim.



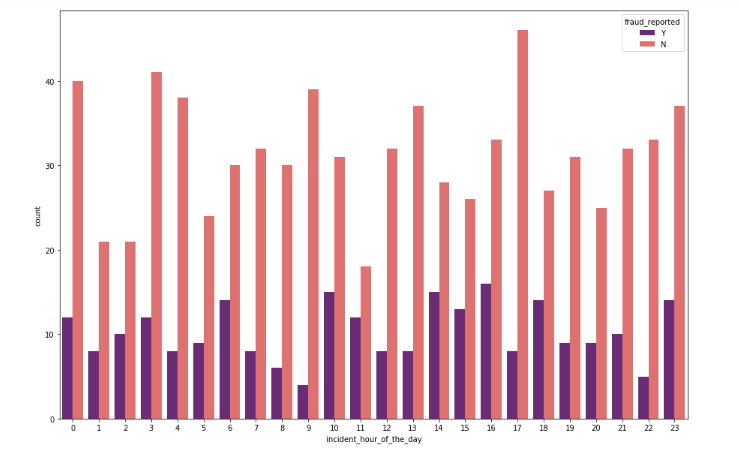
Using the violin plot we checked the Capital loss with fraud reported. We clearly see that the fraud reported in the capital loss ranges between -40000 to -60000



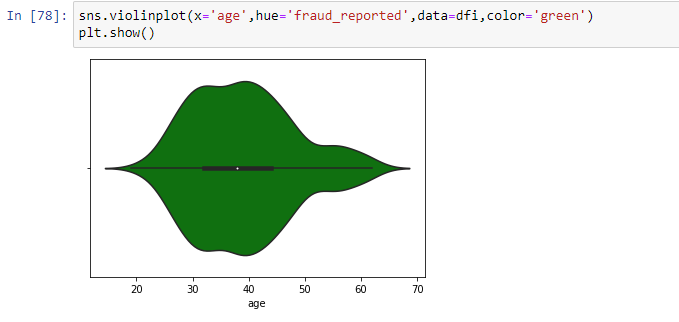
We have more than 1 witness reported the fraud claim. High count is, there is no fraud claim, few people has reported the fraud claim.



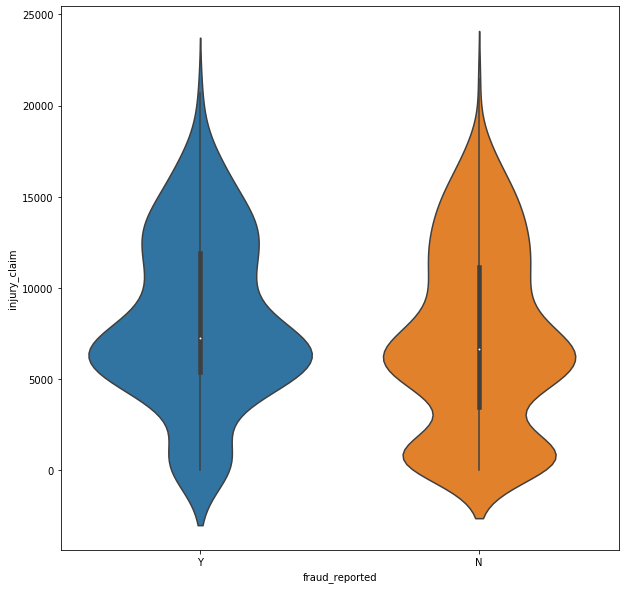
We have compared the Policy deductable and fraud reported. We see that policy deductable count 1000 has more than others.



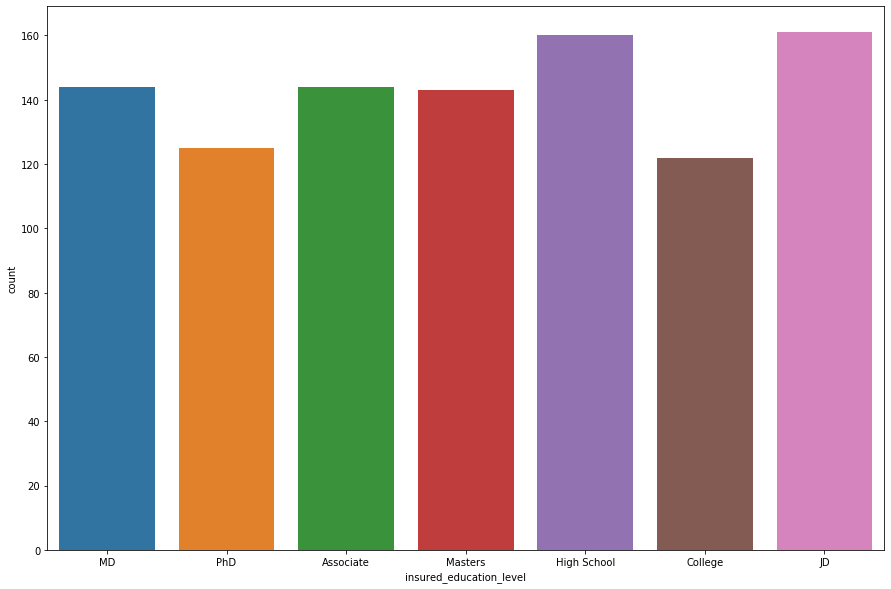
We have checked the incident hour of day; in which hour more fraud claim is happening. In the peak hour there is high fraud claim.



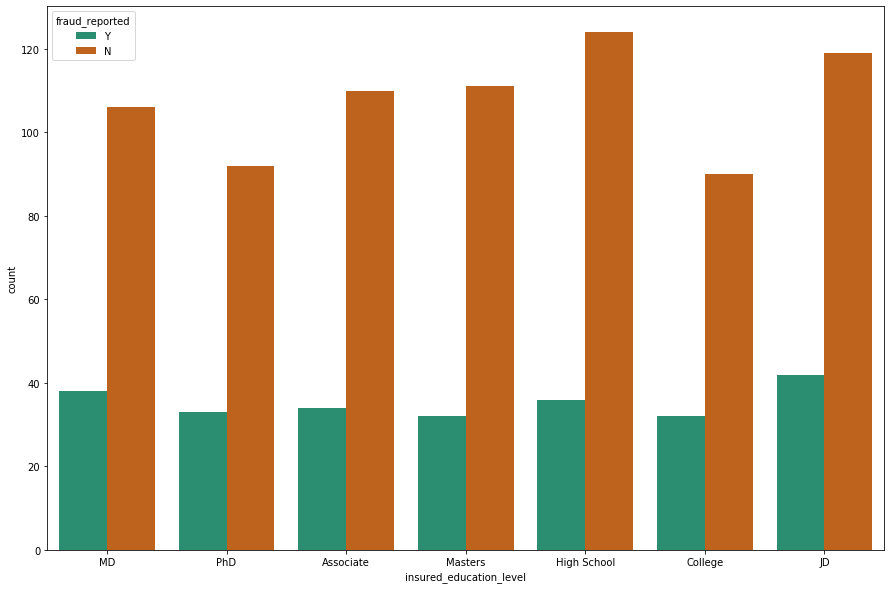
We compare the age with fraud reported, we clearly see that in the range of 30-40 years there is high number of frauds reported.



