**Insurance Claims- Fraud Detection**

Automobile Industry’s Claims Fraud Increases day by day. So, that in any Industry or Company it is necessary to predict the Claims weathers is Fraud or Not. It is not reliable for the company to check each and every Claim because it can be time Consuming and Costly in this case. We can go with Machine Learning models and also play Vital Role in any business environment.

With the Help of Different Attributes about the Claims, Insurance customers and other Circumstances we can find weathers this Claim is Fraud or not.

**Problem Statement:**

**Business Goal:**

The main goal of this project to find the Fraud of the Claims by Machine Learning Algorithms. The main issue behind the ML is variety of the fraud and very less amount of known Fraud.

Here we have 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 opted.

**Data Analysis:**

Here we have Dataset which includes insurance policy along with customers details and accident details.

Data Set Contains 100 Rows And 40 Features:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1000 entries, 0 to 999

Data columns (total 40 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 months\_as\_customer 1000 non-null int64

1 age 1000 non-null int64

2 policy\_number 1000 non-null int64

3 policy\_bind\_date 1000 non-null object

4 policy\_state 1000 non-null object

5 policy\_csl 1000 non-null object

6 policy\_deductable 1000 non-null int64

7 policy\_annual\_premium 1000 non-null float64

8 umbrella\_limit 1000 non-null int64

9 insured\_zip 1000 non-null int64

10 insured\_sex 1000 non-null object

11 insured\_education\_level 1000 non-null object

12 insured\_occupation 1000 non-null object

13 insured\_hobbies 1000 non-null object

14 insured\_relationship 1000 non-null object

15 capital-gains 1000 non-null int64

16 capital-loss 1000 non-null int64

17 incident\_date 1000 non-null object

18 incident\_type 1000 non-null object

19 collision\_type 1000 non-null object

20 incident\_severity 1000 non-null object

21 authorities\_contacted 1000 non-null object

22 incident\_state 1000 non-null object

23 incident\_city 1000 non-null object

24 incident\_location 1000 non-null object

25 incident\_hour\_of\_the\_day 1000 non-null int64

26 number\_of\_vehicles\_involved 1000 non-null int64

27 property\_damage 1000 non-null object

28 bodily\_injuries 1000 non-null int64

29 witnesses 1000 non-null int64

30 police\_report\_available 1000 non-null object

31 total\_claim\_amount 1000 non-null int64

32 injury\_claim 1000 non-null int64

33 property\_claim 1000 non-null int64

34 vehicle\_claim 1000 non-null int64

35 auto\_make 1000 non-null object

36 auto\_model 1000 non-null object

37 auto\_year 1000 non-null int64

38 fraud\_reported 1000 non-null object

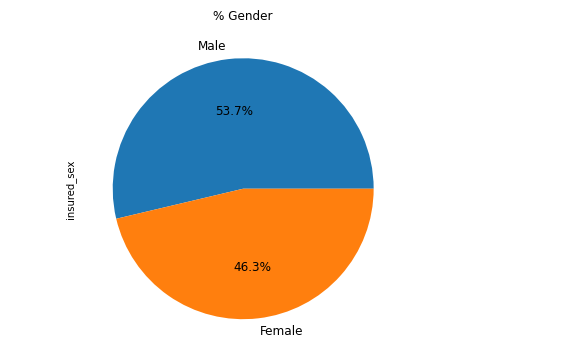
39 \_c39 0 non-null float64

dtypes: float64(2), int64(17), object(21)

