**INSURANCE CLAIM**

**FRAUD DETECTION**



**Submitted by**

**Alivia Dasgupta**

**Introduction**

Insurance fraud detection refers to the identification and prevention of fraudulent activities related to money or property insurance. It involves the use of various software-based solutions to analyze historic patterns and incidents to predict future occurrences. The software performs statistical analysis using artificial intelligence (AI), machine learning and traditional rule-based fraud analytics models.

Insurance fraud detection is commonly used by organizations for:

1. Fraud analytics,
2. Authentication,
3. Governance, risk and
4. Compliance to safeguard databases and identify anomalies and vulnerabilities.

All these steps are involved along with the algorithm on the partial training datasets and then it is tested on the test datasets, finally it is then examined by some random splits. The data in the datasets are handled by certain rules which are as follows:

* 1. Problem Definition.
  2. **Data Analysis.**
  3. **EDA Concluding Remark.**
  4. **Pre-Processing Pipeline.**
  5. **Building Machine Learning Models.**
  6. **Concluding Remarks.**

**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.

# Task:

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.

# Insurance Fraud Detection Market Trends:

The increasing occurrence of insurance frauds across industries is one of the key factors driving the growth of the market. Insurance fraud detection systems are widely used to identify any cover-up of the evidence, misinterpretation of the incident or inflating the severity of the loss caused.

Moreover, the rising incidence of inaccurate claims, fake medical records, postdated laws, abductions, deaths and other customer frauds is also providing a thrust to the market growth. In line with this, organizations are widely using artificial intelligence (AI) and the Internet of Things (IoT)-enabled fraud detection solutions for running automated business rules, self-learning models, text mining, image screening, network analysis, predictive analytics and device identification, which is also contributing to the growth of the market.

## Importance of machine learning suited to fraud detection?

SUPER FAST

When it comes to fraud decisions, you need results FAST! Machine learning is like having several teams of analysts running hundreds of thousands of queries and comparing the outcomes to find the best result - this is all done in real-time and only takes milliseconds.

As well as making real-time decisions, machine learning is assessing individual customer behavior as it happens.

SCALABLE

Every online business wants to increase its transaction volume. Machine learning systems improve with larger datasets because this gives the system more examples of good and bad.

EFFICIENT

Remember that machine learning is like having several teams running analysis on hundreds of thousands of payments per second. The human cost of this would be immense - the cost of machine learning is just the cost of the servers running.

Machine learning does all the dirty work of data analysis in a fraction of the time it would take for even 100 fraud analysts

ACCURATE

Machine learning models are able to learn from patterns of normal behavior. They are very fast to adapt to changes in that normal behavior and can quickly identify patterns of fraud transactions.

This means that the model can identify suspicious customers even when there hasn’t been a chargeback yet.

**Objective:**

The techniques in the machine learning is to improve the accuracy of detection on various imbalanced datasets. A machine learning system works by:

1. INPUT DATA
2. EXTRACT FILES
3. TRAIN ALGORITHM
4. CREATE A MODEL

All these steps are involved along with the algorithm on the partial training datasets and then it is tested on the test datasets, finally it is then examined by some random splits. The data in the datasets are handled by certain rules.

## About Dataset:

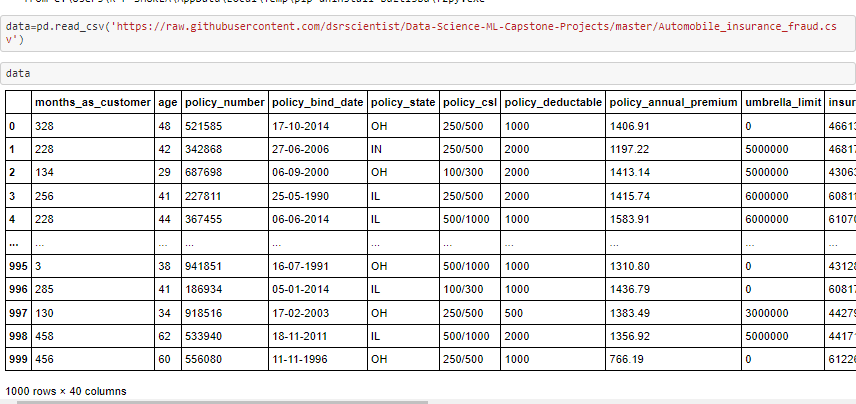
The dataset is about the auto claim insurance dataset along with the customer details for which they have claimed their insurance. Model is predicted whether the claim is authentic or fraud

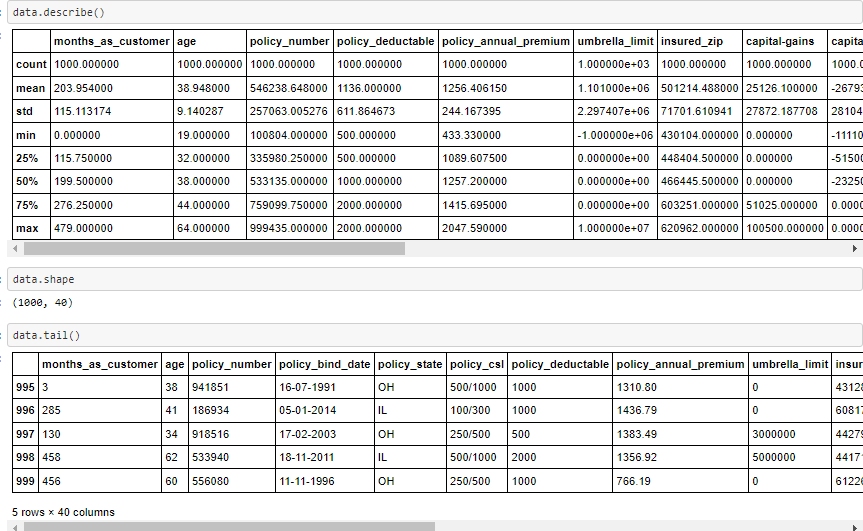
## DATA ANALYSIS

**Importing Libraries:**



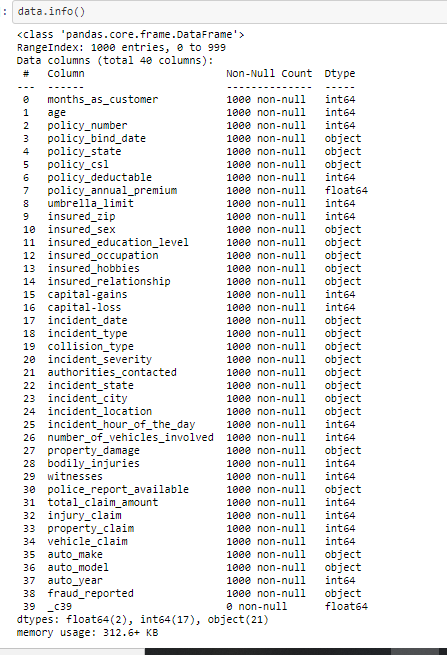
**Extracting dataset**



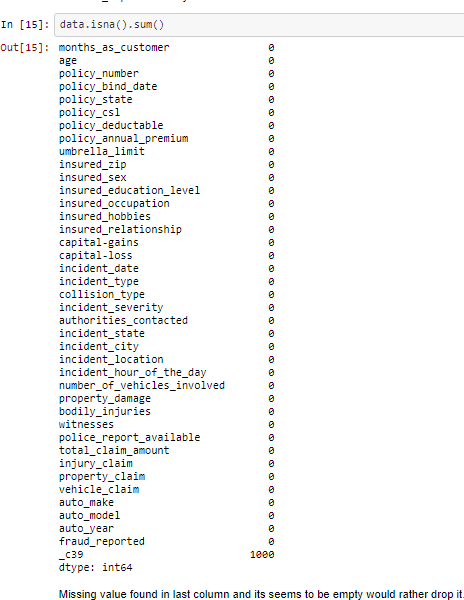


The given dataset contains:

1000 rows and 40 columns. Using this dataset, we will be training the Machine Learning models on 80% of the data and the models will be tested on 20% data.



