**INSURANCE CLAIMS FRAUD DETECTION**

**What are we going to discuss today?**

Today we have been offered insurance for almost everything, you know what even for life, e.g. Health Insurance, Property Insurance, Automobile Insurance, even Insurance for mobiles, Life Insurance, Electronic gadgets what-not, the list goes on…, Let us know the challenges faced by the insurance companies.

What even insurance companies have challenges? Why not every business model has its own loop holes and challenges. Let us see the challenges and how can we help them with our knowledge of Machine Learning and Data Science.

**Business Problem:**

The most common and biggest problem faced by Insurance Companies is FRAUD CLAIMS, which is a huge loss to company. So we will try to find out whether the claim is Genuine or Fraud based on the data from an Auto Insurance Company by building a Machine Learning model.



Image source: Google – Credits to respected owner created

Before getting started with the project in building Machine Learning model, even before viewing the data it is always best practice to know about the problem / business domain and have some idea of it. Which is helpful in our entire project phase. So I recommend to have a quick check with our best buddy, you’re right google the best source for basic insights.

**Glimpse of Automobile Insurance Domain:**

In an automobile insurance the company have some certain terms and conditions, checklists where the insurance will be covered and where it doesn’t take responsibility or pay any amount if it doesn’t meet their terms like : Not eligible for claim in case of accident when drunk and driven, etc.

*The Fraud Claim:* It is a false claim by user even if they don’t meet the terms of company but will show as the Genuine and make claim to the company that they followed all their terms and is eligible for claim for their own benefits / to not occur loss.

**DATA SOURCE:**

The source of data to me is from DATA TRAINED Institute.

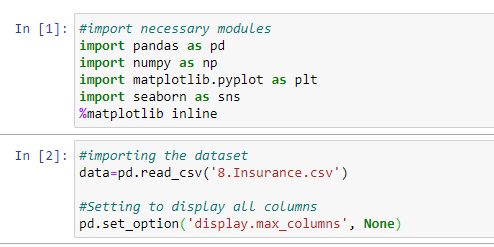
File format: .csv

The data is collection of different parameters and checks with the claim status whether fraud or not collected from different insurance companies and compiled into one file.

**Step by Step process:**

**1. Import the dataset:**

Importing the dataset in jupyter notebook and saving to a variable



Cell [1]: we imported some required modules to get data and create visualizations.

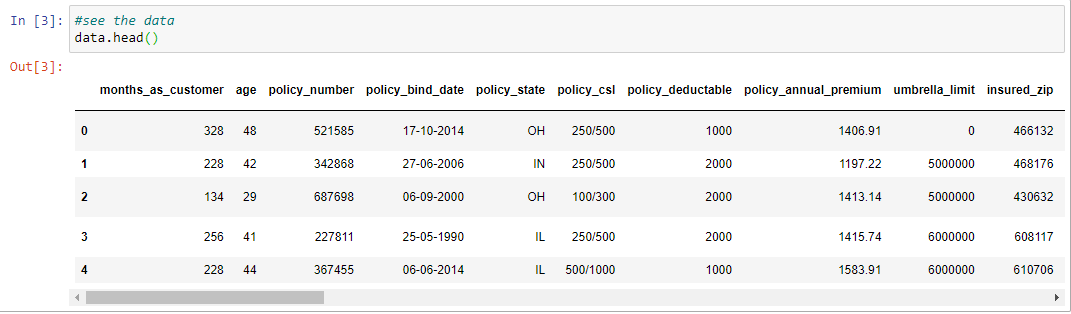
Cell [2]: We imported dataset into variable ‘data’,

8. Insurance.csv is the file name which I stored in my local machine at current working directory.

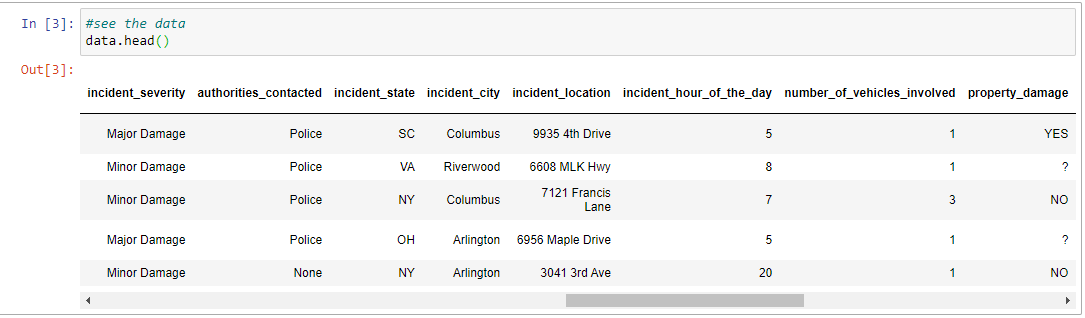
Then we set an option present, to display all the columns of the dataset when we viewed it.