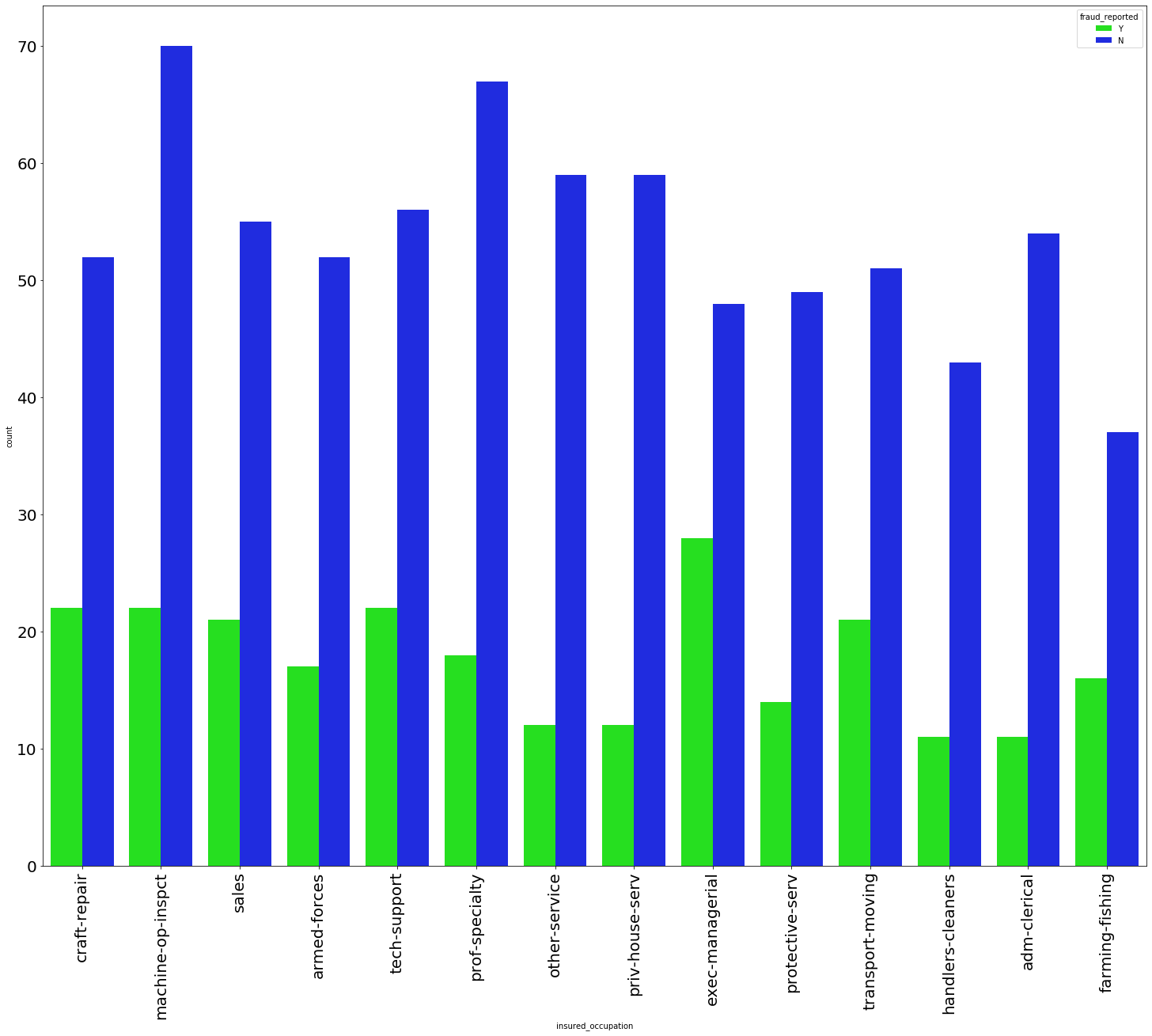
In the injury claim between 5000-10000 there is high number of frauds reported.



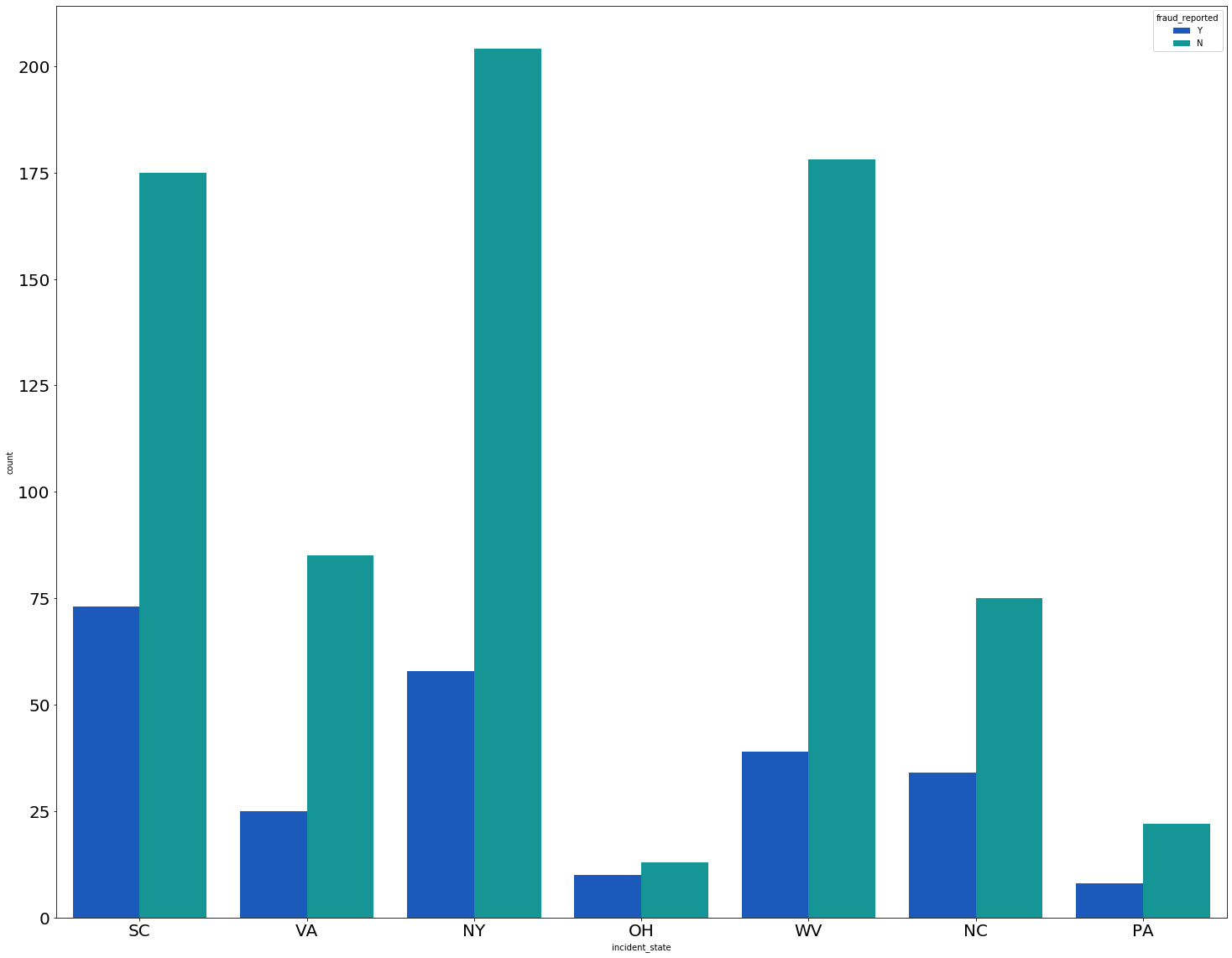
If we see the education level of insured person who has taken insurance from the company. Most of the insured person has completed JD and High School.



