# Diving Into Data Analysis: by Sapna Jolly



## Insurance Claim Fraud Detection Project

**In association with Data Trained Academy: Batch: 1838**

**Introduction:**

Fraud is one of the largest and most well-known problems that insurers face in the insurance industry. This article focuses on claim data of automobile insurance.

Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for the purpose of financial gain. Fraud may be committed at different points in the transaction by applicants, policyholders, third-party claimants, or professionals who provide services to claimants. Insurance agents and company employees may also commit insurance fraud. Common frauds include “padding,” or inflating claims; misrepresenting facts on an insurance application; submitting claims for injuries or damage that never occurred; and staging accidents.

**THANKS …** to Data Science and Machine Learning, which has been very useful in many industries that have managed to bring accuracy or detect negative incidents. Here in this blog, I have created a Machine Learning model to detect if the claim is fraudulent or not. Here various features has been used like, insured information, insured persons, personal details and the incident information. In total the dataset have 40 features. So using all these previously accuired information and analysis done with the data I have achieved a good model that has 92% accuracy. So let’s see what are the steps involved to attain this accuracy.

Various visualization techniques have also been used to understand the co-linearity and importance of the features.

*Note: Various jargons are used in the article assuming the fact that the reader is aware of the language used in data science.*

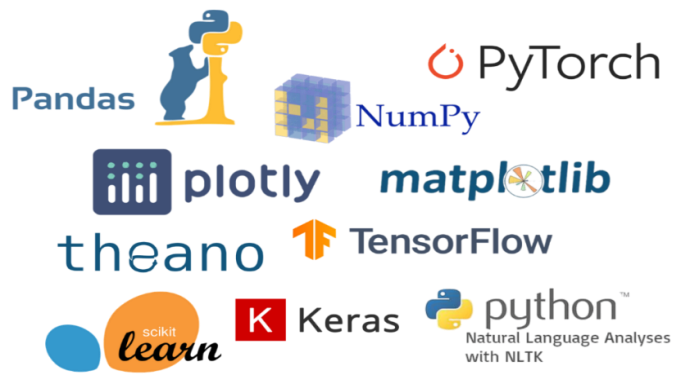
# Hardware & Software Requirements & Tools Used:

### Hardware required:

* Processor: core i5 or above
* RAM: 8 GB or above
* ROM/SSD: 250 GB or above

**Software requirement**:

* Jupyter Notebook

**Libraries Used**:

* Python
* Numpy
* Pandas
* Matplotlib
* Seaborn
* Date Time
* Scikit Learn

Heading forward we will try to understand the problem statement and the dataset.

**Problem Definition:**

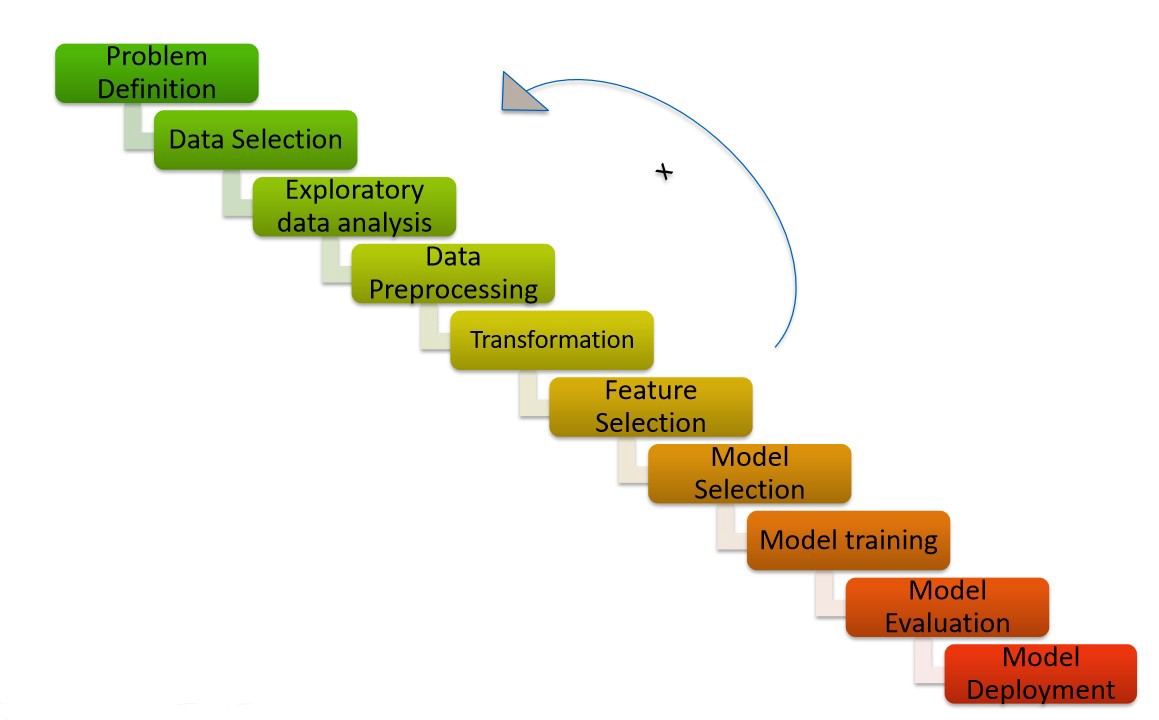
Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem .In this project; we are provided a dataset which has the details of the insurance policy along with the customer 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.