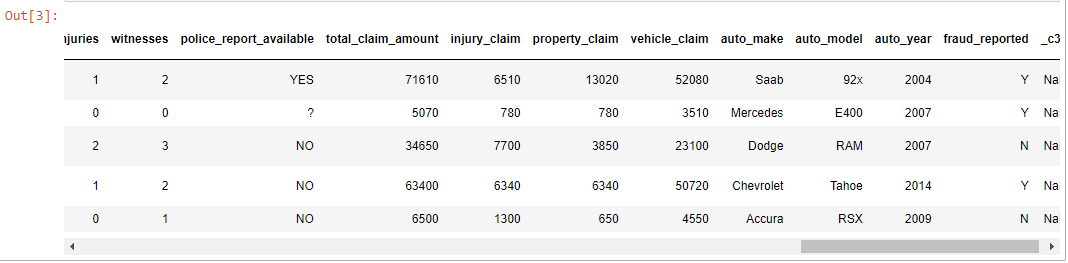
**Data Analysis:** Narrow method of checking assumptions for model fitting, hypothesis testing by cleaning, filling missing values and making transformations as required.

**Viewing the data:**



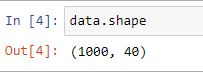
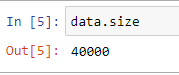






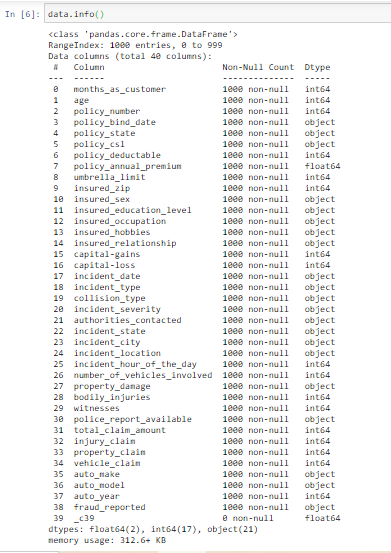
We are now seeing the top 5 rows of our dataset. The data we have consists of columns with values categorical and numerical type also few columns with ‘?’ in them.

Let us see the shape and size of dataset:

There are 1000 rows and 40 columns, with a size of 40000 values (product of rows and columns)

**Column details:** Important step in EDA

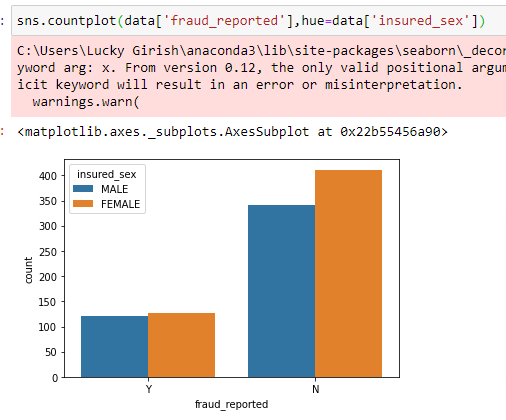


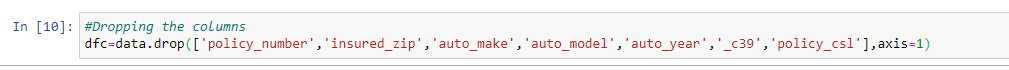
We got the brief of the data columns like – column name, number of non-null values and its data type. This is important step to see whether there are any datatype mismatches and to fix it. So we need to carefully observe the type of data present and verify here.

Observations:

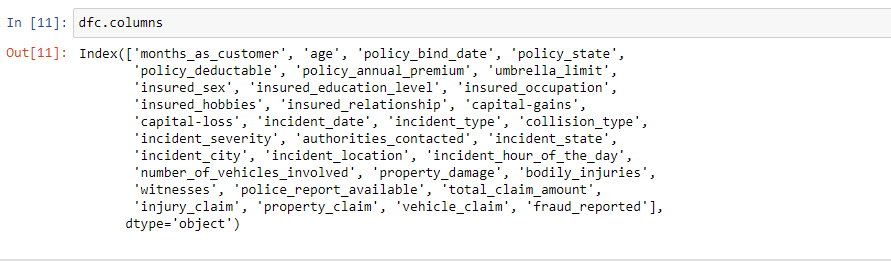
* Our data is mixture of integer, float and object data types.
* Our target variable ‘fraud\_reported’ is object type. So we use ‘Classification Model’.
* We see there are no null values in our brief data, but we found ‘?’ in some columns, so we need to replace them in the best possible way.
* Column policy\_csl is shown as object but we can see it contains numerical data. There might be chances of ‘?’ or any other special character or text present in that column. So it is shown as object type. We need to find that and convert its type.
* Observe column 39: ‘\_c39’- it has 0 non null values, i.e. all the values are null. So we can drop that column straight away.
* **Assumptions:** 
  + - * Assuming that, for making a fraud it is personal intention and that doesn’t depend on what kind of auto make, model and year. So I opt to drop these columns also.
      * Policy\_number: It is a unique id given to the customer, to track the subscription status and other details of customer. So we can drop this column, before dropping check if there are any duplicates in that column, if there are duplicates remove those rows by keeping one uniquely then drop the column.
      * Insured\_zip: It is the zip code where the insurance was made. It doesn’t give any information or useful to us.so we will drop this column too.
      * Policy\_csl is basically [Combined Single Limit](https://www.iruhl.com/blog/understanding-combined-single-limit-csl-split-limit-liability/), i.e. how much credible and what their premium covers, it is also not necessary. So we will drop it.

We have seen that fraud reported by male & female both are almost same count.



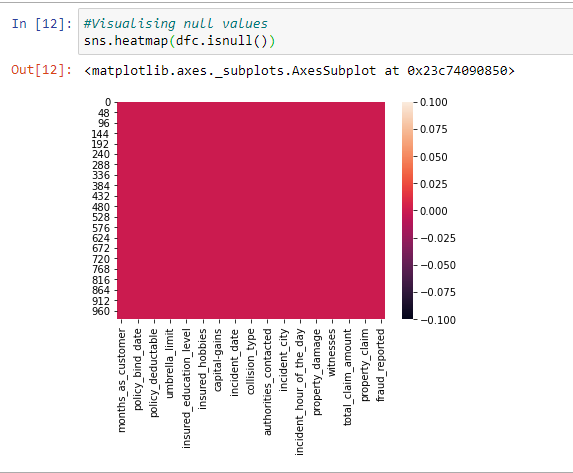


We dropped the columns and copy of it saved into new variable ‘dfc’, the initial variable data will have all columns. Where as in dfc we have columns that aren’t dropped.



This gives list of columns present in the variable dfc.

Checking for null values:



There are no null values. We are going to replace the ‘?’ with mean, median or mode or any other based on the column.

We see ‘?’ present in many columns:

for e.g. we deal with one column collision type-

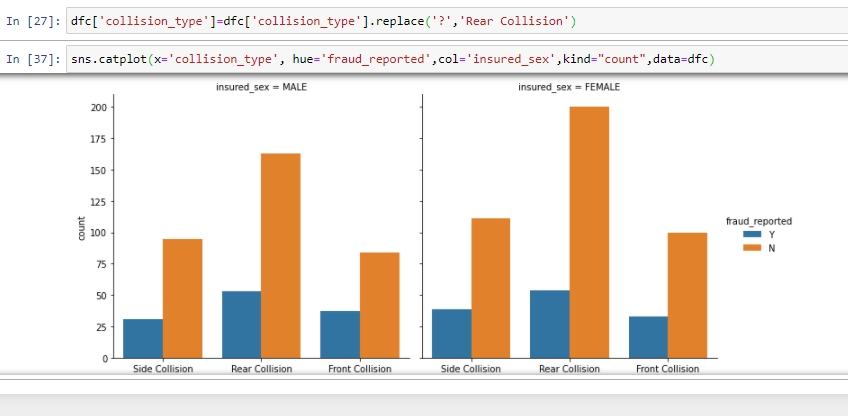
We can see there are 4 unique values in this column, as shown below. Front, Rear, Side and ‘?’ Collisions. ‘?’ is there in 178 rows. This might not be filled, it is a manual error. Because the collision should be in one of the category, and must note while claimed.

The best, thumb hand rule to deal this kind in case of categorical values is to fill with the mode of column.