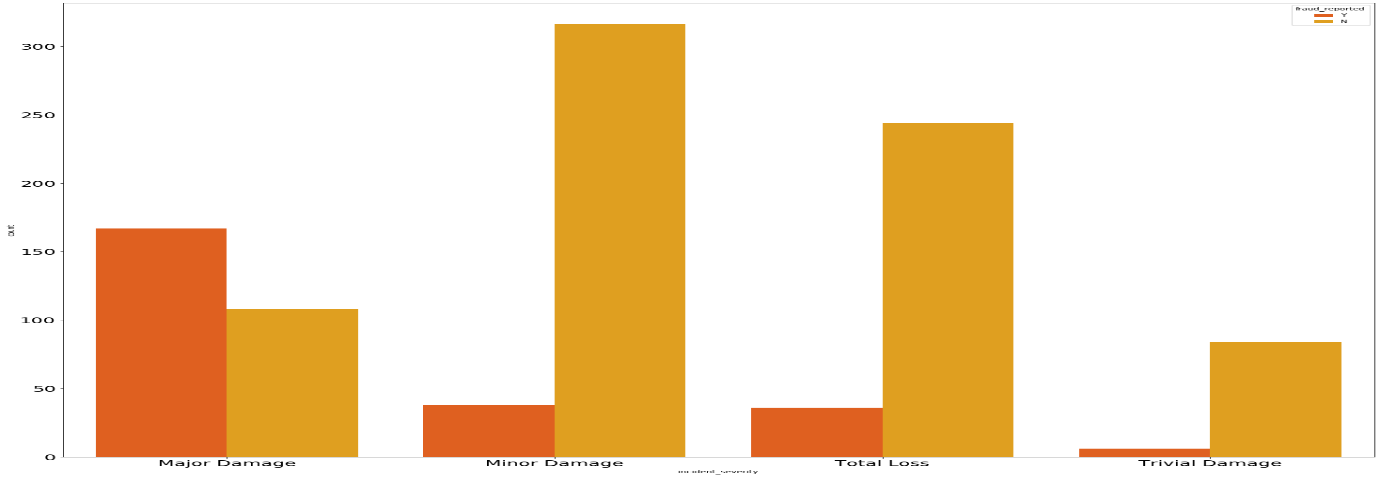
If we compare the education level with fraud reported, those have completed JD has higher chance of fraud claim.



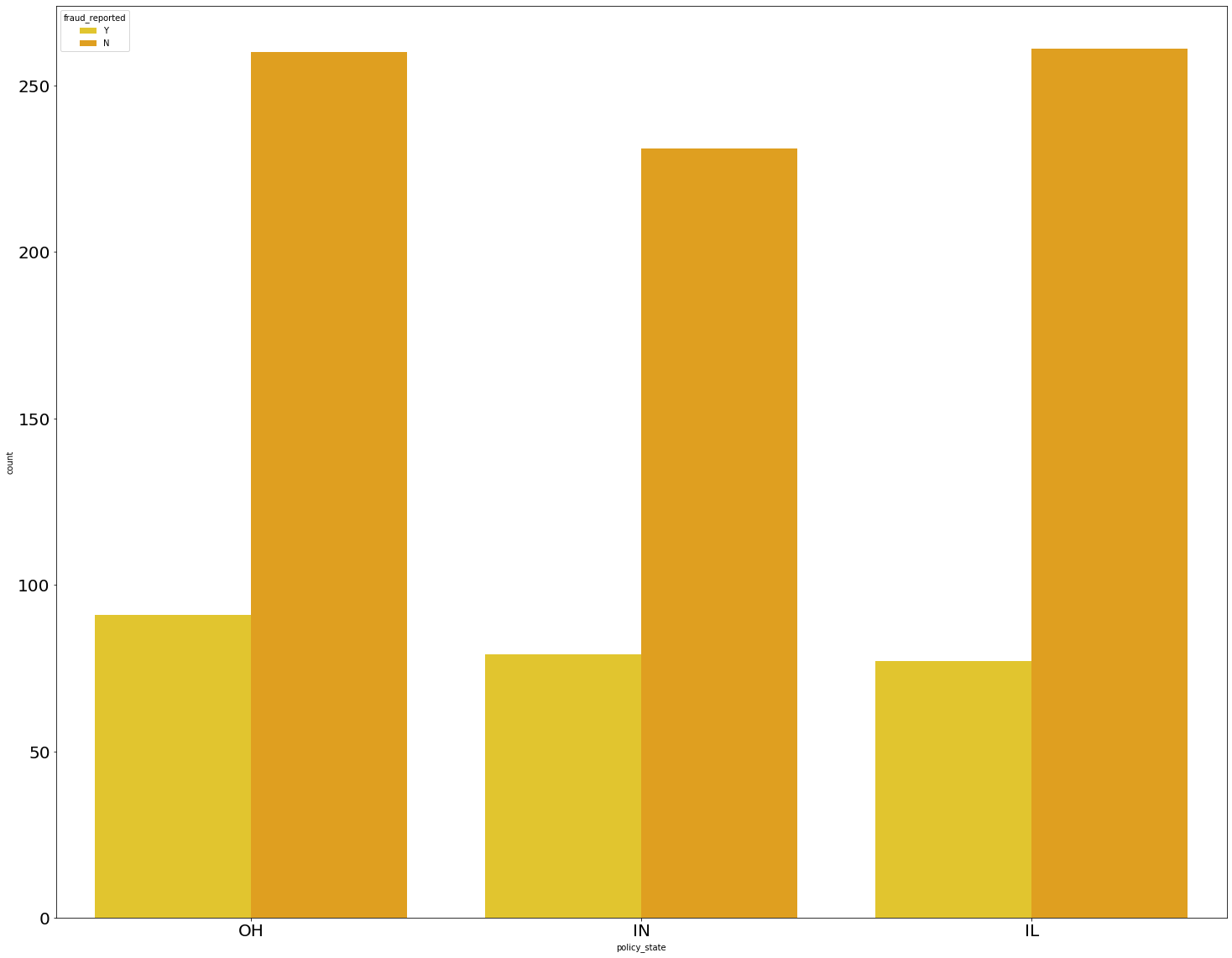
Majority of insured person occupation is Machine-on-inspect, when we checked the insured person occupation and fraud reported. The diagram shows the person who is in executive-managerial has high chance of fraud claim.



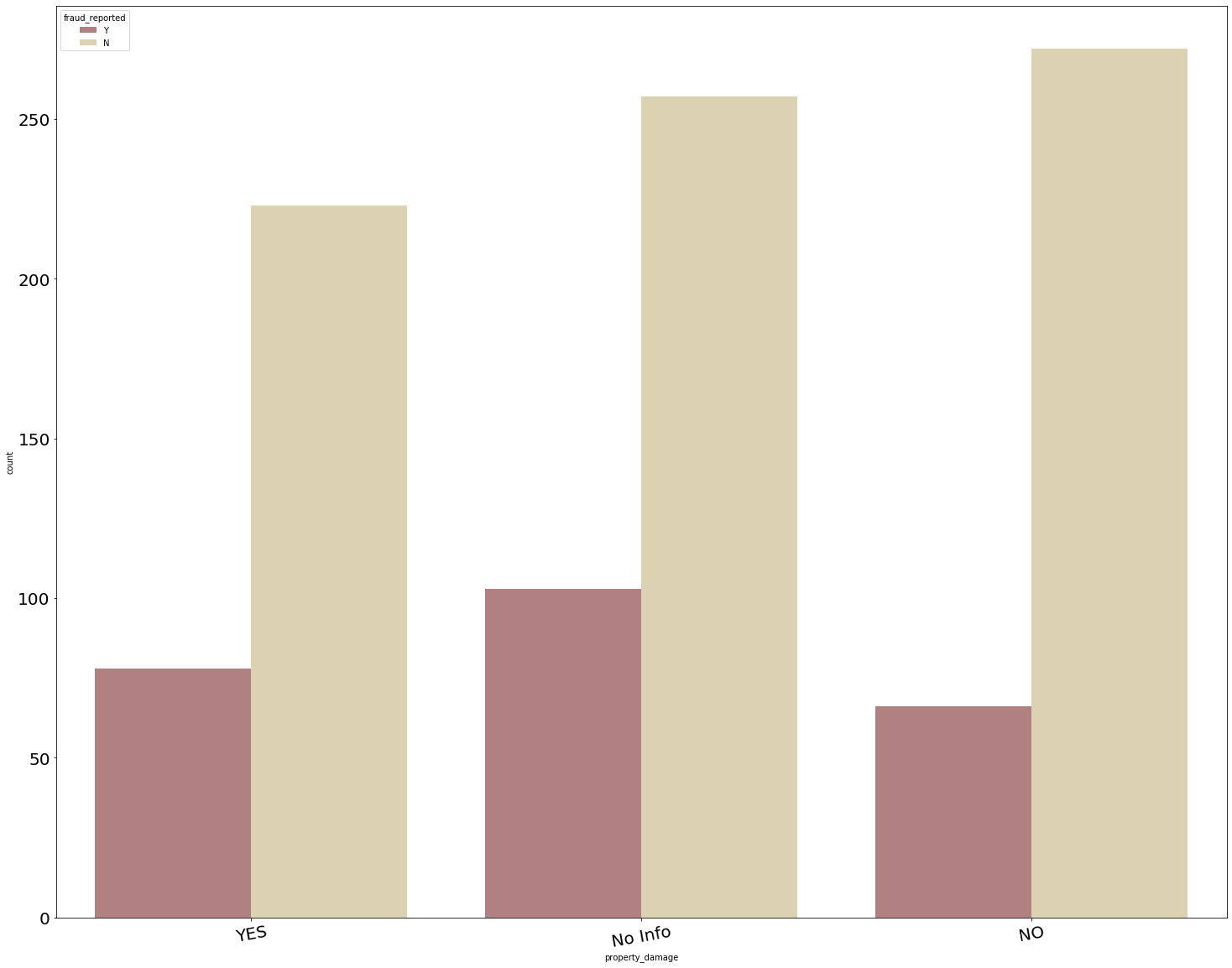
Majority of insured person belongs to NY state. The person who belongs to SC State there is high chance of fraud claim. The PA state have very less fraud claim, that is because there are only few insured people belongs to PA state.