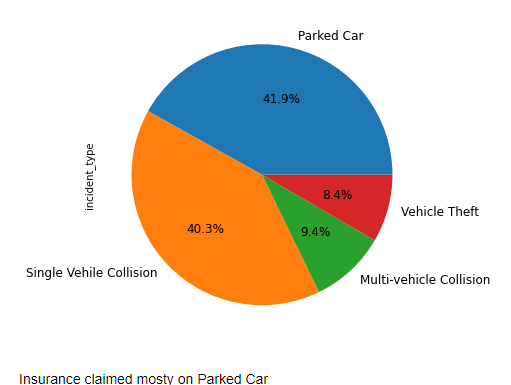
memory usage: 312.6+ KB

**Exploratory Data Analysis:**

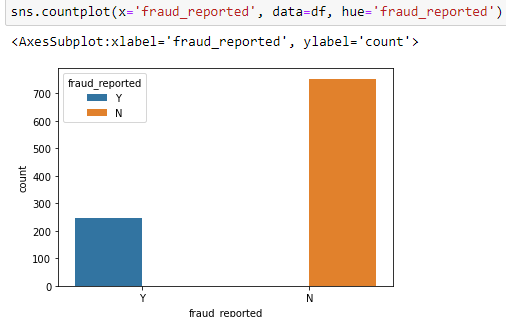
Dependent Variable/Target Variable: This Data Set having Fraud Reported as target variable.in this data set having 247 as Fraud and 753 as No Fraud.so it looks like imbalanced problem. Any ways we can balance the data in Data Pre-processing part.



Mostly Males are Insured more.

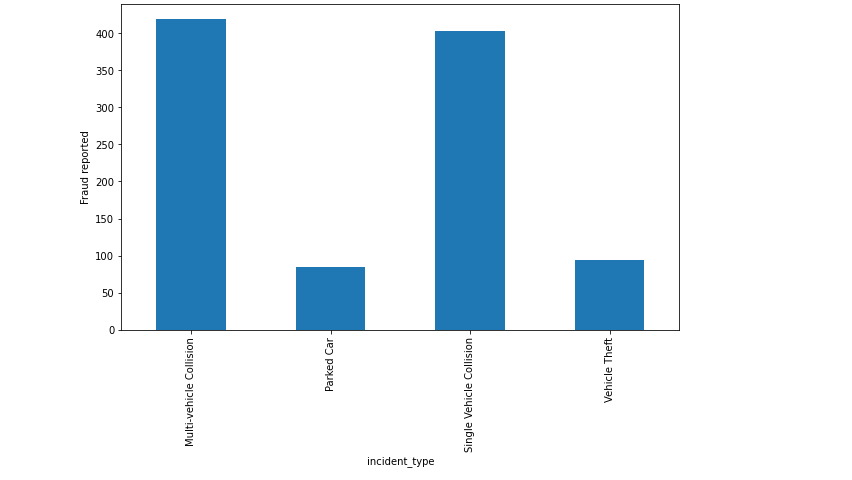


Now we will calculate weather Company is in Loss or Profit due to Fraud claims.



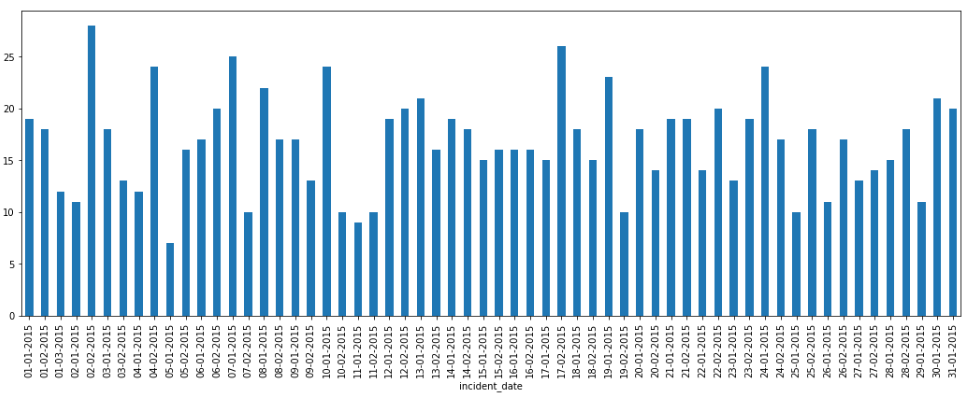
Here we see that Company is in Big Loss. That’s why it is very necessary to concern on this

Fraud Claims.