So we will replace these ‘?’ with the most repeated type of collision type: Rear Collision

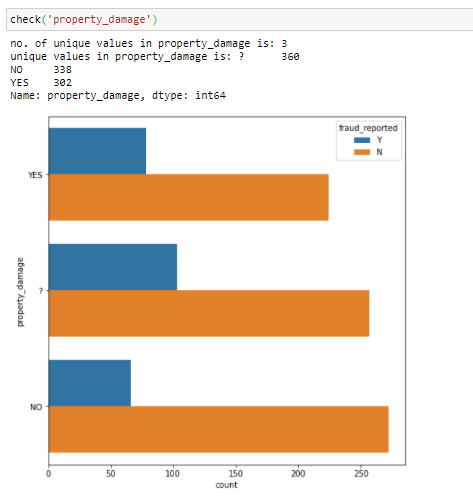


Replacing and counts after replaced wrt to fraud claim or not:

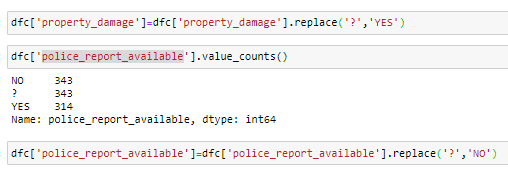


More collisions done by females than males. Rear collisions are max. From above we can also see, what kind of collision made by male or female and whether it is fraud claim or not. Almost count of all the fraud claims made are same irrespective of gender. Side collisions reported by females are fraudulent.

Similarly we do fill the ‘?’ with mode of that column or by analysing the column. I had filled them with mode.



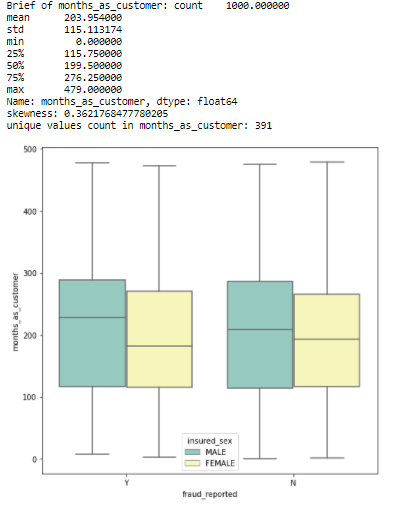
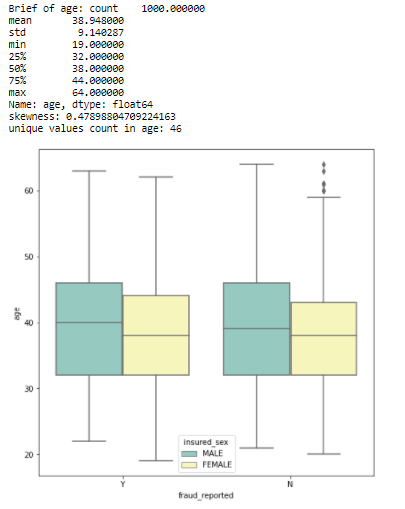
In Property damage ‘?’ is replaced with YES. Since we observed that ‘?’ have more fraud claims. By comparing with other data, claims with yes are more fraud. So we replaced that with YES.



Similarly police report is also filled with NO

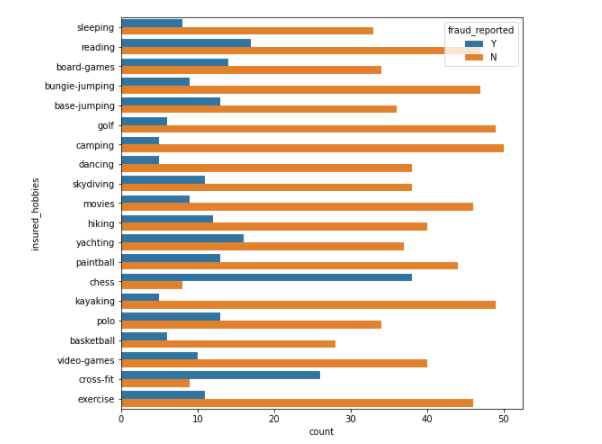
Let us see some important observations:

Average age and months being a customer: 38.9 yrs. old, 203.95 months as customer

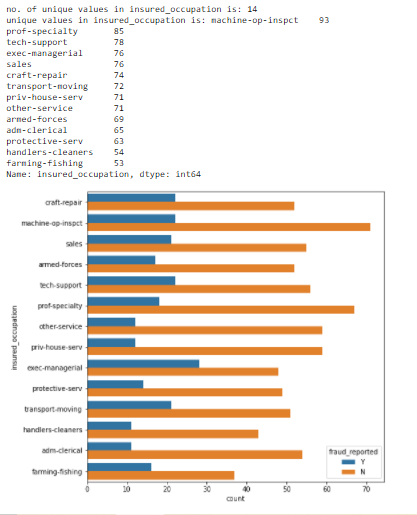
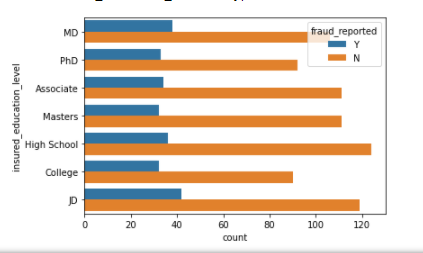


We can see males had done frauds from the age of 22/23 approx., females had done from teens, approx. from 18 yrs. Frauds committed by clients from their 1st month onwards.