There are no null values in the given dataset. There are 64 float, 17 integer & 21 object Data types.



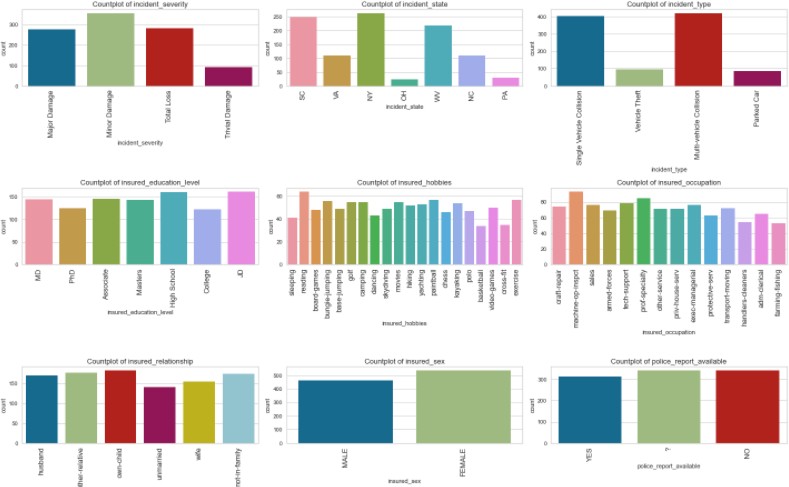
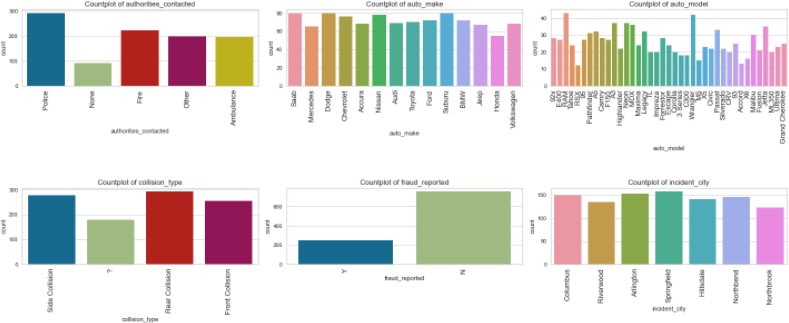
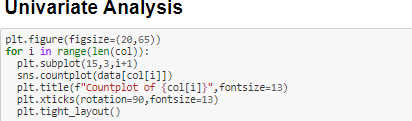
* \_c39 is the missing value found in last column. It must be dropped by me.

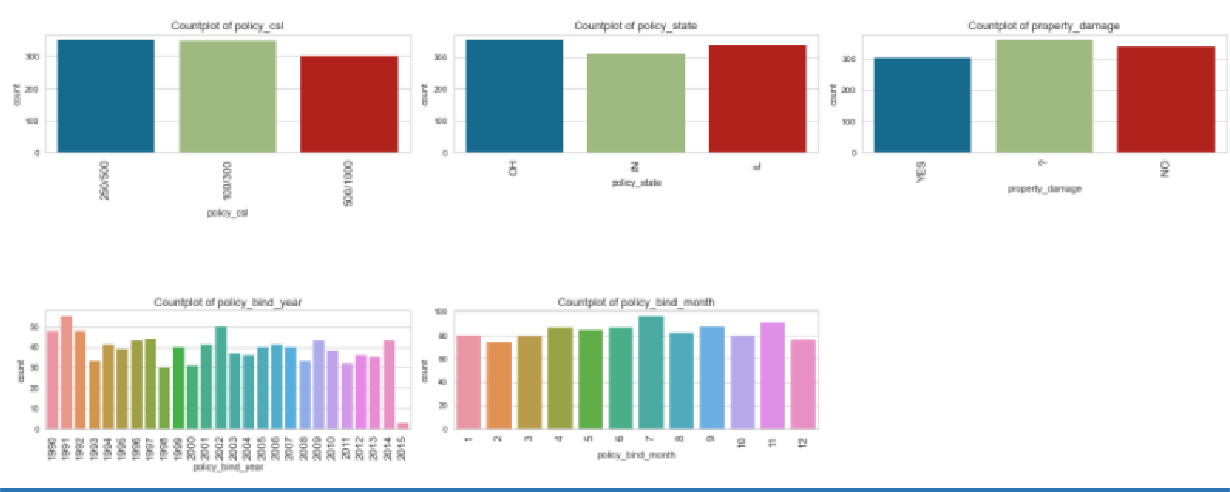
df.drop (['\_c39','policy\_bind\_date','incident\_location','incident

\_date'], axis=1, inplace=**True**)

# DATA PRE-PROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

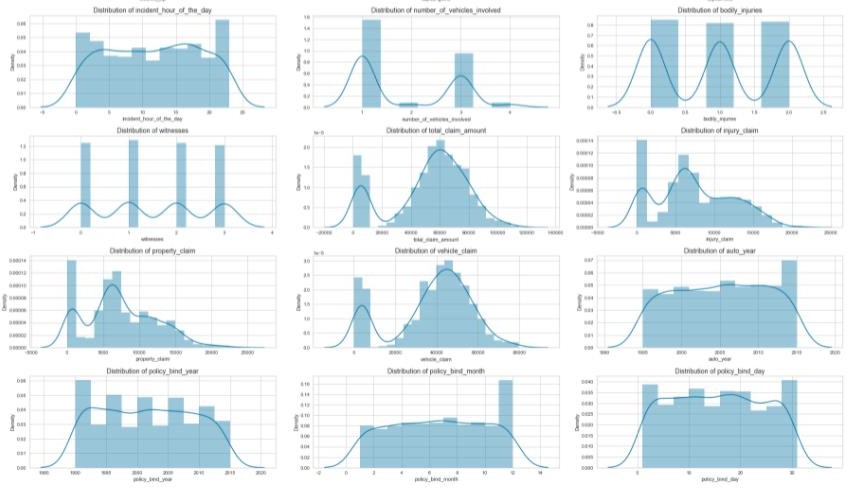




##### OBSERVATIONS MADE VIA THESE COUNTPLOTS:

1. From count plot authorities contacted Police has the highest count followed by fire. In most of the cases people have contacted police first.
2. From count plot automaker Saab, Suburu, Dodge has the highest count in production of automobiles.
3. From count plot auto modal RAM and Wrangler has the highest count.
4. From count plot collision type rear collision has highest count.
5. From count plot Fraud Reported maximum number of frauds hasn’t been reported or claimed.
6. From count plot incident city Springfield has highest count whereas Northbrook has least count among all.
7. From count plot incident severity Minor damage has highest count means mostly people claim insurance for minor damage.
8. From count plot incident state New York has the highest count means most of the accident happens there.
9. From count plot incident type Multi-vehicle Collision and Single- vehicle Collision has the highest count means most of the accident happens with multiple vehicles and single vehicle and very less accident happens with park cars
10. From count plot insured education JD and High school has highest count.
11. From count plot insured hobbies reading is most popular among all others.
12. From count plot insured occupation machine-op-inspect has highest count means most of the people who claim insurance has this occupation and people who have farming-fishing occupation has less claim insurance.
13. From count plot insured relationship own-child has highest count means most of the people who have claimed insurance has child and unmarried has the least count.
14. From count plot insured sex male are less and female are high means people who have sex female has claimed insurance more than male.
15. From count plot policy class 250/500 and 100/300 has same high count and 500/100 has less count.
16. From count plot policy state IN has less count and Il and Ch has same high count.
17. From count plot property damage? and no has high count and yes has less count.
18. From count plot policy bind year most of the people have taken policy in 1991 and 2002 and only few people have taken policy in 2015.