In this problem we will be looking into the insured person details and the incidents and analyze the sample to understand if the claim is genuine or not.

**Lets deep dive step by step in the data analysis process.**

In order to build a Machine Learning Model, we have a Machine Learning Life Cycle that every Machine Learning Project has to touch upon in the life of the model. Lets a sneak peek into the model life cycle and then we will look into the actual machine learning model and understand it better along with the lifecycle.

****

Now that we understand the lifecycle of a Machine Learning Model, lets import the necessary libraries and proceed further.

**Importing the necessary Libraries:**

To analyze the dataset or even to import the dataset, we have imported all the necessary libraries as shows below.

Pandas has been used to import the dataset and also in creating data frames.

Seaborn and Matplotlib has been used for visualization

Date Time has been used to extract day/month/date separately

Sklearn has been used in the model building

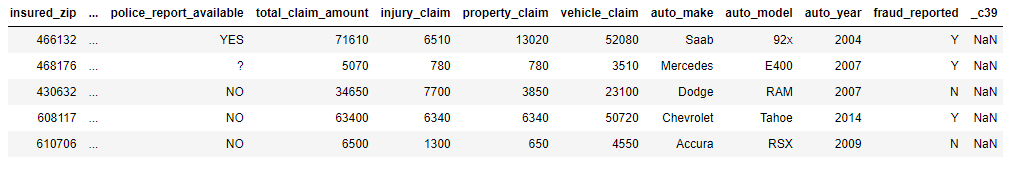
****

**Importing the Dataset**

Let’s import the dataset first.



I have imported the dataset which was in “csv” format as “df”. Below is how the dataset looks.



By observing, the dataset we could make out that the dataset contains both categorical and numerical columns. Here "fraud reported" is our target column, since it has two categories so it termed to be "Classification Problem" where we need predict if an insurance claim is fraudulent or not. As it is a classification problem, hence we will be using all the classification algorithms while building the model that we will see as the blog proceeds.

Also by doing a simple code ‘df.shape’ we also figured out how many rows and columns we have. We have got the result that we have 1000 rows and 40 columns. PCA can be done, however I decided not to lose any data at this time as the dataset is comparatively small and the first lesson of a data scientist is ‘Data is Crucial’ hence proceeded will all the data.



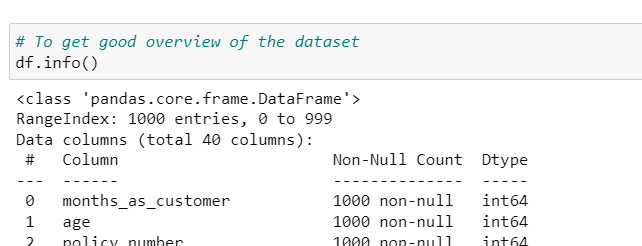
As per the lifecycle of the machine learning model we have already completed point 1 and 2. Now lets move on to the point 3,4, 5 and 6 which is the most crucial part of any machine learning model, as the best way possible data will be analyzed and cleaned the better model accuracy we will get, or the model can remain over fitting or under fitting. We will discuss further why all the steps are used.