Hobbies: People with hobby of chess and followed by cross-fit had done more fraud claims.

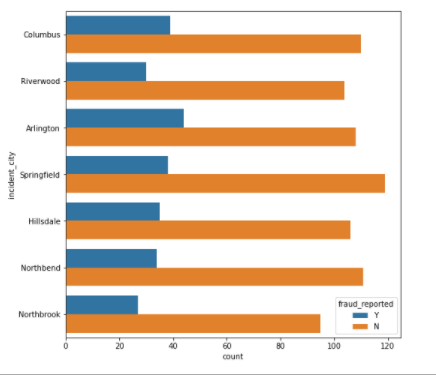
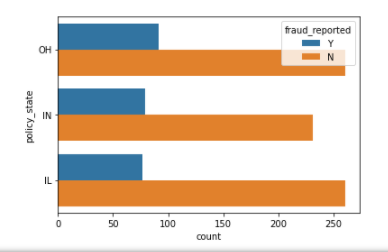


Education level: We can see, those who studied JD, MD has done more frauds comparatively.

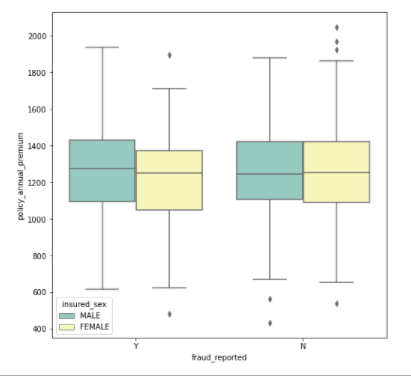


Occupation: People of job exec.manager are doing more frauds.

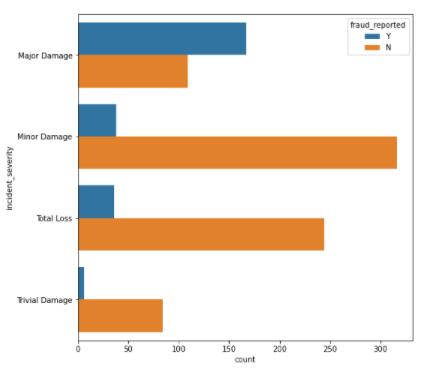
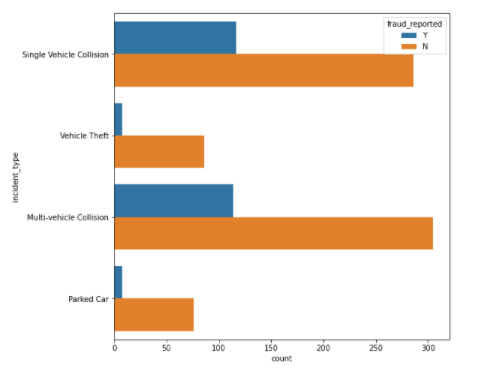
More frauds happened in state OH and in cities: Arlington, Columbus



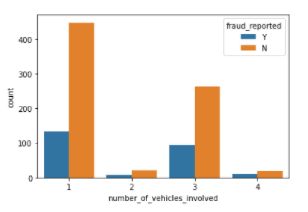
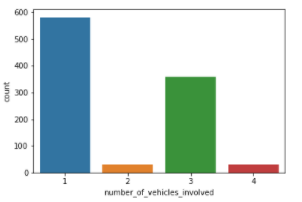
Policy premium: We seen that people with premium of above average are committing frauds



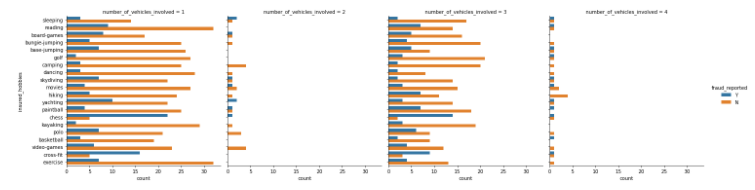
People who claimed incident type as Single vehicle collision or multiple vehicle and major damage collision are more fraudulent claims.