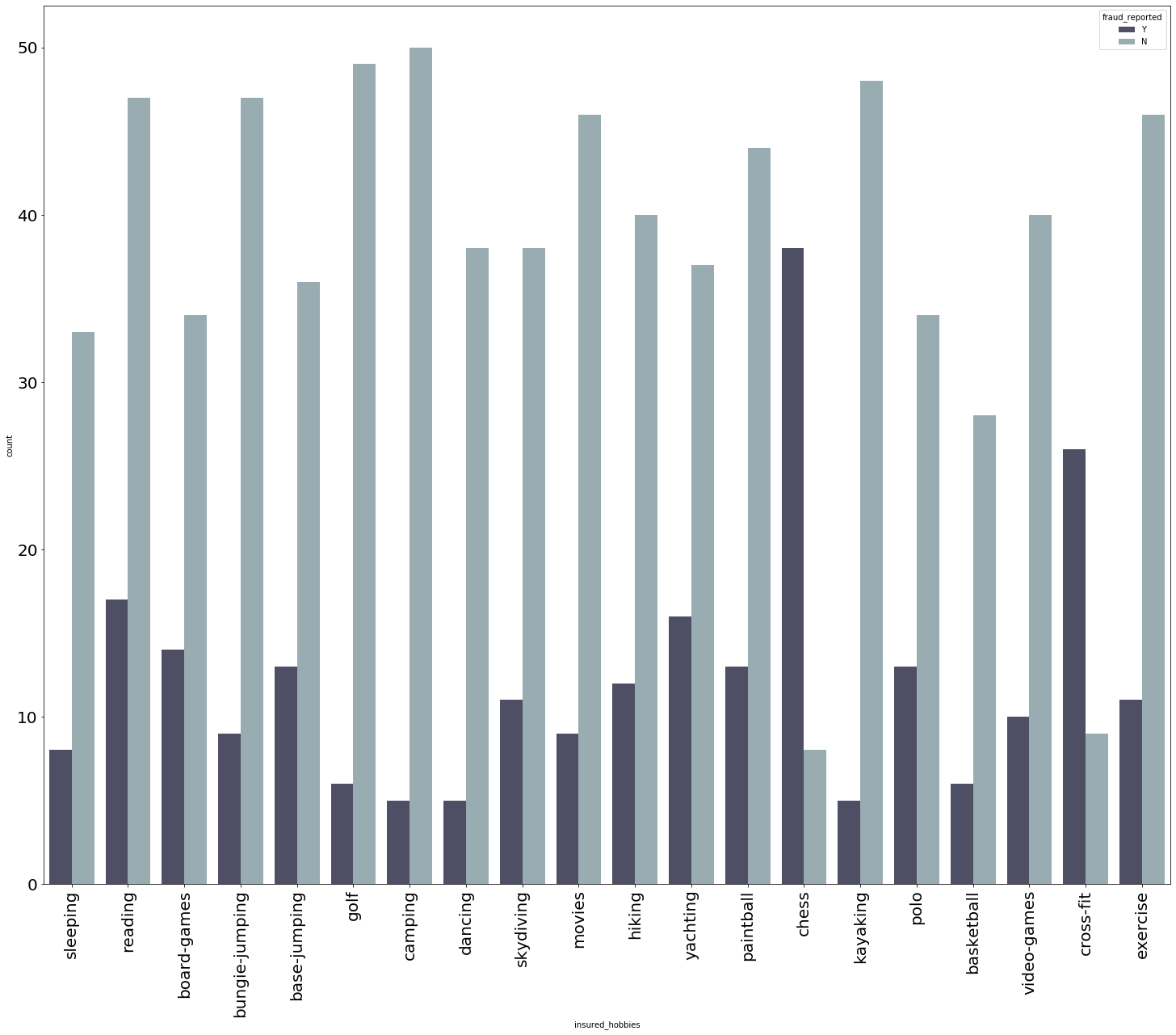
There is a high count in Major Damage and also there is high chance of fraud reported in the Major Damage category. If there is light damage happen to vehicle, they claim huge amount from insurance company for Major Damage. Sometimes they even damage their own vehicle and submitted the insurance claim form. For that purpose, we can build the Machine learning model to predict which is fraudulent claim or not.