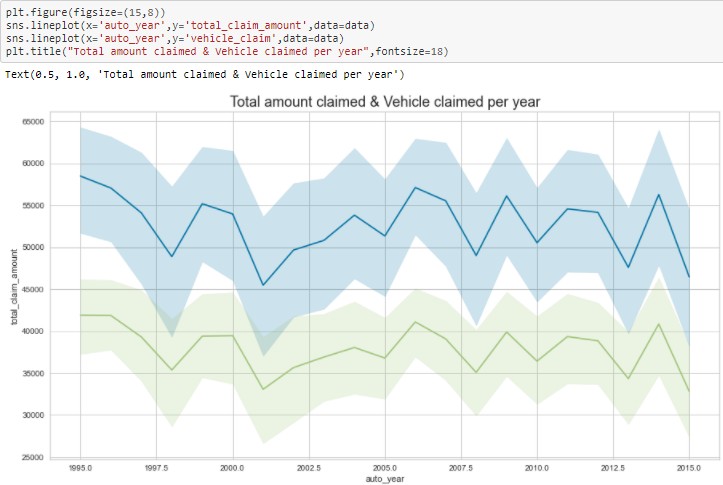
**DISTRIBUTION PLOTS OF COLUMN 1**



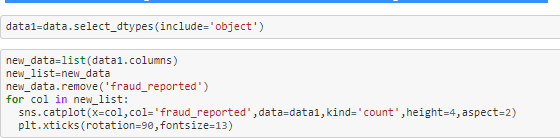
##### OBSERVATIONS MADE VIA THESE DISTPLOTS:

1. From months\_As\_customers most of the people lies in 0-100 and there are less number of people who lies in b/w 300-500 who are loyal customers.
2. From distribution of age most of the people lies between 30-45 and less people are in between 50-60.
3. From policy\_annual\_premium it is normally distributed.
4. From policy deductible I can say the value is between 500-1000 and 17000-20000
5. From capital gain 0-10000 has highest peak and with capital loss 0 to -10000 has high peak rest of the distribution are same both features.
6. From total\_claim\_amount 0-10000 has highest peak and rest of all values are normally distributed.
7. From distribution of property claim most of the people have claimed that 0-1000 values and from 4000-10000 these are 2nd highest people who have claimed for these values and there are very few people who claimed for 20000-25000 values.
8. From distribution of vehicle claim there are many people who claimed for 0-10000 and rest of the value has normal distribution.

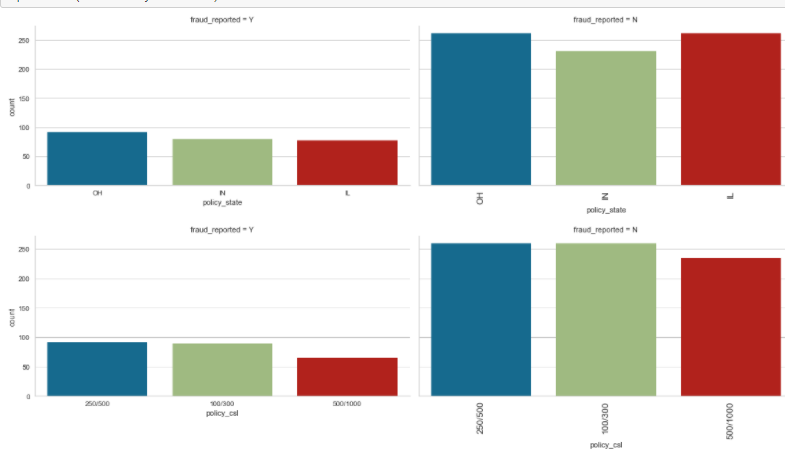
# Bivariate analysis

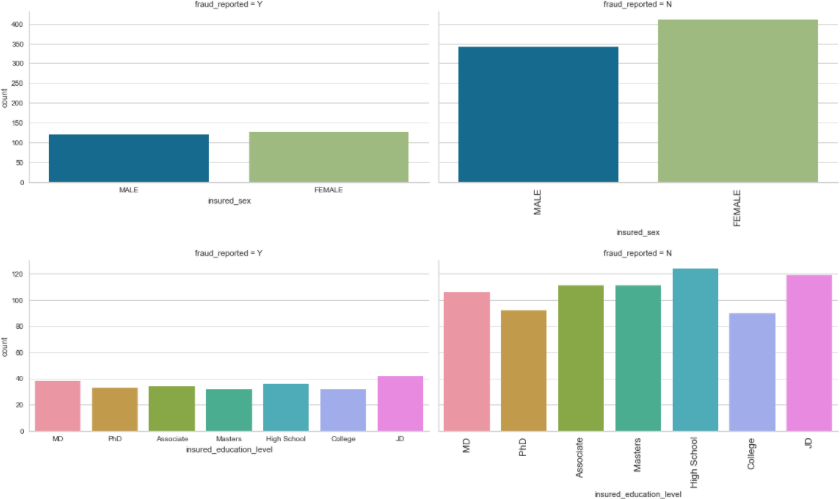


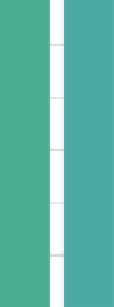
Total amount claimed has high value and vehicle claimed has low count although both have same distribution.



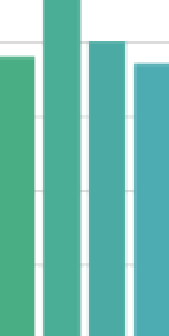
**OUTPUT**

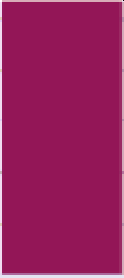
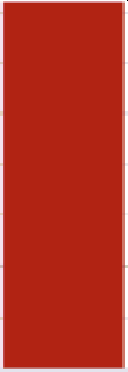
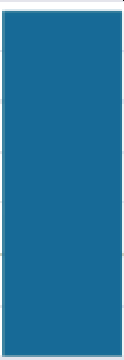


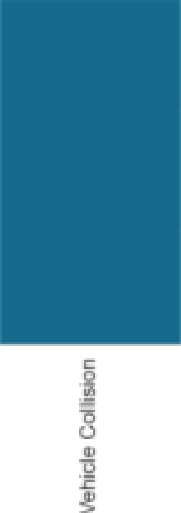
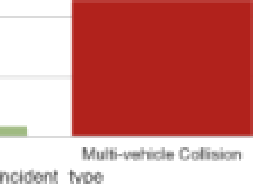
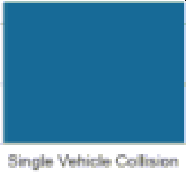