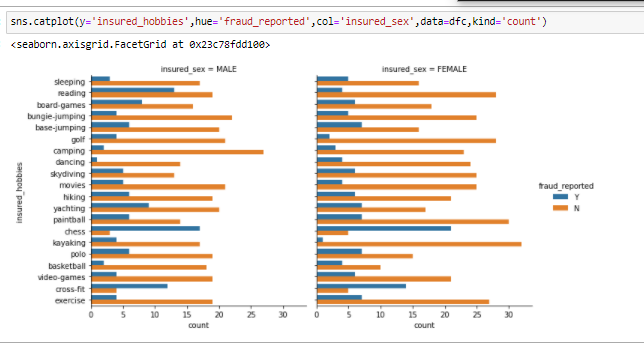


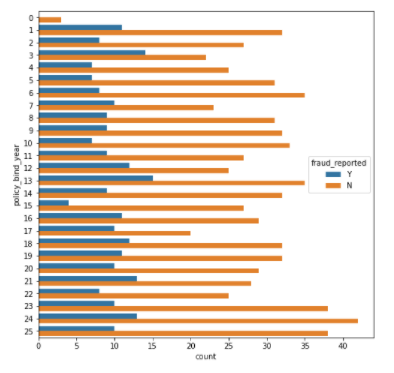
Majority of the incidents are 1 and 3 vehicles, also 1 involved are more fraud claims, but if we see carefully even though 2 and 4 vehicles involved claims are less, 50% of them are fraud.



Observing the fraud basis on no. of vehicles and Hobbies , Gender and Hobbies.







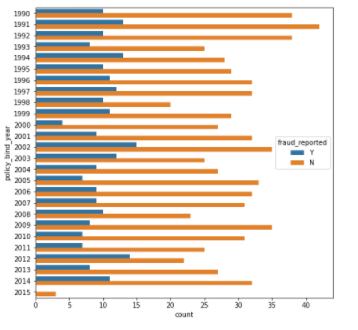
Based on number of years of policy binded

Similarly we observe and get insights from other columns.

We removed two more columns, incident date and incident location, which are not necessary, have almost unique values in each row.

We made a new column policy bind year from policy bind date and dropped the original column.

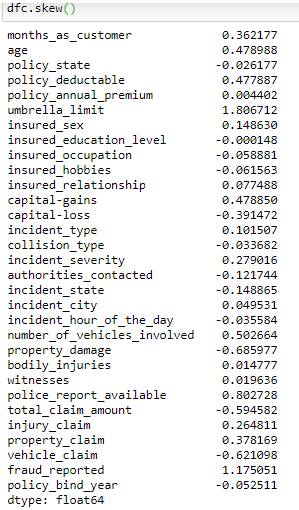
From that we observed policy been made from the years 1990 to 2015, and more frauds are of those who made insurance in year 1991, 2002, 2012.

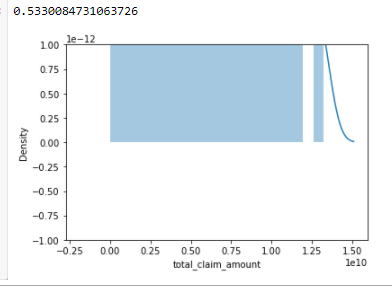
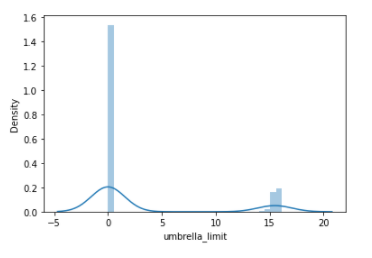


So to make it simple, we found the difference of number of years from 2015.

We made all the columns encoded and find the correlation.

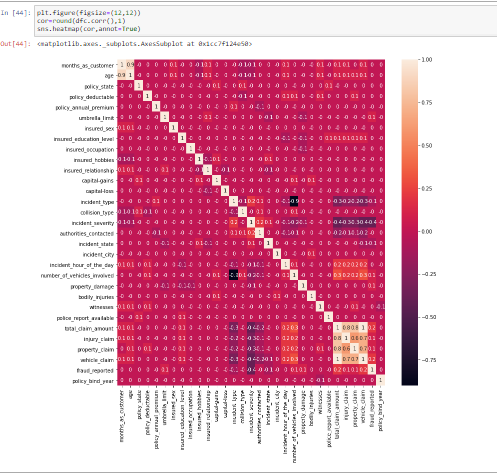
Checking the skewness and reducing skewness in columns umbrella limit, total claim amount





Skewness reduced to 1.4 from 1.8, Skewness reduced to 0.5 from -0.59

Plot of correlation:



Observations:

* Age & months as customer are highly correlated, we will drop the column age.
* Total claim amount, injury claim, property claim, vehicle claim are correlated to each other. So we will drop all those columns except Total claim amount
* Incident hour of the day doesn’t have any correlation with fraud. So we will drop it.\
* There are many columns which have no correlation with fraud reported

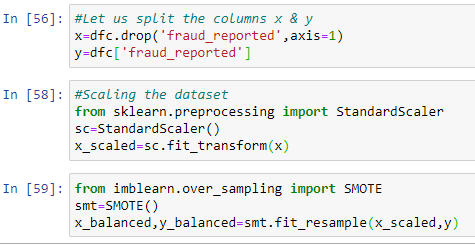
Let us drop these columns, find the VIF to know the multicollinearity and drop other columns.