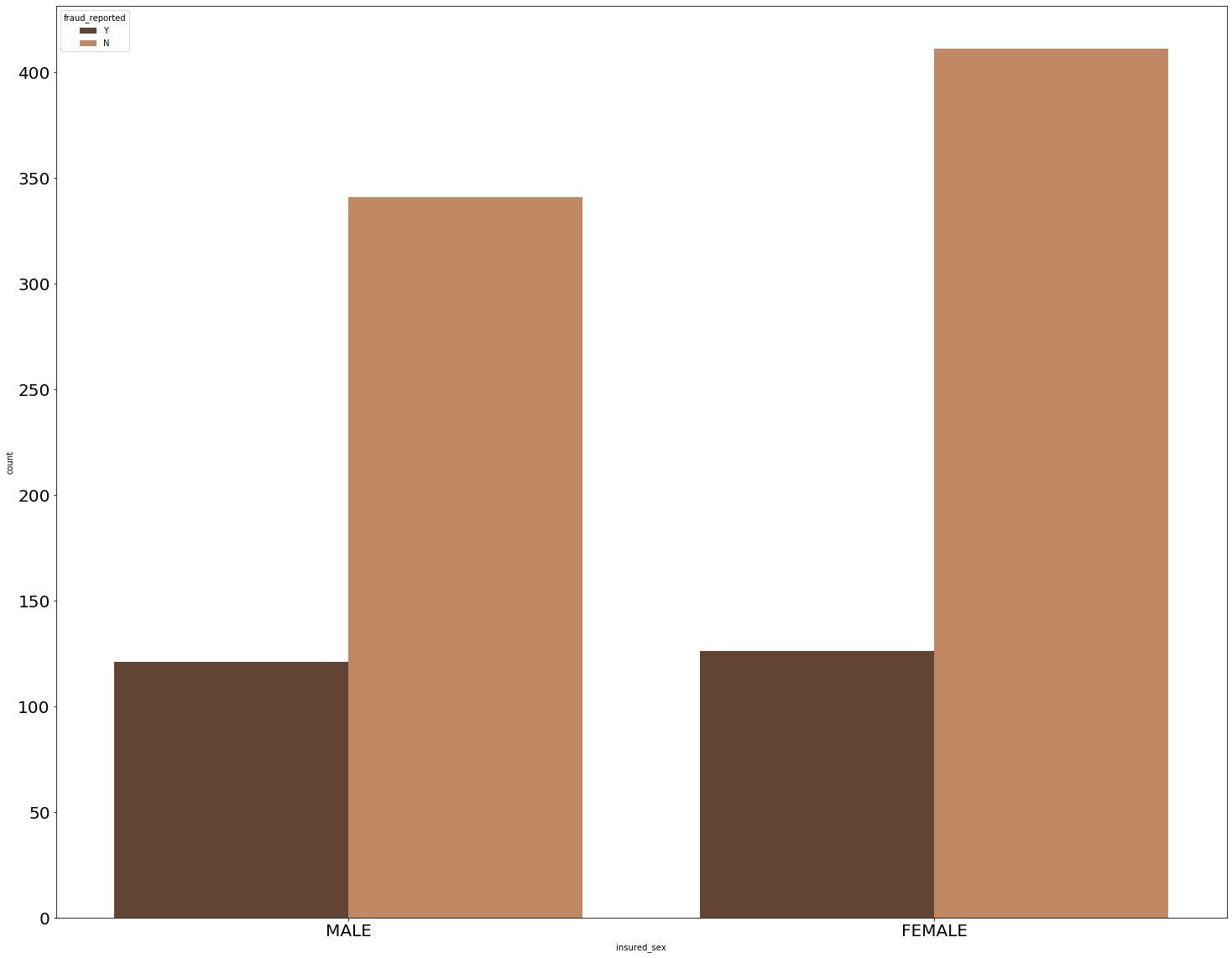
There is almost equal fraud reported in all the policy state. There is a little bit high fraud reported in OH policy state.



In the Property damage column, there are many rows that don’t have any information. 301 Insured persons has reported the property damage and 338 insured persons has not reported any property damage.

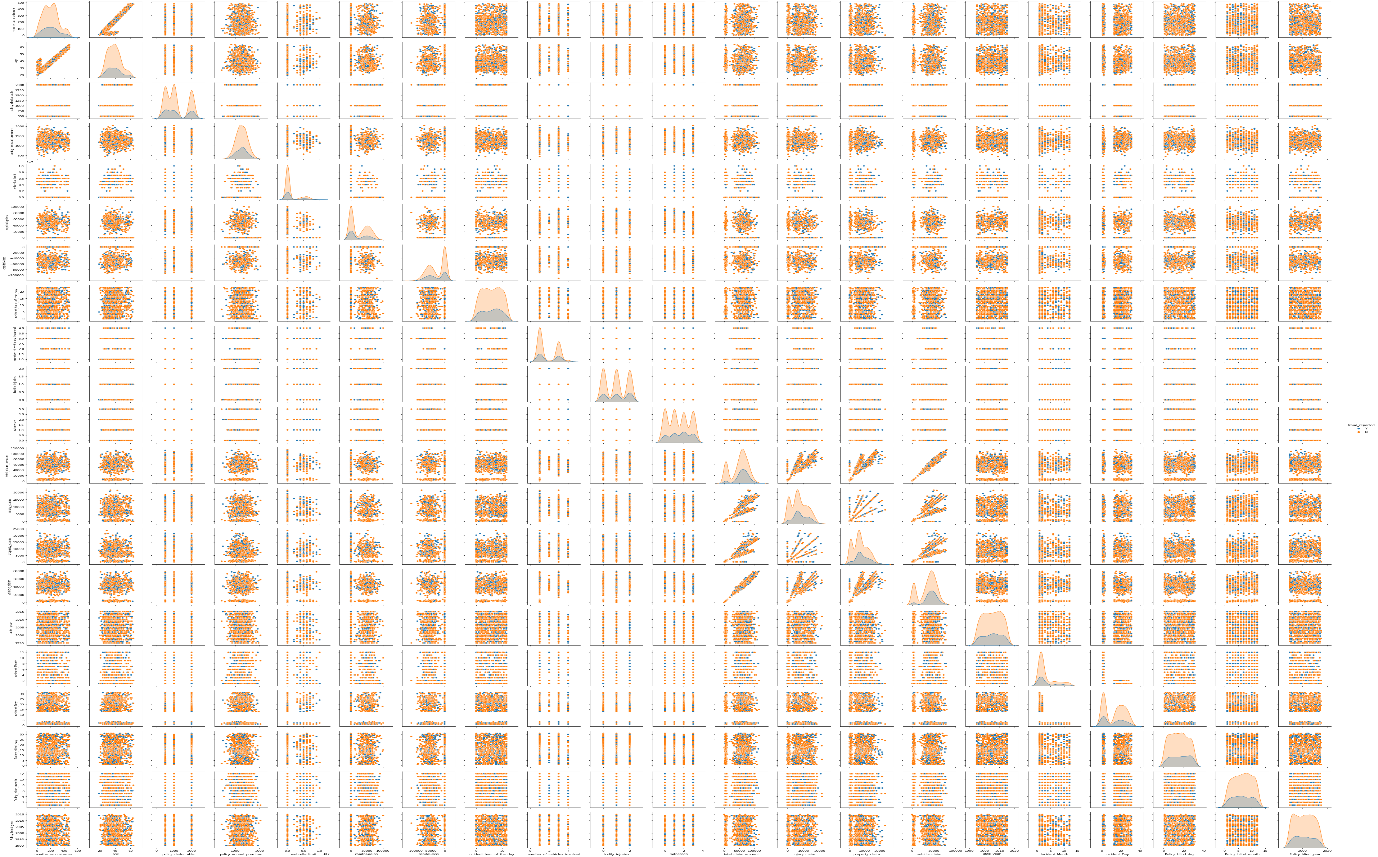


Let’s see what are the hobbies insured person have. Majority of the insured person have the hobbies of reading books. Then comes next Paintball and exercise. Only few insured persons are interested in playing basketball. If we checked the fraud reported with Hobbies. We see in the diagram, those who are interested in playing chess there is a high chance of fraud reported. Obviously, chess is mind game, we have to think each move carefully. So, the fraud claims also need to do carefully, only those people who strategically and logically strong they can do the fraud claim.



There is equal chance of fraud claim in both the gender.

Checking pairplot of each feature.



In the pair plot we see each feature with fraud reported. In the diagram the diagonal it is normally distributed. The data are scatter.

**EDA Concluding Remark**

* Checked the null value using the heatmap, we found there is no null value present in data.
* We have converted the date datatype from object to datatime datatype and we have split into day, month and years.
* We have converted the categorical data to numerical data using the Label Encoder.
* We dropped the irrelevant columns from our data set.
* We checked the unique value, data info, shape of data, column name, statistical summary of data using describe method.
* We have checked the count of each feature and visualize them using bar chart, violin plot, count plot.
* We have visualized each feature with fraud reported and analysis in which category there is high chance of fraud reported.
* We have checked the correlation of data.
* In some column there is a ‘?’ present we have replaced it with ‘No info’
* Checked the normal distribution of feature to find whether the data contained skewness or not.
* Using the pair plot we have checked the relations of each feature with fraud reported.