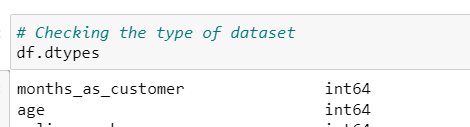
**Exploratory Data Analysis and Data Preparation:**

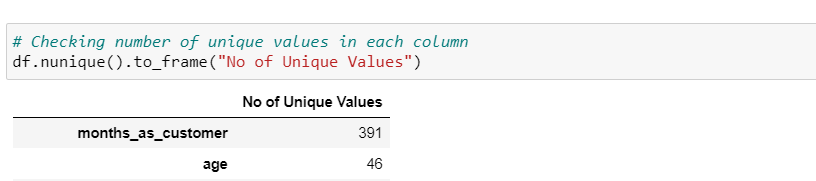
In this part we will firstly be exploring the data with some basis steps and then further proceed with some crucial analysis, like feature extraction, imputing and encoding.

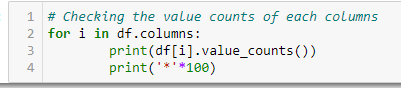
Let’s start with checking shape, unique values, value counts, info etc…..

After doing the analysis if we find any unnecessary columns in the dataset we can drop those columns.

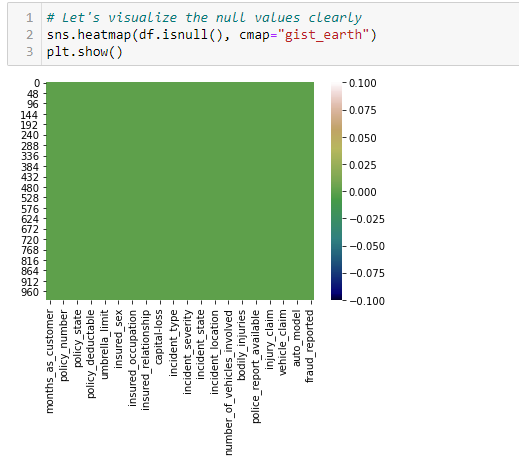








After doing basic analysis, we are checking for the null values. Will mention all the further analysis below

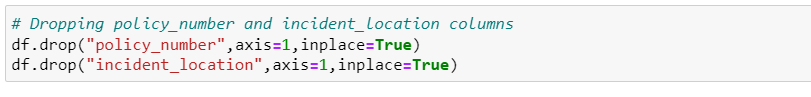


**Observations :**

* First we can see that we donot have any null values in the dataset.
* Second, the dataset contains 3 different types of data namely integer data type, float data type and object data type.
* Third, after analyzing it is seen that c39 column has only entries those are all NaN. Keeping all entries NaN is useless hence dropping that column



* Fourth, we can observe the columns policy\_number and incident\_location have 1000 unique values which means they have only one value count. So it not required for the prediction so we can drop it.



* Fifth, by looking at the value counts of each column we can realize that the columns umbrella\_limit, capital-gains and capital-loss contains more zero values around 79.8%, 50.8% and 47.5%. I am keeping the zero values in capital\_gains and capital\_loss columns as it is. Since the umbrella\_limit columns has more that 70% of zero values, let's drop that column.

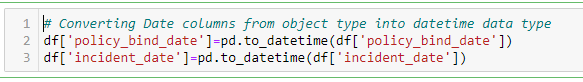


* Sixth, the column insured\_zip is the zip code given to each person. If we take a look at the value count and unique values of the column insured\_zip, it contains 995 unique values that mean the 5 entries are repeating. Since it is giving some information about the person, either we can drop this or we can convert its data type from integer to object for better processing.