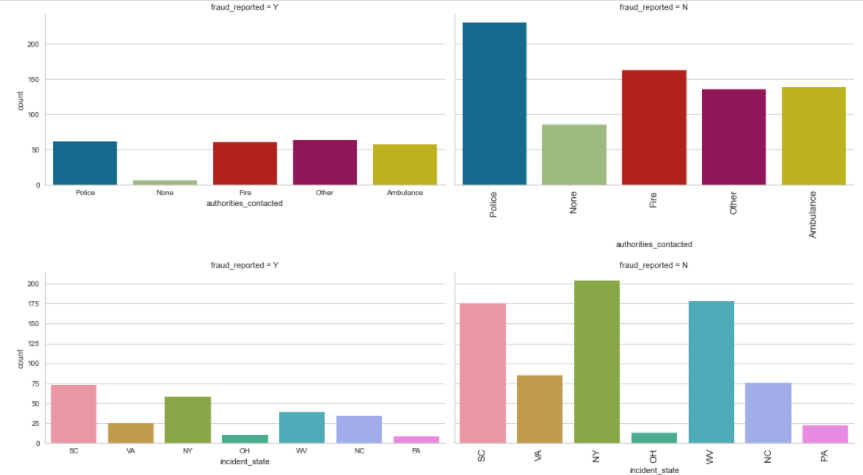


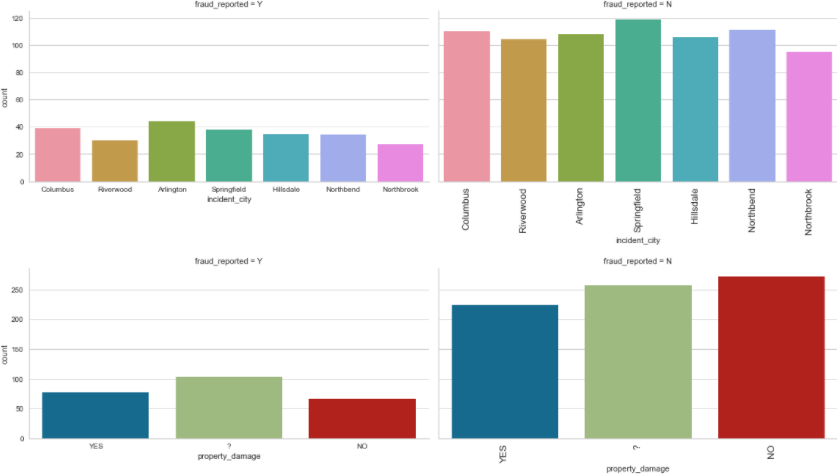
    

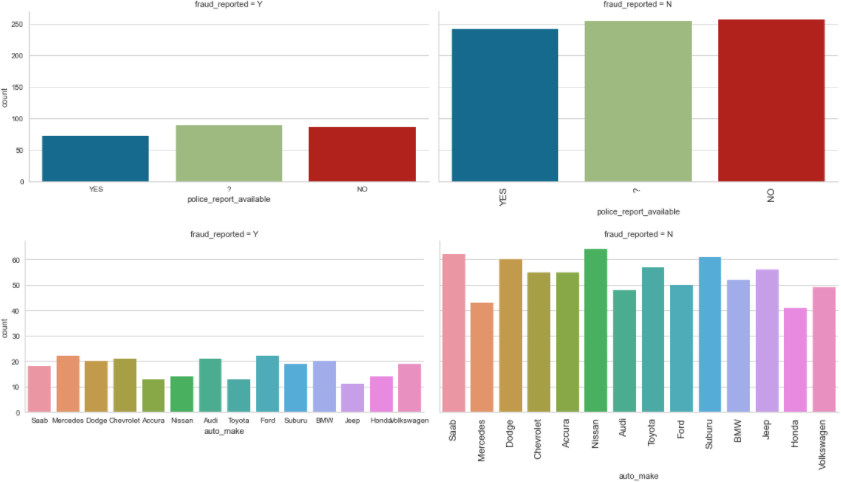
alll)

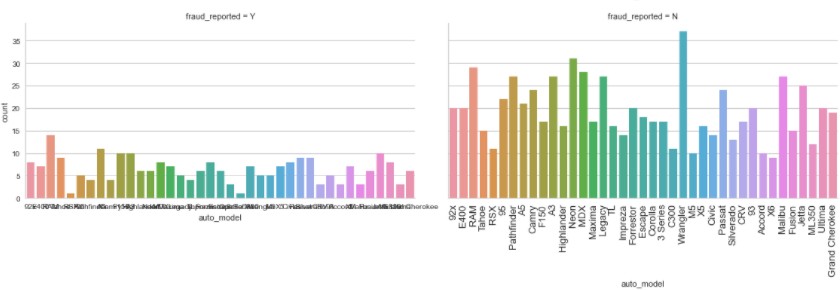












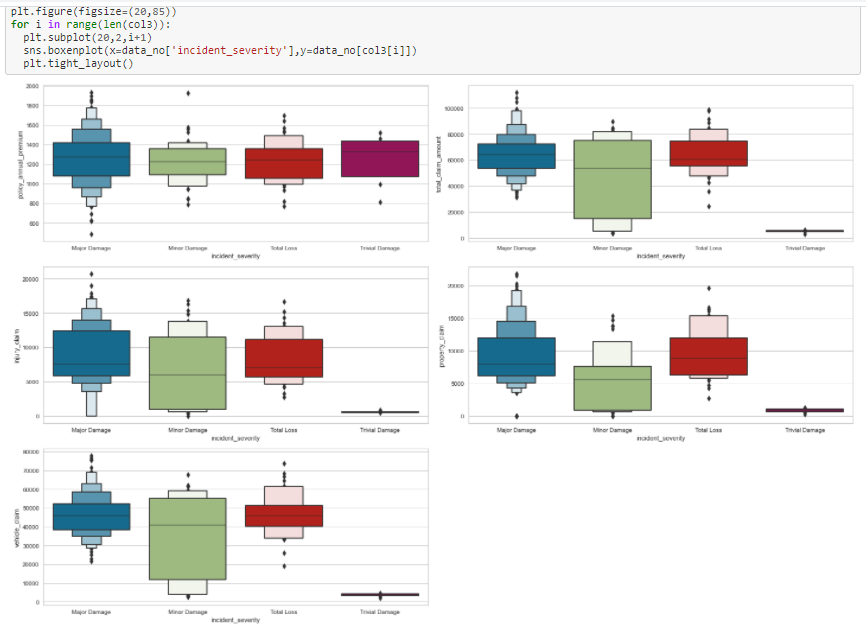
**OBSERVATIONS**

1. People who reported fraud have high value of CH of policy state.
2. People who reported fraud have equal count on both male & female. 3-Mostly People who reported fraud their hobbies are chess.
3. Mostly people who reported fraud mostly have relationship status other-relative.
4. Mostly people who reported fraud have incident type status single and multi-vehicle collision
5. Mostly people who reported fraud have collision type status of Rear collision and very less people who claim unknown (?)
6. Mostly People who reported fraud have incident severity status major damage and very less people who have status trivial damage
7. Mostly People who reported fraud have incident state SC who claimed more fraud
8. Mostly People who reported fraud have incident city ARLINGTON who claimed more fraud.
9. Mostly People who reported fraud have property damage? followed by YES.
10. Mostly People who reported fraud have automaker Mercedes, Ford followed by Audi.



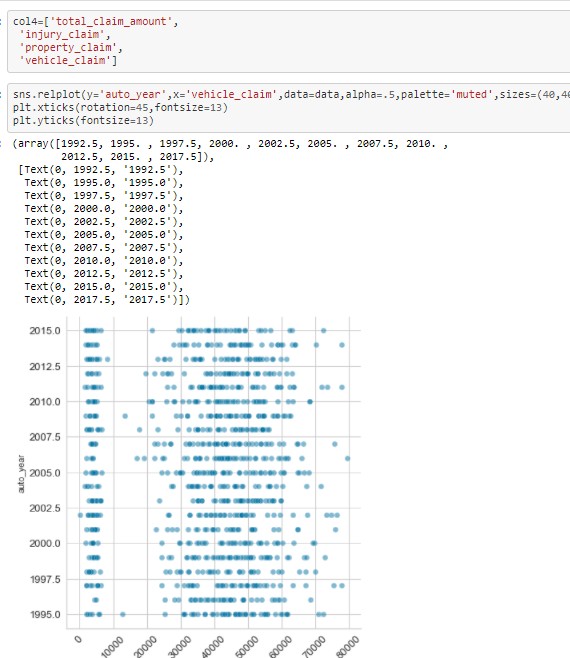
#### OBSERVATION

1. People who have claimed fraud insurance they have policy\_annual\_premium whose mean value is between 1300 and max value is approx. 1800 and where collision type is? People mostly have claimed 25% to 75% and for front and collision type people have claimed mostly 25% of the value.
2. From total\_claim\_amount vs fraud for side and front collision people have claimed maximum of 90000 and for front collision value is around 75000.
3. From injury claim where is fraud is yes Side and rear collision have same max value 15000 and mean value of 700 and for injury claim people have claimed mostly 75% to max value and less people are there who claimed for less than 25%
4. From property claim who have fraud side collision have 9000 mean value and 17000 max values and front and rear collision almost have approx. values and for side collision mostly people have claimed more than 50% of value same with front and real collision most of the people have claimed large value that is greater than 50%.
5. From vehicle claim where fraud is rear and side have almost same mean value and front side have a different value.
6. From the values that is mentioned above that means people have claimed these much of money from company in name of fraud means value represent the average value that people have claimed and max value represent the value that the maximum money that people have claimed in name of fraud.