**Outliers removal and selecting 99.7% data:using zscore method.**

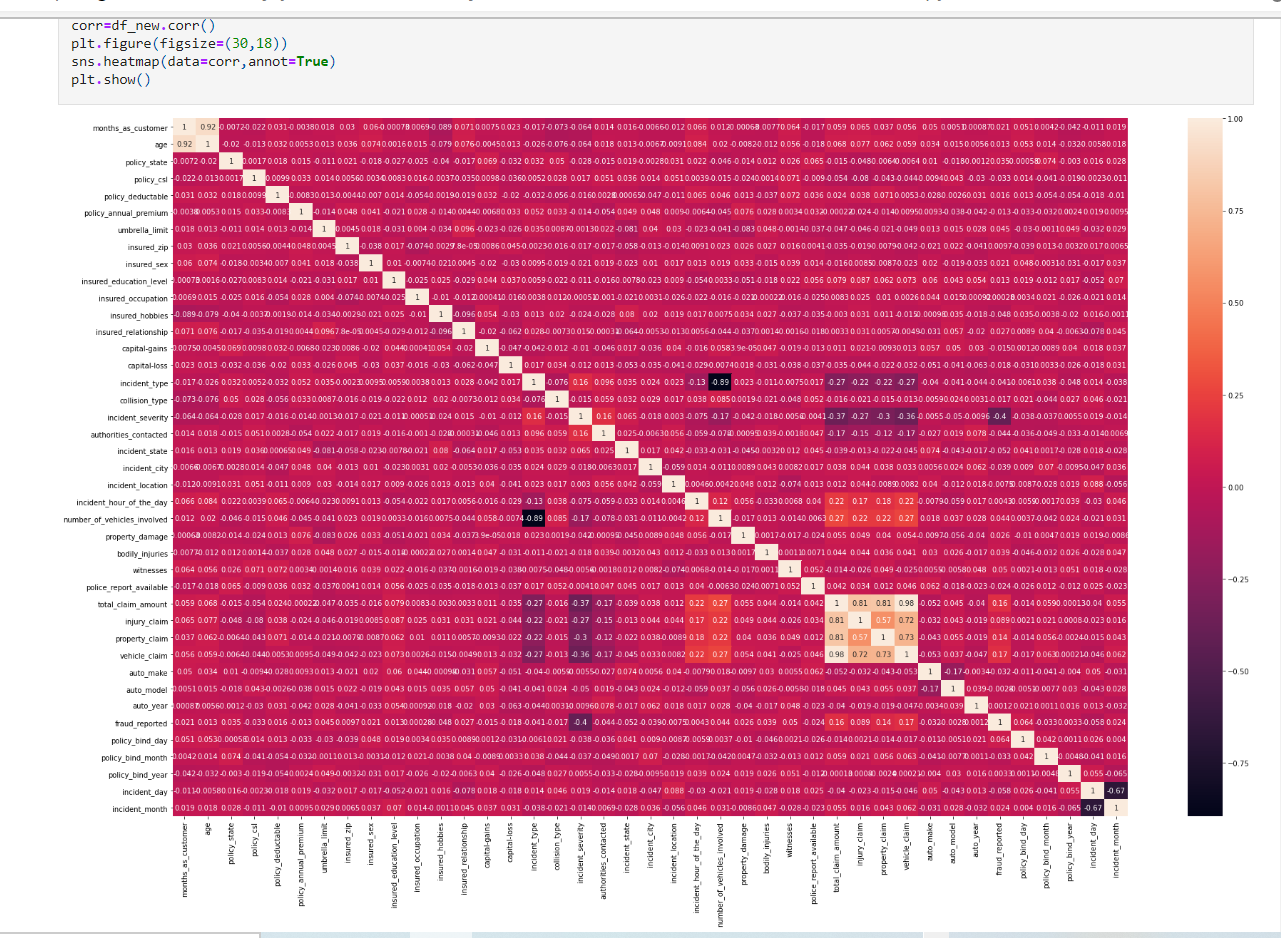
****

We checked both the data before removing the Outliers and after removing Outliers.

Before removing the Outliers, the shape is 999 rows and 39 columns. After removing the Outliers, the shape is 980 rows and 39 columns.

We checked the data loss how much data removed in the Outliers process. It has removed **1.9%** of data in Outliers.

**Correlation: checking features for multicollinearity among them using heat map.**

****

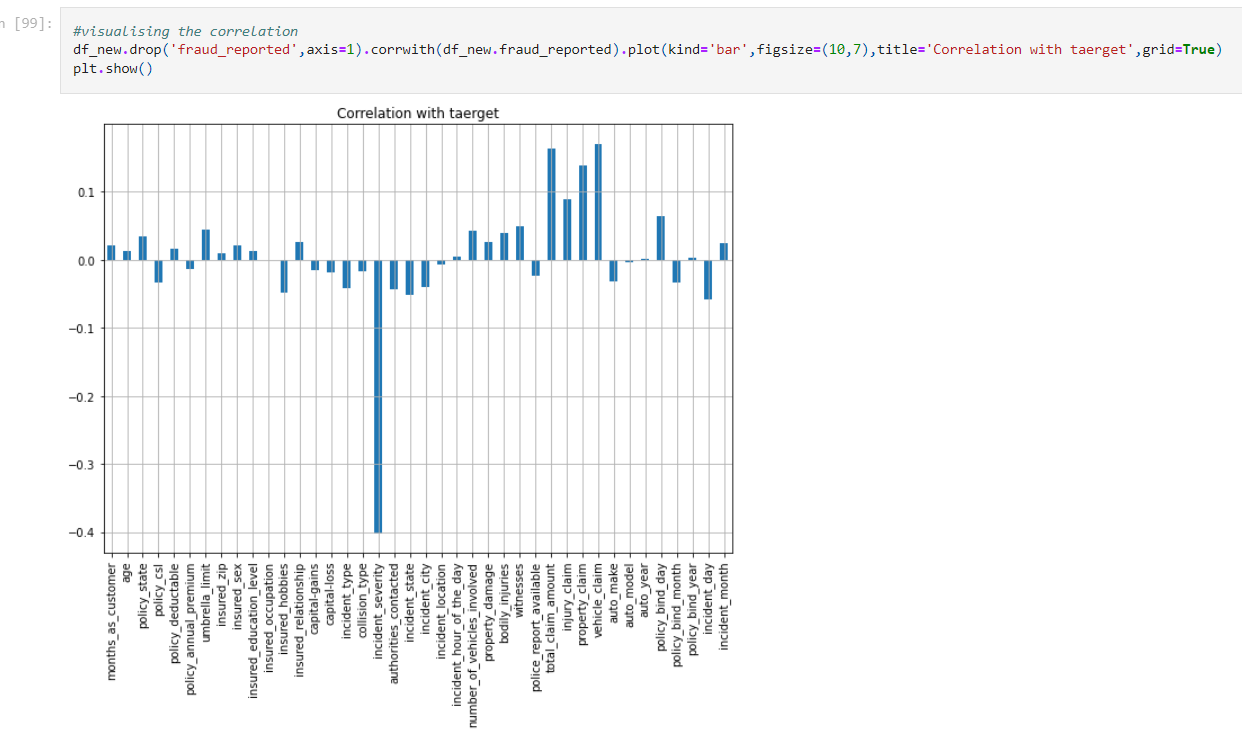
We checked the correlation of data, there is positive correlation, in few columns there is negative correlation. There is NaN value in Incident Year. So, we can drop the Incident Year column.

Visualizing the correlation using the heatmap. Total\_claim\_amount is correlated with vehicle claim,property\_claim,injury\_claim.

**Checking correlation between features and target:**

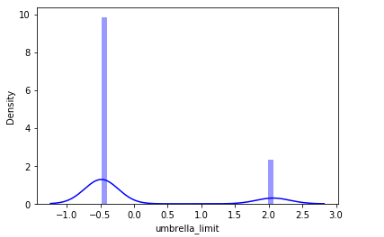
We can clearly see from below image , that incident\_severity column is highly negative correlated with target.

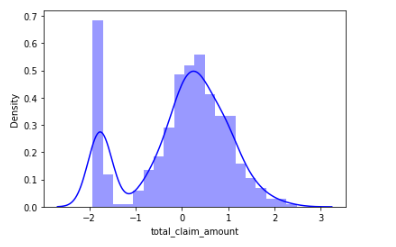
And,total\_claim\_amount , property\_claim, vehicle\_claim are positivly correlated with target.