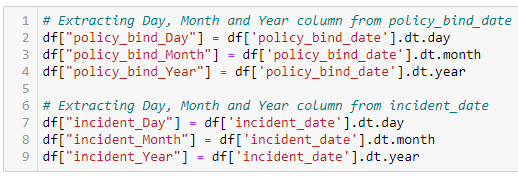


**Proceeding to Feature Extraction:**

The policy\_bind\_date and incident\_date have data type as object, which should be in datetime data type that means the python is not able to understand the type of this column and giving default data type. We will convert this object data type to datetime data type and we will extract the values from these columns.



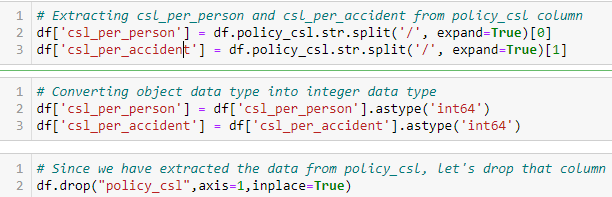
Now that we have converted object data type into datetime data type. Now let's extract Day, Month and Year from both the columns



After we have extracted Day, Month and Year columns, from both policy\_bind\_date and incident\_date columns. So we can drop these columns.



Again, from the features we can see that the policy\_csl column is showing as object data type but it contains numerical data, maybe it is because of the presence of "/" in that column. So first we will extract two columns csl\_per\_person and csl\_per\_accident from policy\_csl colums and then will convert their object data type into integer data type.



After extracting we have dropped the policy\_csl feature.

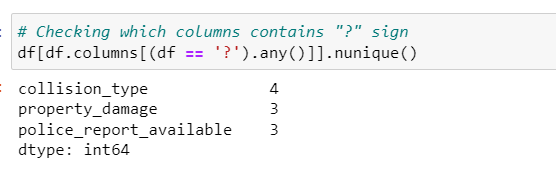
Also we have observed that the feature ‘incident-year’ has one unique value throughout the column also it is not important for our prediction so we can drop this column.



**Moving on to Imputation:**

Imputation is a technique to fill null values in the dataset using mean, median or mode. YES…. I know you might be thinking that we did not get any null values while checking for the null values, however from the value counts of the columns we have observed that some columns have "?" values, they are not NAN values but we need to fill them.

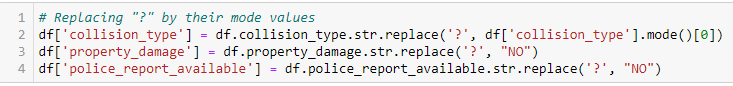
So, let’s begin……



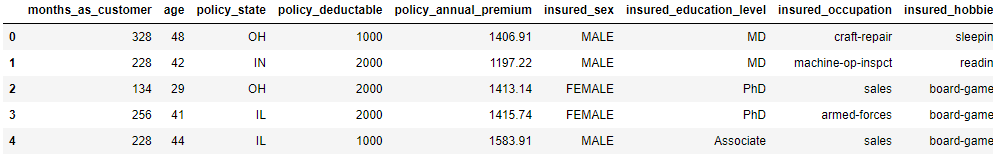
These are the columns which contains "?" sign. Since these column seems to be categorical so we will replace "?" values with most frequently occurring values of the respective columns that is their mode values.



The mode of property\_damage and police\_report\_available is "?", which means the data is almost covered by "?" sign. So we will fill them by the second highest count of the respective column.

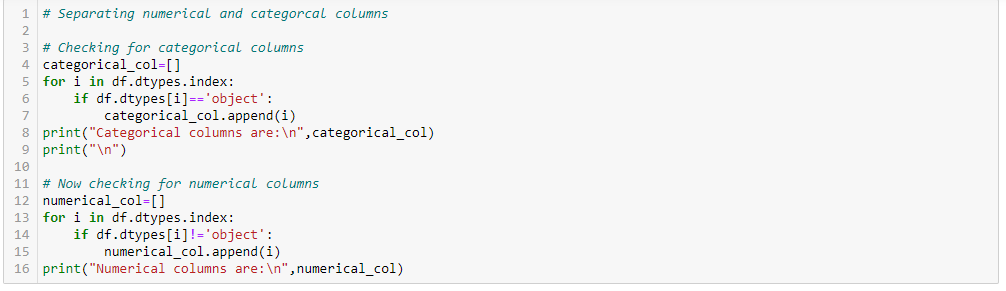


Now after all the data cleaning until now, the dataset looks like this…..