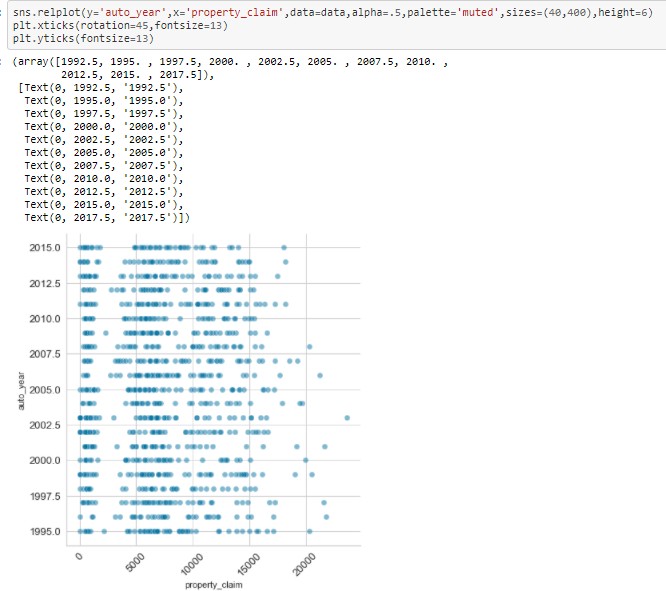


### OBSERVATIONS:

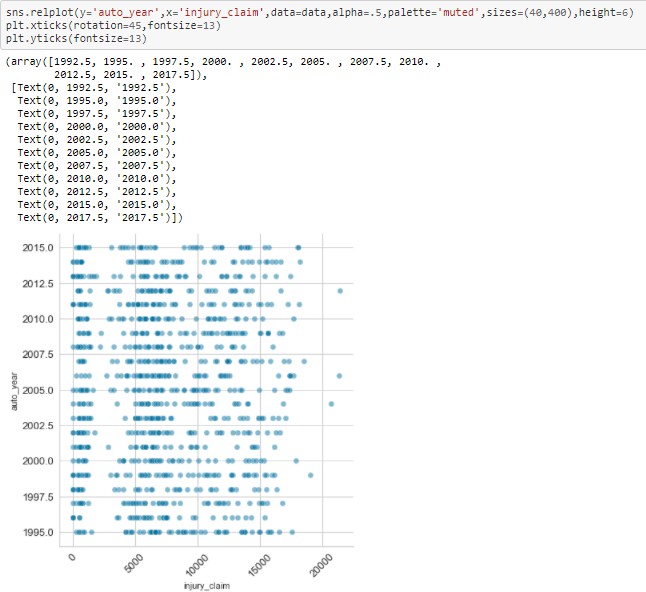
1. From policy annual premium minor damage most of the people have claimed value between 25-75%. Also, there are some outliers who have claimed for minimum and maximum values & for trivial damage most of the people have claimed for the value that is less than average.
2. From total claim amount for minor damage mostly people have value that is less than average.
3. From injury claim for major damage people have claimed more than the mean and for major damage most of the people have claimed 25% or 75% of value and for total loss most of the people have claimed more than mean value.
4. From property claim for major damage the most of the value lies above mean means for major damage and people get more money & for minor damage people get most of the value that is less than mean or average.
5. From vehicle claim for minor damage maximum people get money that is less than mean less no. of people get money that is greater than mean value.



Vehicle claim value is not proportional to year, it remains same by the years increasing.



Property claim value is not proportional to year, it remains same by the years increasing.

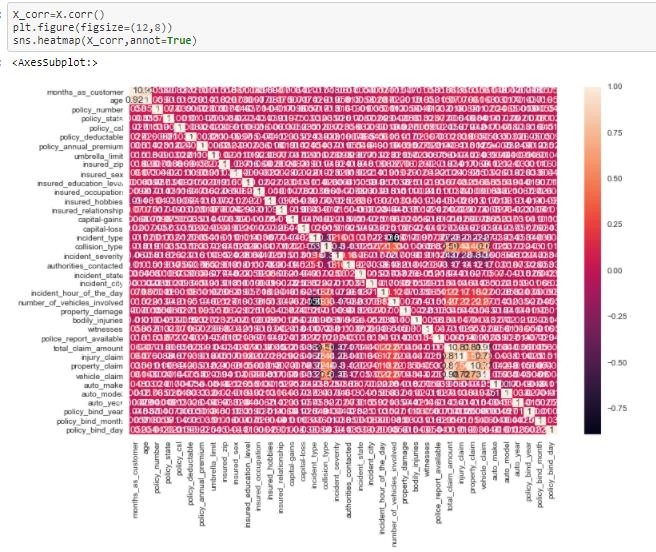


Injury claim value is not proportional to year, it almost remains same by the years increasing.



Total claim amount value is almost same as the year increasing.

CORRELATION GRAPH

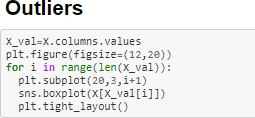


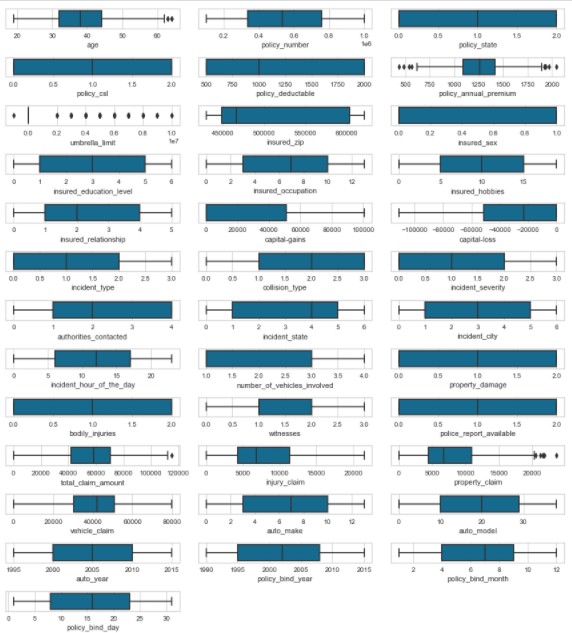
OBSERVATION:

From above heat map months as customer and age are having correlation more than 90% while rest of all features are having good correlation.

I should drop month as customer.