****

**Feature Selection:** Using Kbest features selecting24 best features.

****

We have removed the skewness using **yeo-johnson** method and transform the skewed data to scaled data.

  
After removing the skewness, the normally distribution curve look like above.

**Pre-Processing Pipeline.**

* Removing skewness.
* Scaling features using standard scaler.
* Oversampling target using SMOTE.

****

The target variable fraud\_reported is not equal. We have balanced the data. So, we have used the over sampling technique.

Now the data is balanced, our target value is equal 740.

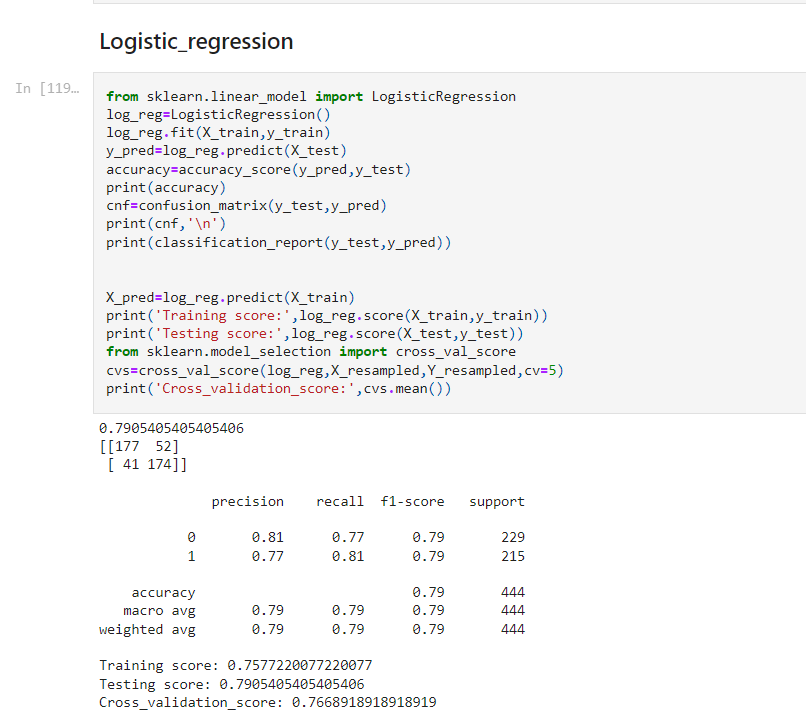


We see the data is balanced using the count plot.

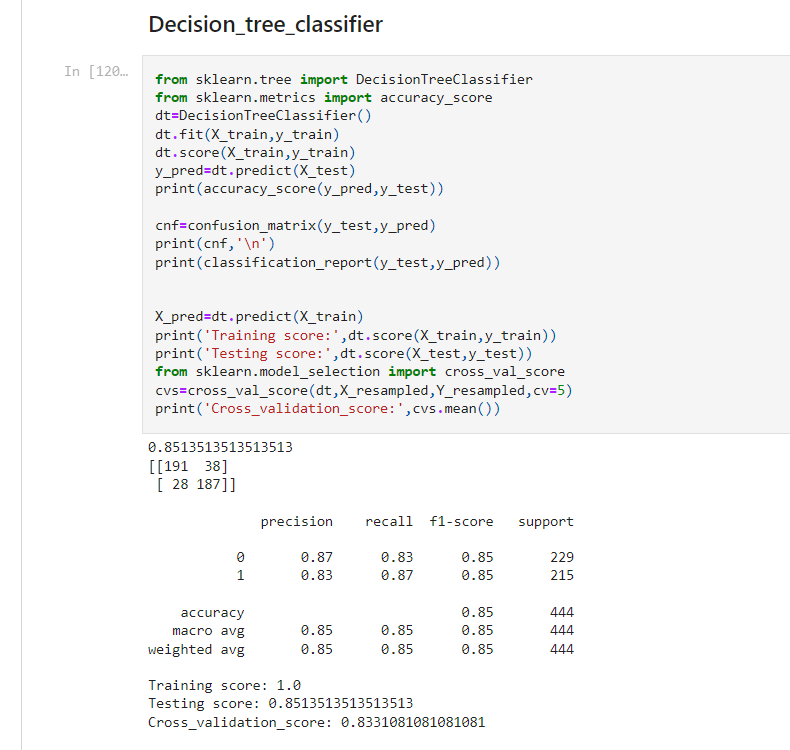
After dropping the total\_claim\_amount, we removed all the high inflation factor from our data set. There is no multi-collinearity exist.

**Building Machine Learning Models**

Logistic Regression: Using the Logistic Regression Model, our accuracy score is 79%. And cross validation score is 76%.

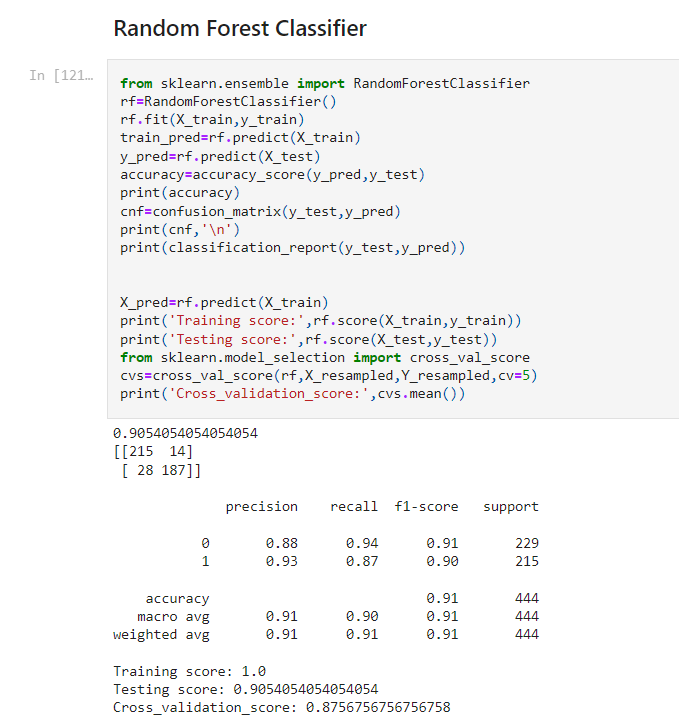
****

Decision Tree Classifier:



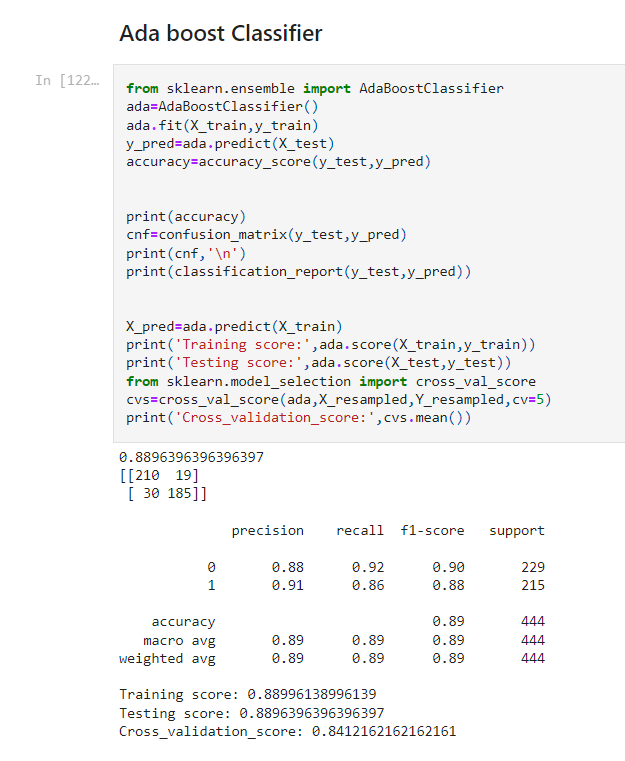
In Decision Tree model, our accuracy score is 85%. And cross validation score is 83%.