****

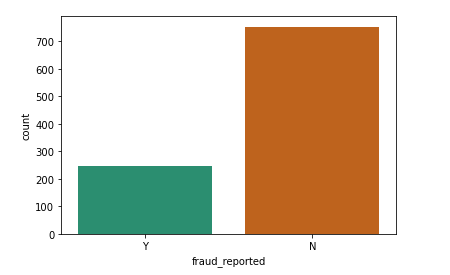
**Preparing for Visualization**

First we will look into the categorical and numerical columns so that we can visualize the features accordingly.

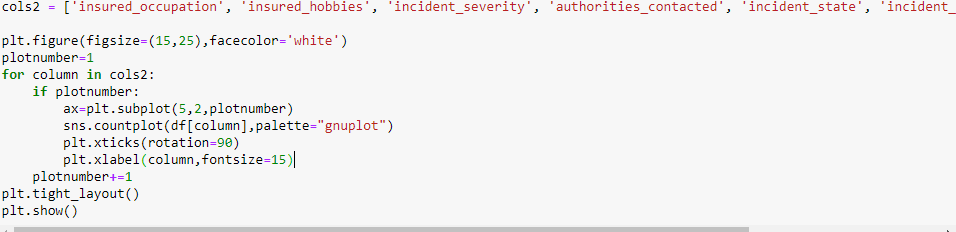


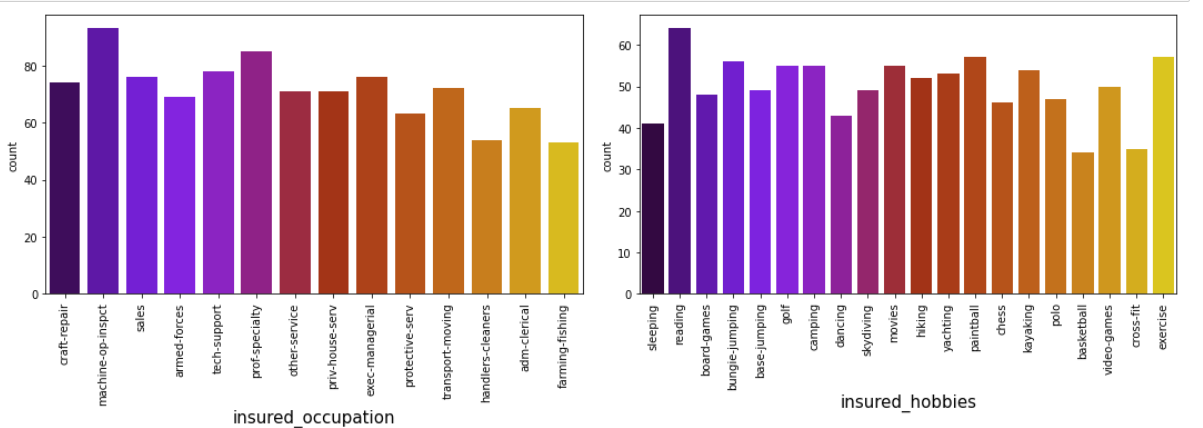
**Visualization**

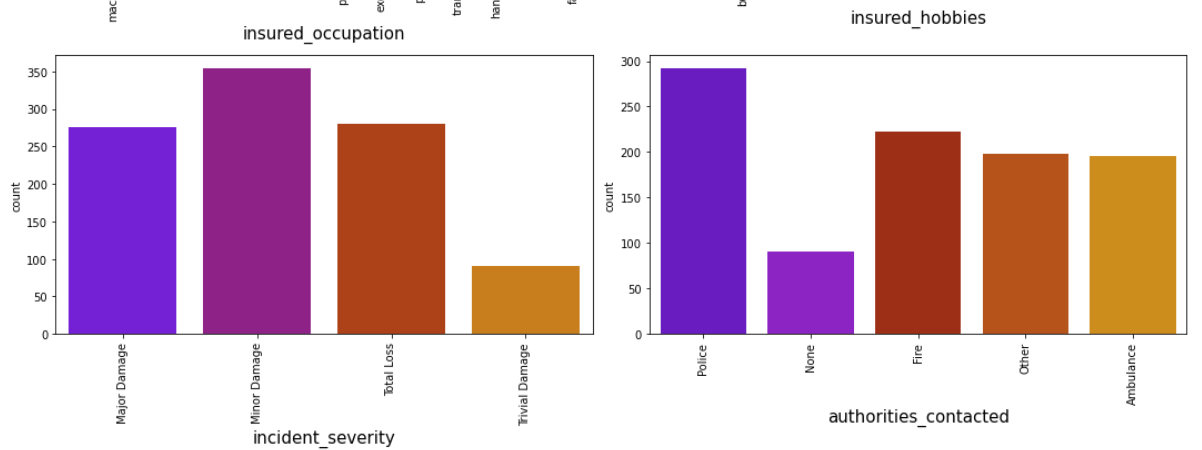


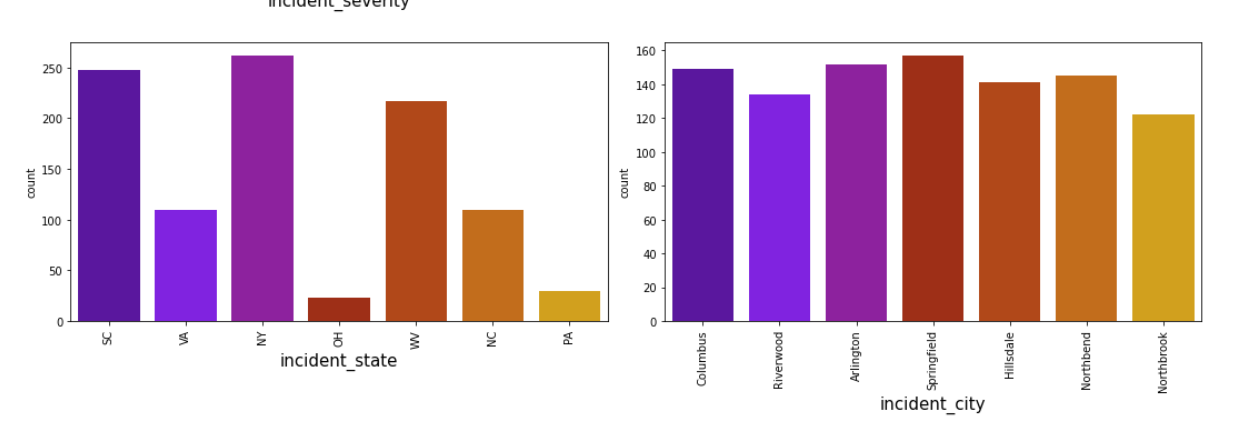


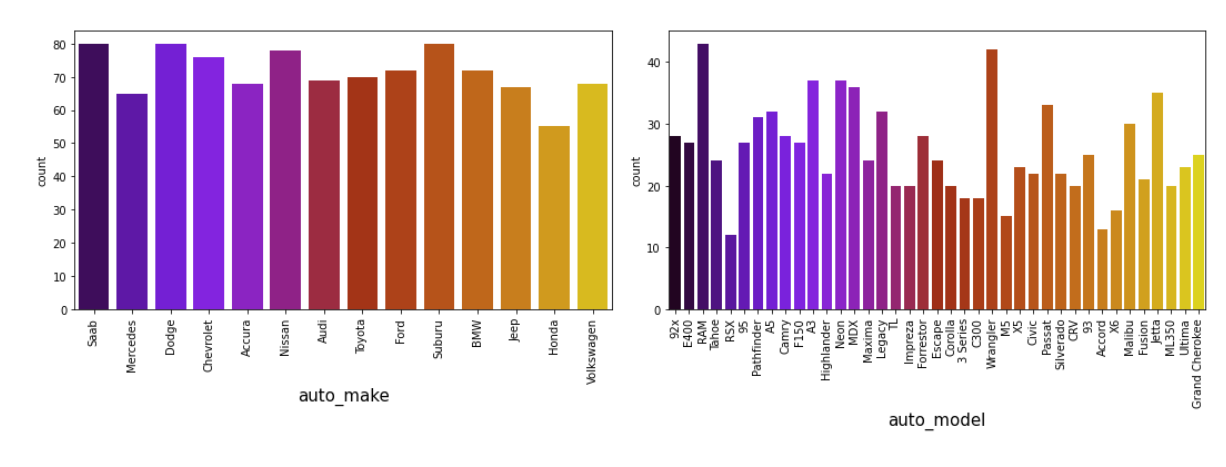
By looking into the plot we can observe that the count of "N" is high compared to "N". Which means here we can assume that "Y" stands for "Yes" that is the insurance is fraudulent and "N" stands for "No" means the insurance claim is not fraudulent. Here most of the insurance claims have not reported as fraudulent. Since it is our target column, it indicates the class imbalance issue. We will balance the data using oversampling method in later part.