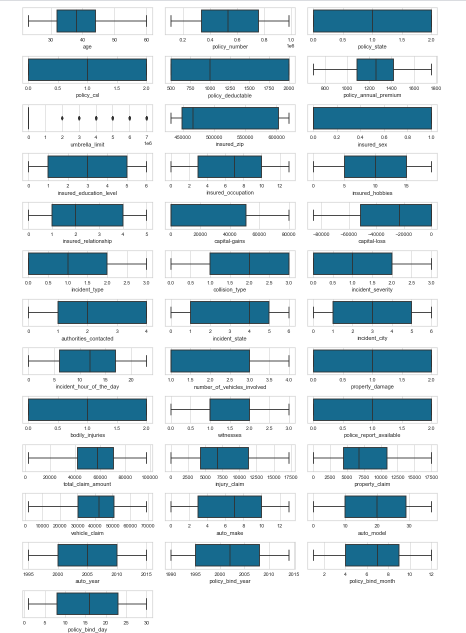
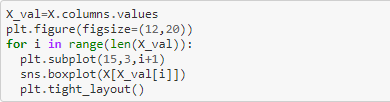
# OUTLIERS BEFORE REMOVING



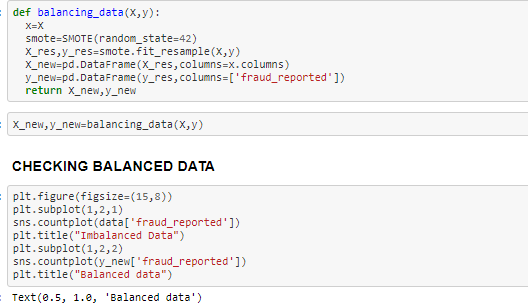


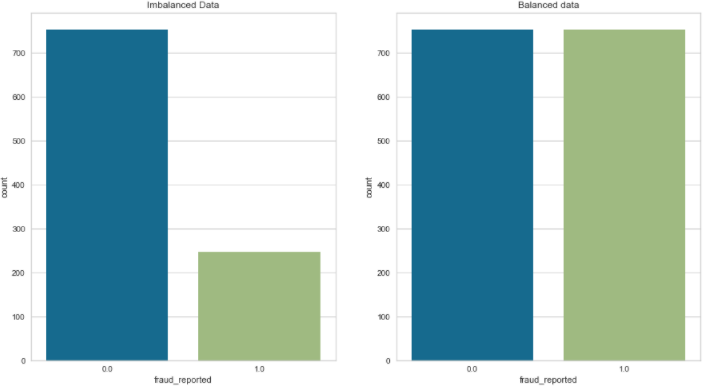
OUTLIERS SHOULD BE REMOVED

**AFTER REMOVING OUTLIERS**



**BALANCING DATA**





**SKEWNESS**

Skewness is a quantifiable measure of how distorted a data sample is from the normal distribution**.**

|  |
| --- |
| **age 0.461109** |
| **policy\_number 0.036105** |
| **policy\_state -0.026177** |
| **policy\_csl 0.088928** |
| **policy\_deductable 0.477887** |
| **policy\_annual\_premium -0.046551** |
| **umbrella\_limit 1.712094** |
| **insured\_zip 0.816445** |
| **insured\_sex 0.148630** |
| **insured\_education\_level -0.000148** |
| **insured\_occupation -0.058881** |
| **insured\_hobbies -0.061563** |
| **insured\_relationship 0.077488** |
| **capital-gains 0.437885** |
| **capital-loss -0.366324** |
| **incident\_type 0.101507** |
| **collision\_type -0.193345** |
| **incident\_severity 0.279016** |
| **authorities\_contacted -0.121744** |
| **incident\_state -0.148865** |
| **incident\_city 0.049531** |
| **incident\_hour\_of\_the\_day -0.035584** |
| **number\_of\_vehicles\_involved 0.502664** |
| **property\_damage 0.106418** |
| **bodily\_injuries 0.014777** |
| **witnesses 0.019636** |
| **police\_report\_available 0.052967** |
| **total\_claim\_amount -0.646051** |
| **injury\_claim 0.213656** |
| **property\_claim 0.247848** |
| **vehicle\_claim -0.674588** |
| **auto\_make -0.018797** |
| **auto\_model -0.088625** |
| **auto\_year -0.048289** |
| **policy\_bind\_year 0.050075** |
| **policy\_bind\_month -0.029321** |
| **policy\_bind\_day 0.017380** |
| **dtype: float64** |

We could see most of the skewness is present in "umbrella limit". It is with the ordinal data, so we will ignore the skewness.

**SLIP TEST AND TRAIN DATA**



**MACHINE LEARNING MODELS:**

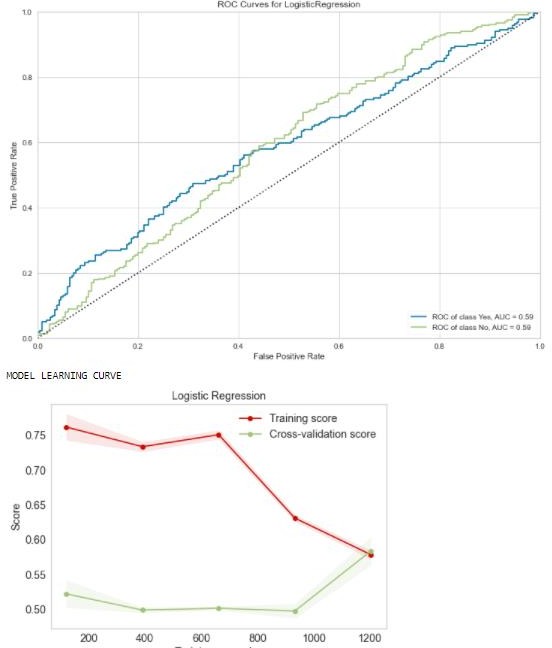
A machine learning model is the output of the training process and is defined as the mathematical representation of the real-world process. The [machine learning algorithms](https://www.educba.com/machine-learning-algorithms/) find the patterns in the training dataset, which is used to approximate the target function and is responsible for mapping the inputs to the outputs from the available dataset.

Since the dataset is large to my system configurations, ensemble techniques will be efficient although I’m testing the results with the below algorithms.:



###### Logistic Regression Model:

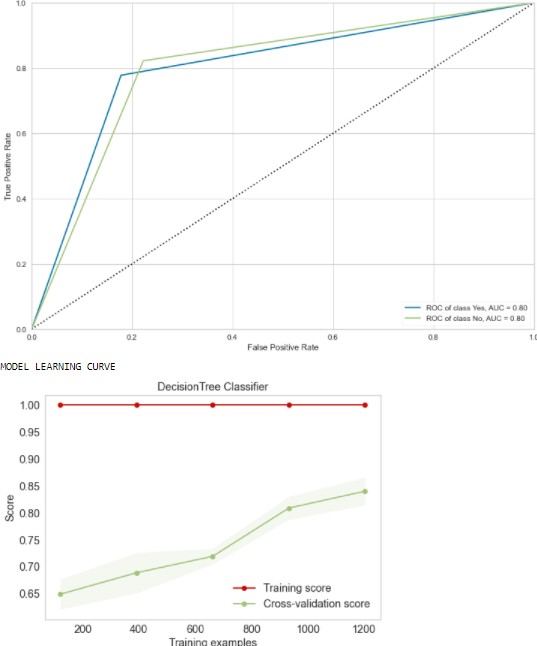
|  |
| --- |
| Accuracy Score 0.5685840707964602 |
| SCORE 0.5685840707964602 |
| Confusion metrics |
|  |
| [125 91] |
| [104 132] |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.55 0.58 0.56 216 |
| 1.0 0.59 0.56 0.58 236 |
|  |
| accuracy 0.57 452 |
| macro avg 0.57 0.57 0.57 452 |
| weighted avg 0.57 0.57 0.57 452 |
|  |
| Cross\_Val\_Score 0.5650810763239533 |



By **Logistic Regression Model,** we were able to get the accuracy score of 0.5685840707964602Cross\_Validation\_Score: 0.5650810763239533

###### Decision Tree Classifier Model:

|  |
| --- |
| Accuracy Score 0.8008849557522124 |
| SCORE 0.8008849557522124 |
| Confusion metrics |
|  |
| [168 48] |
| [ 42 194]] |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.80 0.78 0.79 216 |
| 1.0 0.80 0.82 0.81 236 |
|  |
| accuracy 0.80 452 |
| macro avg 0.80 0.80 0.80 452 |
| weighted avg 0.80 0.80 0.80 452 |
|  |
| Cross\_Val\_Score 0.8426173241512839 |