Random Forest Classifier: In Random Forest Classifier, our accuracy score is 90%. And CVScore is 87%. Which is a good score.

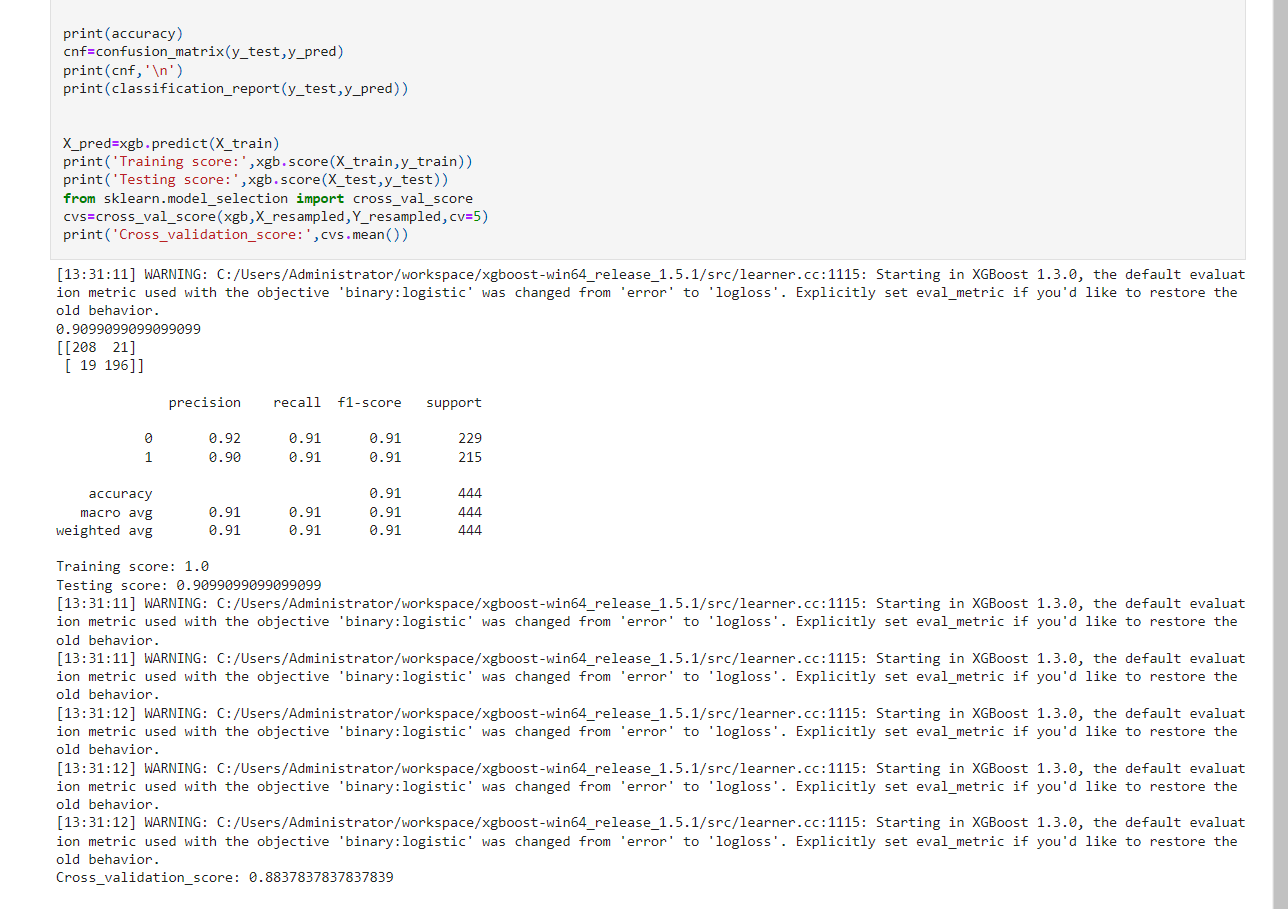


ADA Boost Classifier:

Accuracy score of ada boost model is 88% and CVS is 84%.



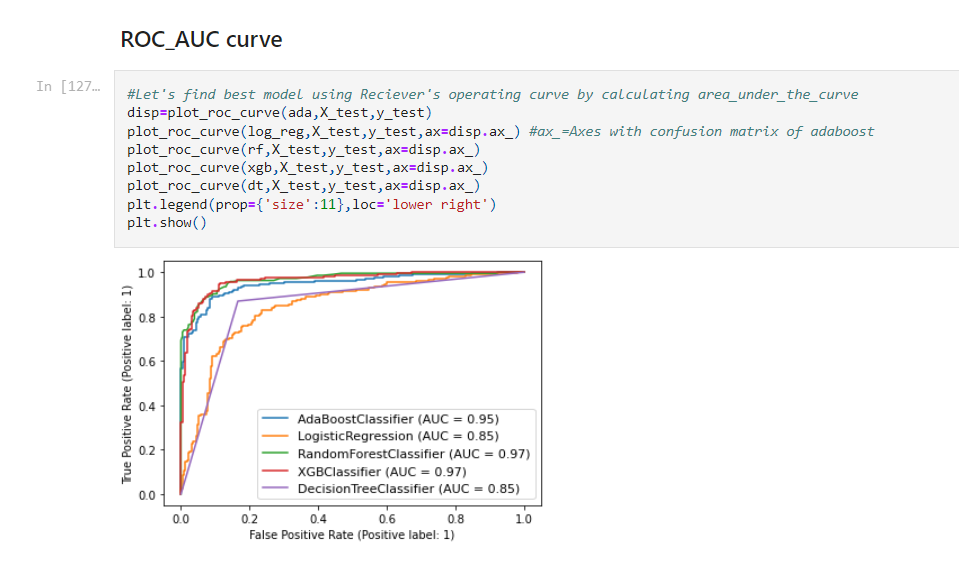
XGBoost Classifier:



Accuracy score of XG boost is 90% and CVscore is 88%.

ROC\_AUC Curve:

Area under ROC curve for random forest and XG boost model is 97%.



We checked cross validation score of various models, XG Boosting Classifier gives good accuracy score and cross validation score. So, we will consider XGBoosting Classifier as our final model. Let’s do the Hyper parameter tuning to increase accuracy.

**Hyper parameter tuning**:.

We have done the hyper parameter tuning using grid search CV and passed some parameter and we trained the model. In Grid Search CV the best score is 90.54%. We will pass this estimator in our final model to check the accuracy.

We have saved the model using Pickle library.

**Concluding Remarks**

In this Insurance Claim – Fraud Detection project we have analysis the data using various plot in Visualization. We have gone through Feature Engineering, Pre-processing the data. We have handled the imbalance data. We have converted the categorical data to numerical data using a Label Encoder.

We have checked the distribution of data and removed the skewness present. We also checked the Outliers and removed it using the zscore method. We have checked the correlation of data . Then we have scaled the data.

We have splitted the data for training and testing. We have given 70% of data for training and 30% of data for testing.

We have dropped the column which were having multi-collinearity issue. We have built various classification model and checked their accuracy score, confusion matrix and classification report.

We have checked the cross-validation score of each model and compare with other models. Picked the best model which gives a good score. In our project XG Boosting Classifier gives a good score. We have done the hyper parameter tuning to increase accuracy. In our scenario the accuracy score didn’t increase.

We have seen that predicted and actual value almost similar; this means our model is working well. Based on this prediction we can predict which insurance claim is fraudulent or not.

The insurance company can check the insured person features and predict the results whether to accept the insurance claim or reject it. This will help to decide and avoid the major loss to the company. Machine Learning model plays very important role in predicting the fraudulent claim.