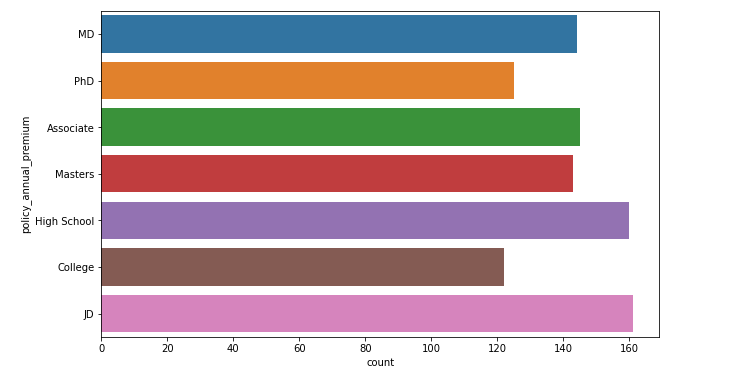


Single vehicle and Multi-Vehicle Collision are High in when compare to Vehicle Thief and

Parked Vehicle Accident.



Here we can observe that the Claim Amount is more in the incident date on 02-02-2015.

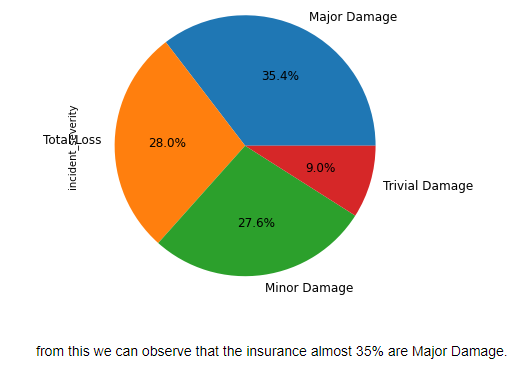


This chart indicated Ex-Managerial having higher Tendency of Fraud and handlers-cleaners

having less Tendency fraud.

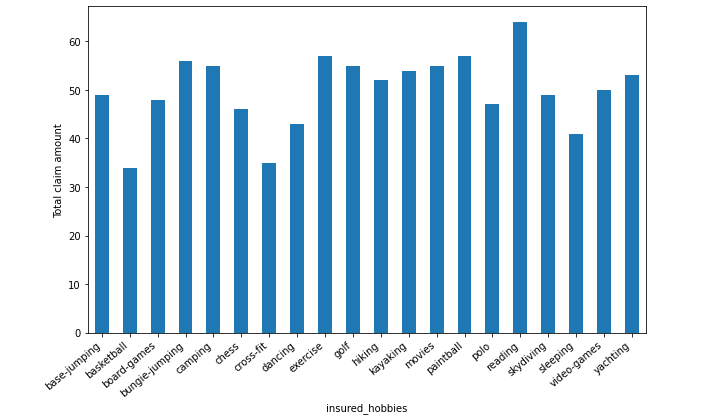
This is very shocking observation highly educated & higher post becomes fraud and less

educated having less fraud.



From this chart we say that Major damage severity having more Fraud. So, that we should check across the claim amount too. There may be possible whose claim for higher amount they have

fraud too.



From this visualization we clearly see that Fraud find in Higher Claim Amount and for lower

amount of claim fraud is none.

**Data Pre-processing Pipeline:**

Data pre-Processing is the process of the transforming the raw data into the useful or understate Data.

In the real dataset there is lots of mixture of data like missing values, incomplete values, Noisy data and much more. It is necessary to pre-process the data before applying into the model.

There are steps in Data Pre-Processing.