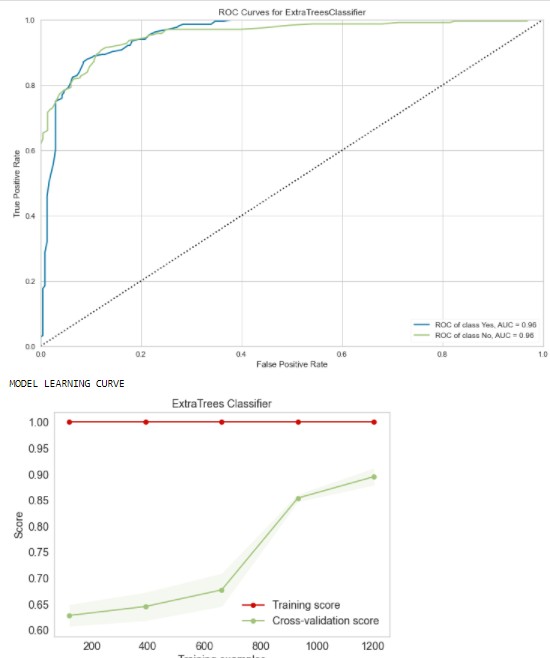


By **Decision Tree Classifier,** we get the accuracy score of 0.80088

Cross\_Validation\_Score: 0.8426

###### Extra Trees Classifier

|  |
| --- |
| Accuracy Score 0.8915929203539823 |
| SCORE 0.8915929203539823 |
|  |
| Confusion metrics |
|  |
| [191 25] |
| [ 24 212] |
|  |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.89 0.88 0.89 216 |
| 1.0 0.89 0.90 0.90 236 |
|  |
| accuracy 0.89 452 |
| macro avg 0.89 0.89 0.89 452 |
| weighted avg 0.89 0.89 0.89 452 |
|  |
| Cross\_Val\_Score 0.8877758465160281 |



By **Extra Trees Classifier model,** we were able to get the accuracy score of 0.8915929203539823. Cross\_Validation\_Score: 0.8877758465160281

###### Random Forest Classifier

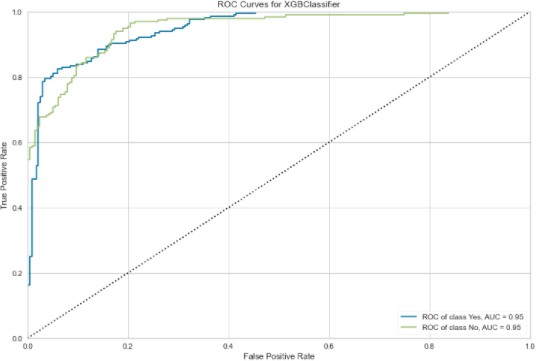
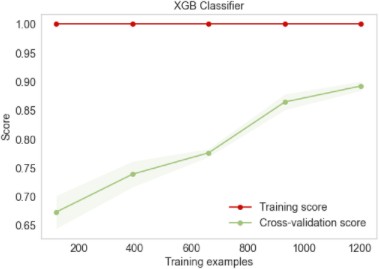
|  |
| --- |
| Accuracy\_Score 0.8628318584070797 |
| SCORE 0.8628318584070797 |
| Confusion metrics |
|  |
| [186 30] |
| [ 32 204] |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.85 0.86 0.86 216 |
| 1.0 0.87 0.86 0.87 236 |
|  |
| accuracy 0.86 452 |
| macro avg 0.86 0.86 0.86 452 |
| weighted avg 0.86 0.86 0.86 452 |
|  |
| Cross\_Val\_Score 0.8725066555191304 |
|  |

By **Random Forest Classifier,** we were able to get the accuracy score of 0.86283.

Cross\_Validation\_Score: 0.87250

###### XGB Classifier Model:

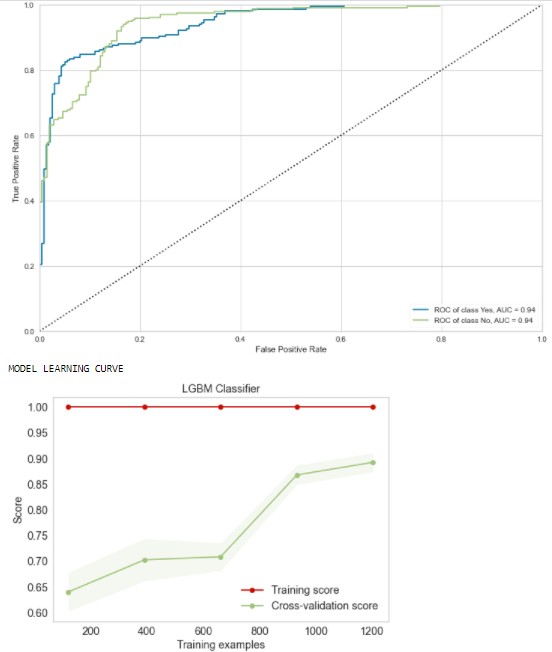
|  |
| --- |
| Accuracy\_Score 0.8716814159292036 |
| SCORE 0.8716814159292036 |
| Confusion metrics |
|  |
| [180 36] |
| [ 22 214] |
|  |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.89 0.83 0.86 216 |
| 1.0 0.86 0.91 0.88 236 |
|  |
| accuracy 0.87 452 |
| macro avg 0.87 0.87 0.87 452 |
| weighted avg 0.87 0.87 0.87 452 |
| Cross\_Val\_Score 0.894411564102 |

 By **XGB Classifier model,** we were able to get the accuracy score of 0.87

Cross\_Validation\_Score: 0.89

###### LGBM Classifier Model:

|  |
| --- |
| Accuracy\_Score 0.8805309734513275 |
| SCORE 0.8805309734513275 |
|  |
| Confusion metrics |
| [183 33] |
| [ 21 215] |
|  |
| CLASSIFICATION REPORT |
|  |
| precision recall f1-score support |
|  |
| 0.0 0.90 0.85 0.87 216 |
| 1.0 0.87 0.91 0.89 236 |
|  |
| accuracy 0.88 452 |
| macro avg 0.88 0.88 0.88 452 |
| weighted avg 0.88 0.88 0.88 452 |
|  |
| Cross\_Val\_Score 0.8857890915491409 |



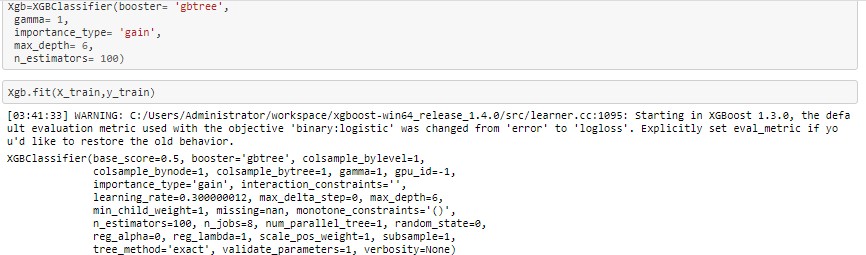
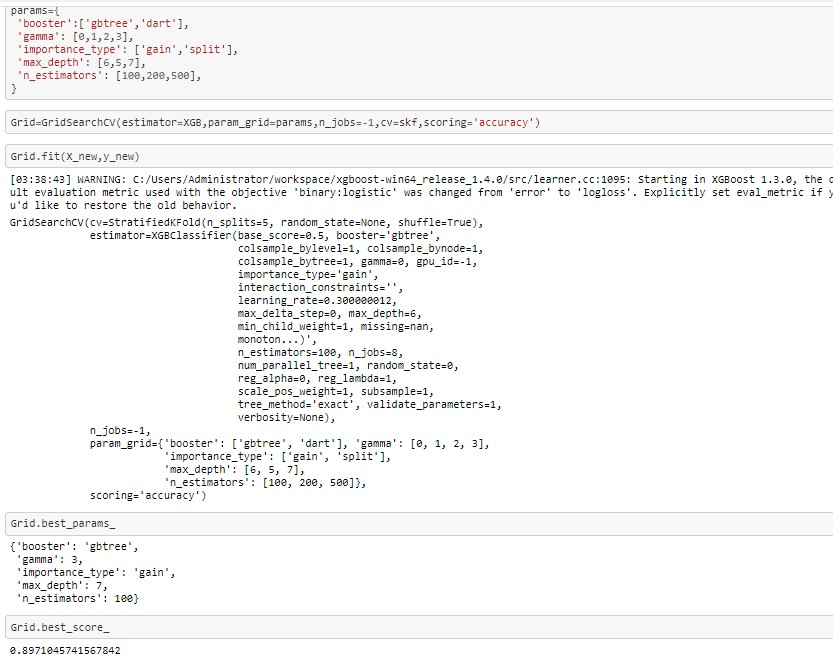
By **LGBM model,** we were able to get the accuracy score of 0.88053. Cross\_Validation\_Score: 0.8857.