By viewing VIF I dropped columns

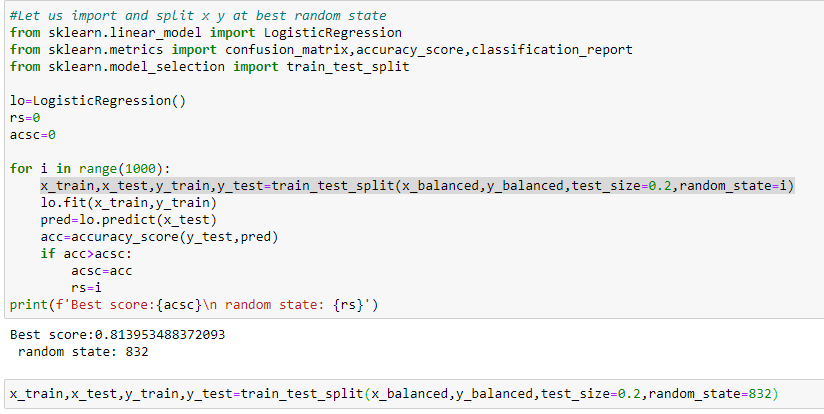
Number of vehicles involved, premium deductables, capital gain and capital loss.

Finally two last steps before making model, Split x & y, then scale the x and finally balance the dataset.

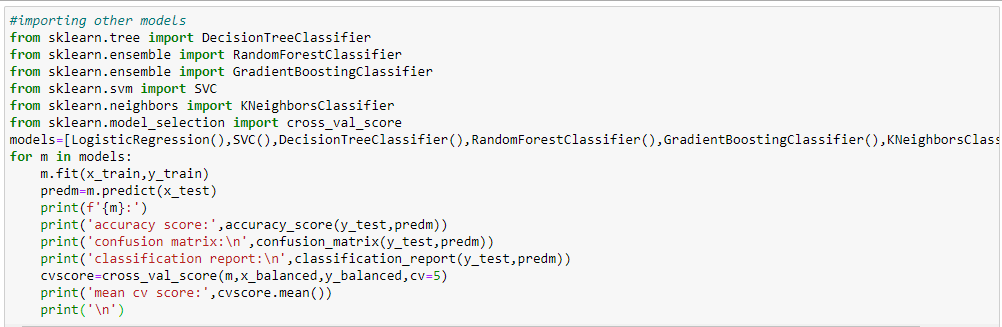


X is features, y is our target.

Found best random state & split.



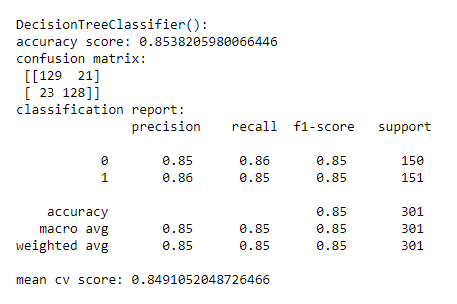
Imported different models to try and select the best one



From above we got results of accuracy score, f1 score, evaluation report and cross validation score of different models –

|  |  |  |  |
| --- | --- | --- | --- |
| Model | F1 score | Cross validation | Difference |
| Logistic Regression | 81 | 74.5 | 6.5 |
| Decision Tree Classifier | 85 | 84.9 | 0.1 |
| Random Forest Classifier | 91 | 87 | 4 |
| Gradient Boosting Classifier | 91 | 88 | 3 |
| K Neighbors Classifier | 72 | 73 | 1 |
| SVC | 89 | 84.5 | 4.5 |

From above we can see, Decision Tree is the best model with least difference between CV score and F1 score. These are the metrics generated:



From confusion matrix we can see



Our model is saying 23 members as not fraud where they are actually frauds. We need to reduce this for better results.

FP is ok because checking whether fraud or not is less cost to company than paying huge amount to the original fraud claim.