1. Data Cleaning: - Removing Outliers, Skewness and imputing Missing Values.

2. Data Transformation: - Like Normalization by applying normalization we can improve the accuracy

and efficiency of the models. And also reduce the errors.

3. Data Reduction: By Reducing the no of features by Feature Selection Process, PCA and VIF.

In Our Project, we find some’?’ In data set than we first replace it with none value and after that we put most frequent.

df.isnull().sum()

months\_as\_customer 0

age 0

policy\_number 0

policy\_bind\_date 0

policy\_state 0

policy\_csl 0

policy\_deductable 0

policy\_annual\_premium 0

umbrella\_limit 0

insured\_zip 0

insured\_sex 0

insured\_education\_level 0

insured\_occupation 0

insured\_hobbies 0

insured\_relationship 0

capital-gains 0

capital-loss 0

incident\_date 0

incident\_type 0

collision\_type 0

incident\_severity 0

authorities\_contacted 0

incident\_state 0

incident\_city 0

incident\_location 0

incident\_hour\_of\_the\_day 0

number\_of\_vehicles\_involved 0

property\_damage 0

bodily\_injuries 0

witnesses 0

police\_report\_available 0

total\_claim\_amount 0

injury\_claim 0

property\_claim 0

vehicle\_claim 0

auto\_make 0

auto\_model 0

auto\_year 0

fraud\_reported 0

\_c39 1000

dtype: int64

**Converting Categorial Data into Numerical**

There are few techniques used for converting categorical data into the Numerical data like

OneHotEncoder and Label Encoder.

In our project I used Label Encoder For converting the data. Label Encoder can be import from

the sklearn library.

**Data Cleaning Process**

First, we have to remove some Unnecessary Features like Policy no, policy\_bid\_data, profit/loss,

incident\_hour\_of\_the\_day, number \_of\_vehical\_involved.

**Removing Outliers:**

It is defined as the points that are far away from the same points. It can be happened because of

The variability of the measurements and may be some errors also. If it is possible, outliers should be removed from the datasets.

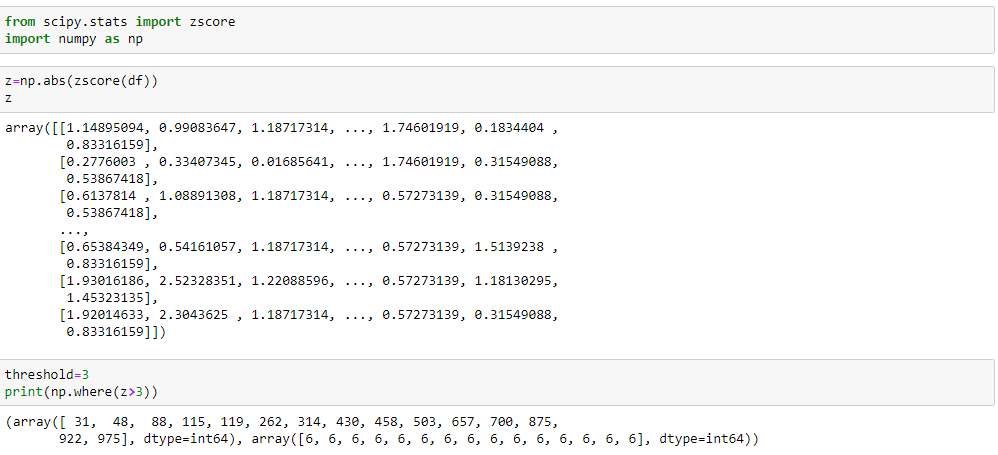
There are servals methods to remove the outliers.

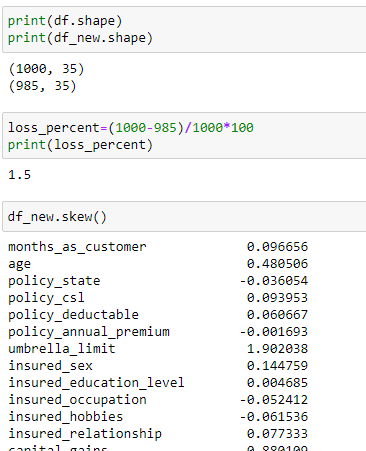
1. Z score.