XGM CLASSIFIER has high accuracy score" and with least difference between the Accuracy Score and the Cross-validation score

**"XGB Classifier" is our best model with 87.34 % Accuracy Score.**



**Hyper-parameter tuning**



pr1nt( "Roe Auc cuRvE- )

plt.Figure(IigsiZe=(12,8))

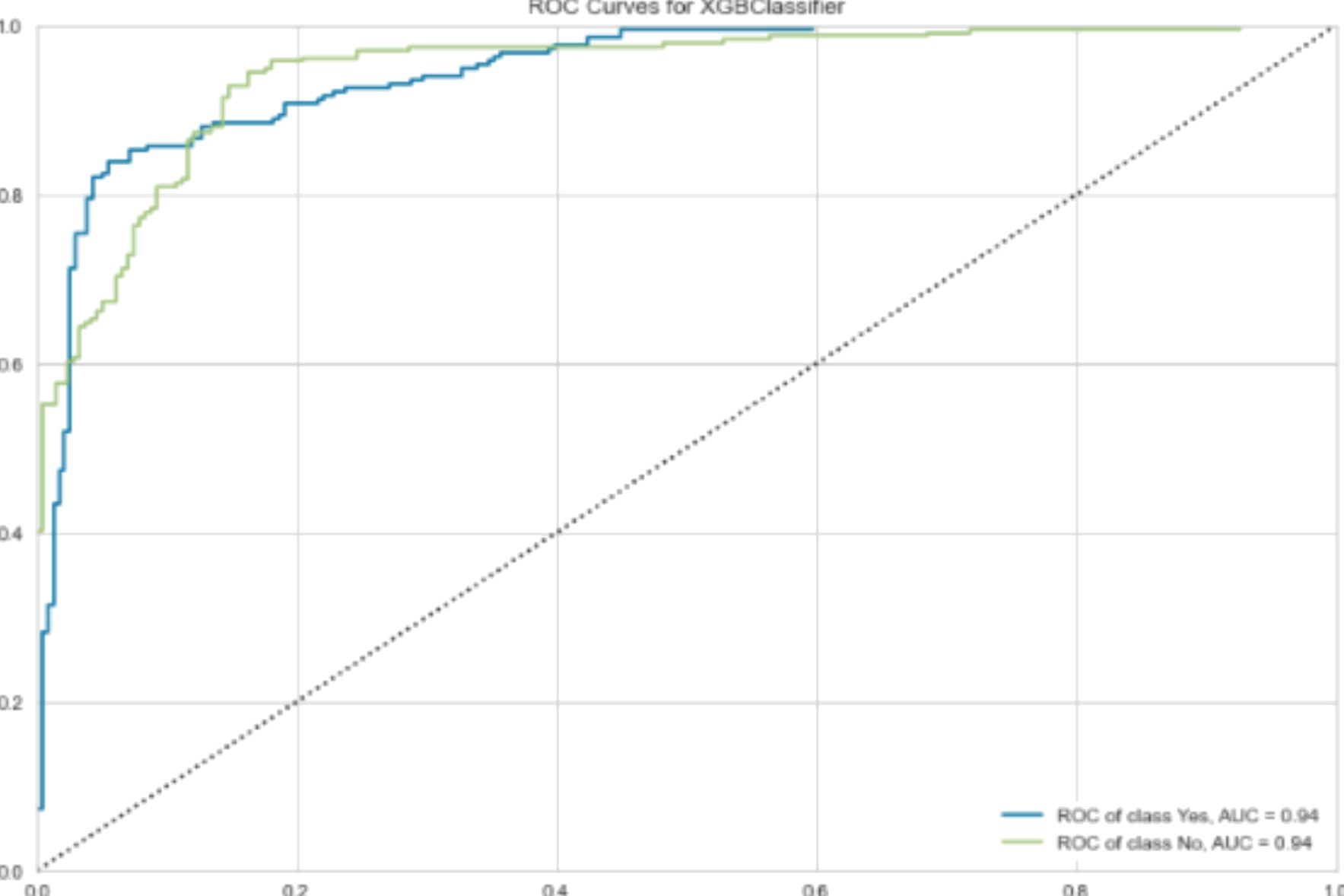
roc\_auc (xgb,x\_trai n,y\_tr ain,x\_te st=x\_t es t,y\_test=y\_test, classe s= [ ' Ye s' , ' No ' ] , n1c ro=FaI se, nacro=F alse)

pr1nt ( "MZ'DEL LEARN iND CURVE " )

skplt.estimators.p1ot\_learnin curve(xgb,X\_n9w,y\_n9w,cv=skf,scoring=’accuracy')

plt.shew()

ROC AUC CURVE



Pg@EL LEARNIfJG CURVE



Final model meoics

y\_pred1cted=xgb. pred1ct(x\_test)

pr1nt( "occur acy\_scare ",accuracy\_score{y\_test,y\_pred1cted})

pr1nt( "cvs" , cross\_va1\_score(xgb,x\_neu,y\_neu, scori ng= ' accuracy ' , cv=skf) . mean () ) pr1nt( "con-Euston metr1c s" }

pr1nt( ' \n ' )

pr1nt( confus1 on\_oatr1x(y\_test, y\_predicted) ) pr1nt( ' \n ' )

pr1nt( "c1ass lflcat1on Repart *"* )

pr1nt( "\n" )

pr1nt( c1assif1cation\_repoW(y\_test:,y\_pred1cted ) )



confusion metrics

C1ass1f1catlm Repos

preci.s ion reca LL £a -score support

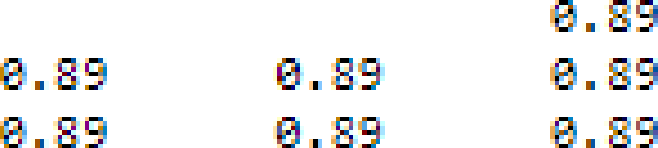
e.e

1.B



weighted avg

B.92 B.87



216









Saving model

: port jobllb

joblib.dump(Zrid,”AutomobiIe\_insurance\_fraud.obj“)

: [ 'Auto■obI 1e\_1nsurance\_€raud. obj ' §

## CONCLUSION

As you can see, financial fraud and machine learning are practically inseparable at present times. By applying various rules and synthetic algorithms, it becomes just a perfect technology for automated financial fraudulent detection.

Unlike the traditional system of analysis, which is mostly performed by human decisions, it allows covering much more information and processes the big data in shorter periods of time, thus saving lots of investments, resources, and time for the financial units.

Fraud detection using machine learning allows creating new rules and more complex algorithms for analyzing various transactions and suspicious financial behavior thus minimizing the risks of financial loss.