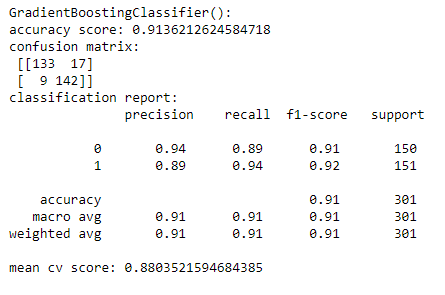
In our case we need least False Negative.

23

21

But wait there is almost 6% difference in Decision Tree classifier, Random Forest or Gradient Boosting. Let us observer their metrics once

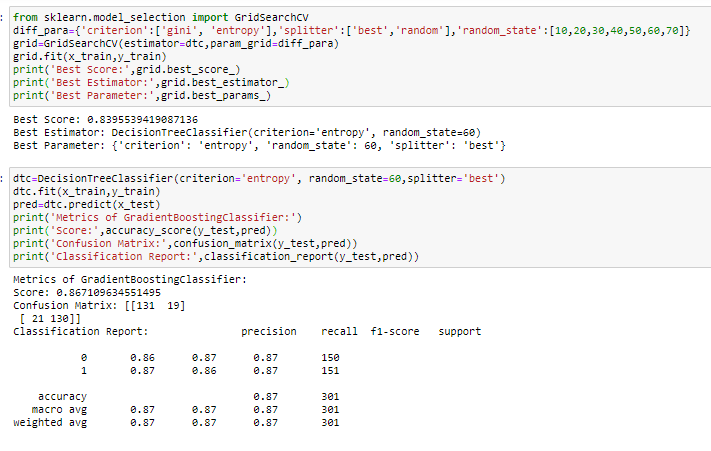
If we see other models: It has 3% difference but giving the least FN

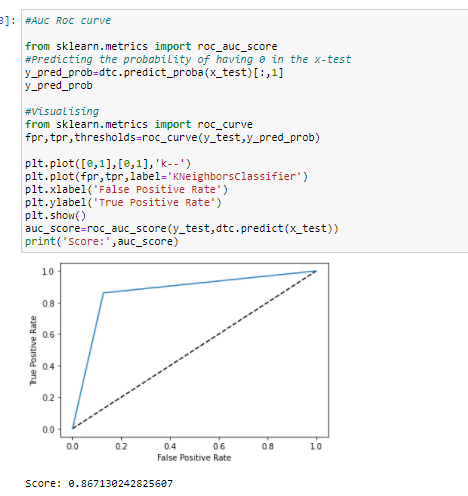
17

9

By tuning the decision tree classifier: The model accuracy increased from 85 to 87% only. With FN as 21.

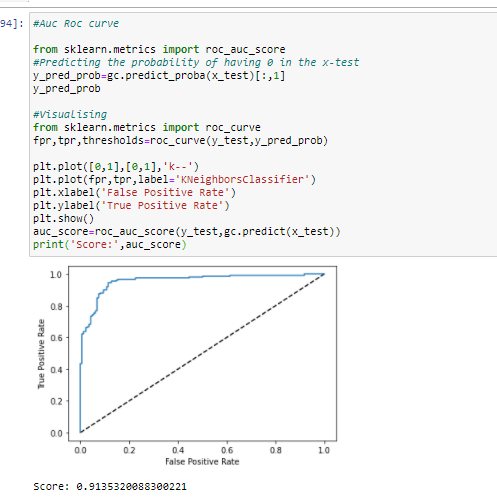


AUC-ROC:



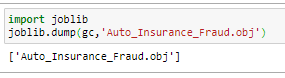
For the Gradient Boosting model, even after tuning there is no improvement in accuracy:

AUC-ROC:



I am considering this as my best model and saving it to predict.

Saving the model:



Conclusion:

Now we made a model, that can predict whether the claim made is Fraud or not with a 91% accuracy.

We have Seen, Gender, number of months as customer, Premium, Hobbies, Education , Type of incident , Incident severity and the policy bind year are the most important columns that help us to predict and understand whether the claim is fraud or not.

* Customers with education of JD are doing more fraud.
* Customers with hobbies of CHESS, CROSS-FIT are doing more fraud.
* Customers who claimed the incident to be single vehicle collision or Multi vehicle collision are more fraud claims.
* Gender is almost same in both the cases of fraud claims
* Premium – Customers who opted for higher premiums done frauds
* Majority of claims made by customers with policy bin year 1991, 2002, and 2012 are frauds.

We encoded the values, categorical columns. Our result is also encoded as 0 and 1 where

0 is NO

1 is YES

Later it will be decoded, when the model is deployed.

Attaching the GitHub link along with this, where you can get dataset, the complete step by step procedure followed jupyternotebook .ipynb file.

https://github.com/Makham-Sai-Girish/Auto\_Insurance\_Fraud-Claim\_detection