2. Quantile Method (Capping the data).

In this project I used Zscore to remove the outliers from the data.

**1. Z Score:** It can call from the SciPy. Stats library. And for most of the case threshold values should be used 3.





Here Percentage of Data Loss Is 1.5.

**2.Quantile Methods:** Inter Quantile Range is used to detect or cap the outliers. Calculate the IQR by scipy.stats.iqr

Multiply Interquartile range by 1.5

Add 1.5 x interquartile range to the third quartile. Any number greater than this is a suspected

outlier.

Subtract 1.5 x interquartile range from the first quartile. Any number lesser than this is a

suspected outlier.

**Splitting Data Into train\_test\_split**

This function is in sklearn. Model selection splitting the data array into two arrays. Train and

Test with this function no need to split train and test manually, by default it makes random

partition and also set the random state. So, that it gives four o/p

like x\_train, x\_test, y\_train, y\_test.

After Doing splitting we have to balanced our data.it can be by SMOTE or oversampling

methods. Like Up Sampling, down sampling.

**Up sampling:** This method used to modify the unequal data into the balanced data by increases

the minory class or rare class. Advantage of this method is to no loss of information but from

that model can be in overfitting.

**Down Sampling:** Like the Up sampling it’s also balanced data but by reducing the size of the

class which is high.

**SMOTE**

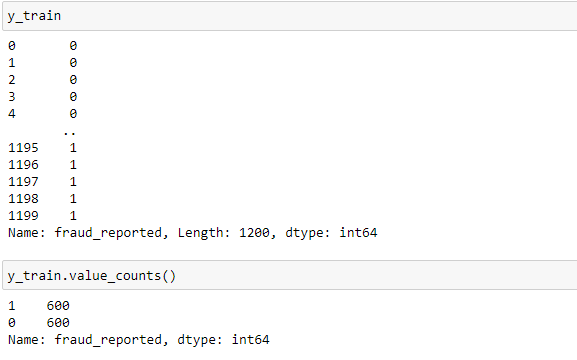
But if we balanced our data before train test split means we balanced from our whole data set or

form x. It means at that time our test data is leak. We have to isolate our test data. Here you

expose it. So, our f1 or recall or precision will be good.

Our model will already know which is positive or negative. And I can also say because of that

there is bias or model Overfitting to prevent this we balanced our data.

****

Now our Data is ready to build a best fit model.

Try Different Models:

**Logistic Regression:**

Logistic regression is the supervised machine learning problem which is used for the

classification problem and used to predict the probability of the classification.it is widely used

for the binary classification problem. It is one of the simplest methods of MI.

**Decision Tree Classifier:**

DTC can be used by both classification and regression both. But mostly it’s used for the classification problem. Its structure is tree based. Where internal nodes represent the features of dataset

and branches represents the decision rules and each leaf nodes represents the outcomes.

**Now let’s understand the Recall Precision and f1-score.**

**Accuracy:**

It can be defined as the ratio of total number of correct classifications divided by total number of

classifications. Accuracy=(TP+TN)/(TP+FP+TN+FN) .

**Recall:**

It is measure of correctly identified positive cases from all the actual positive cases.

It is useful when cost of False Negative is high. Recall=TP/(TP+FN) .

**Precision:**

It is measure of all the positive predictions how many of them actually positive.

Precision=TP/(TP+FP).

**F1-Score:**

It gives the combine result of Recall and Precision F1-score=2\*(Precision\*Recall)/ (Precision + Recall).

**Confusion Matrix:**

It is the table that is used to describe the performance of classification model on set of tests data by using different parameters.

**Hyper Parameter Tunning**

Hyper parameter optimization in machine learning is used to find parameters of given machine learning algorithm that perform best as measured on validation.

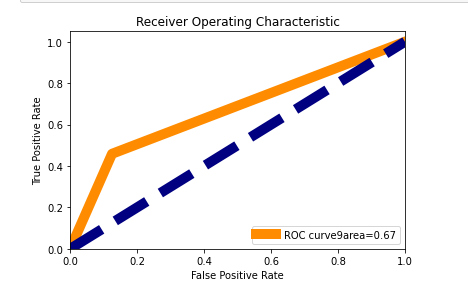
I used GridSearchCV for hyper parameter tunning.

After Hyper Tuning it will increase score from 71 to 77.

**ROC-AUC Curve**

It is the performance measurement of the model at different thresholds.

ROC is the performance score and AUC is the separation score means how much mode classify 0 as 0 and 1 as 1.



**Cross Validation**

This technique is used to check weather our data set is over fitting or under fitting. If model score is high and cv score is less it means model perform well in train dataset but did not perform well in unseen or

test dataset.

Feature selection is the best way to overcome the overfitting problem. There are 3 ways for the

validation. KFold Cross validation score, Hold Out Methods and LOOCV.

**KFold**

In this technique it will rotate the data into the k-fold times. Suppose k=3.

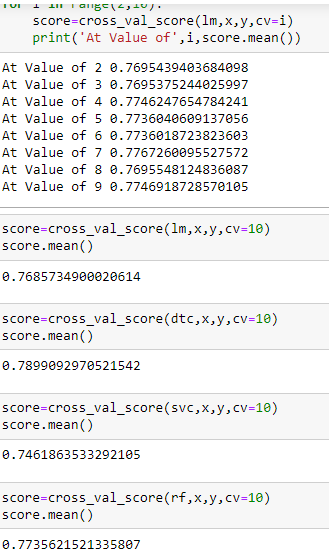
1 2 3 4 5 6 7 8 9

1st Iteration: 1-3 as Test and 4-9 Train

2nd Iteration:4-6 as Test and 1-3 & 7-9 Train

3rd Iteration:7-9 as Test and 1-6 as Train

It means all the data (9 rows) go for training.



**LOOCV**

Leave one out cross validation, it will take one row for test and remaining for training.

Each and every row go for test and its time-consuming processing.

**Concluding Remarks**

From this model we can detect the auto insurance fraud by using this model loss of the company can be reduced. We used different classifiers like Logistic Regression, Decision tree Classifiers and gradient

boosting classifiers. And also used the data Balanced process and also hyper parameter tunning for

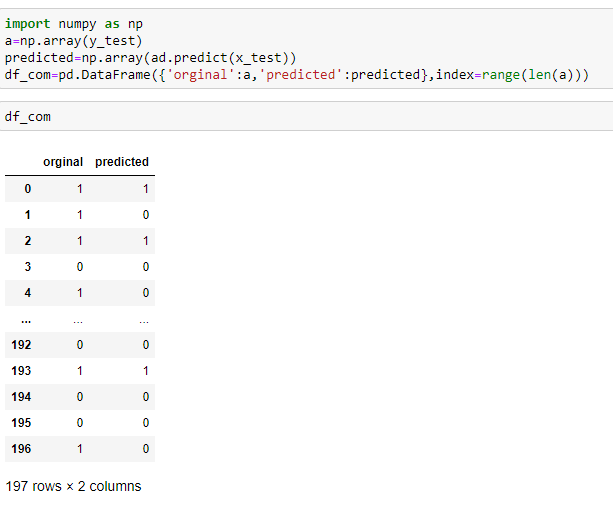
improving score.

We get good score in Gradient Boosting Classifiers.F1 score is 85% and Roc AUC Score is 89. And

the model performance is excellent. Model can distinguish correctly weather the claim is Fraud or

Correctly with high accuracy.

**Saving the Model**

****