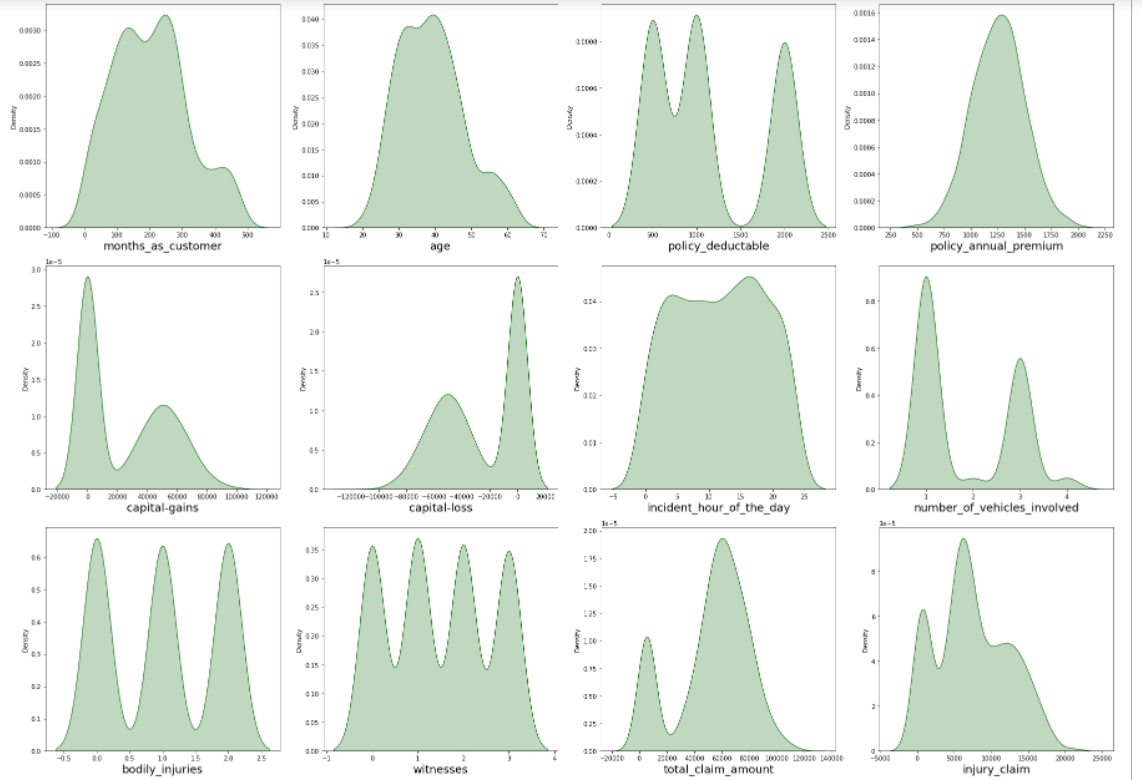


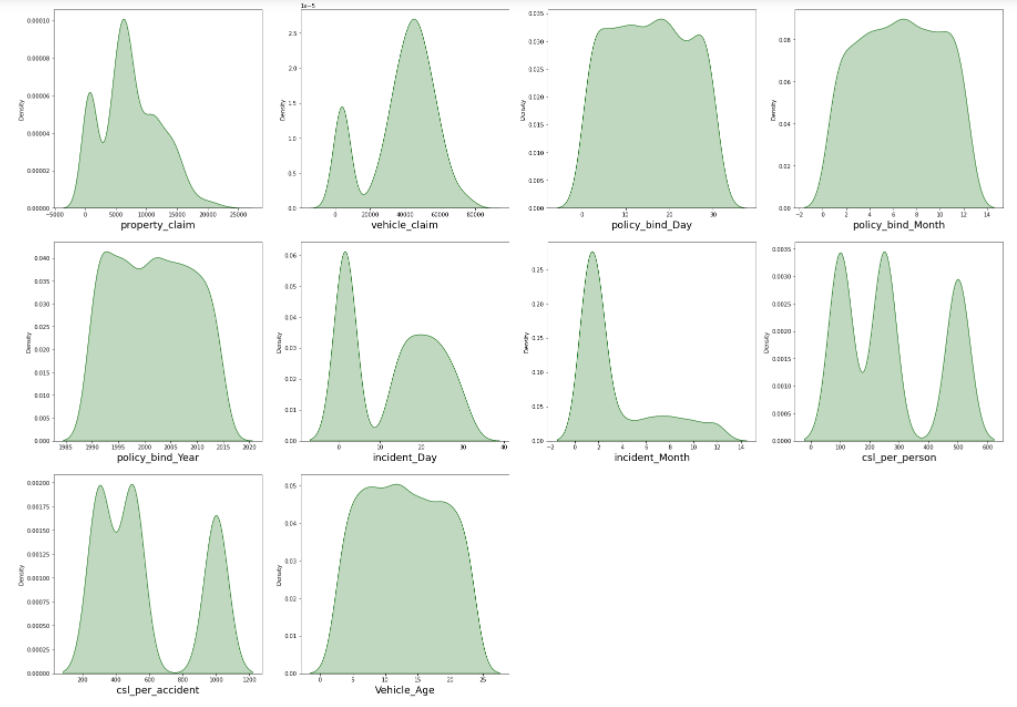


 By looking into the count plots we can observe the following things:

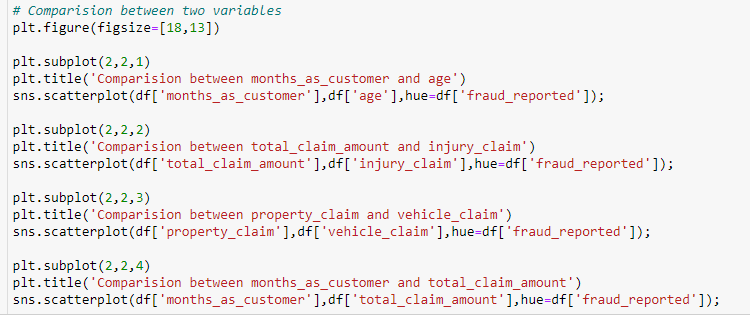
* In the insured occupation we can observe most of the data is covered by machine operation inspector followed by professional speciality.
* With respect to insured hobbies, we can notice reading covered the highest data followed by exercise. And other categories have the average counts.
* The incident severity count is high for Minor damages and trivial damage data has very less count compared to others.
* When the accidents occurs then most of the authorities contacts the police, here the category police covers highest data and Fire having the second highest count. But Ambulance and Others have almost same counts and the count is very less for none compared to all.
* With respect to the incident state, New York, South Carolina and West Virginia states have highest counts. In incident city, almost all the columns have equal counts.
* When we look at the vehicle manufactured companies, the categories Saab, Suburu, Dodge, Nissan and Volkswagen have highest counts.
* When we take a look at the vehicle models then RAM and Wrangler automobile models have highest counts and also RSX and Accord have very less count.

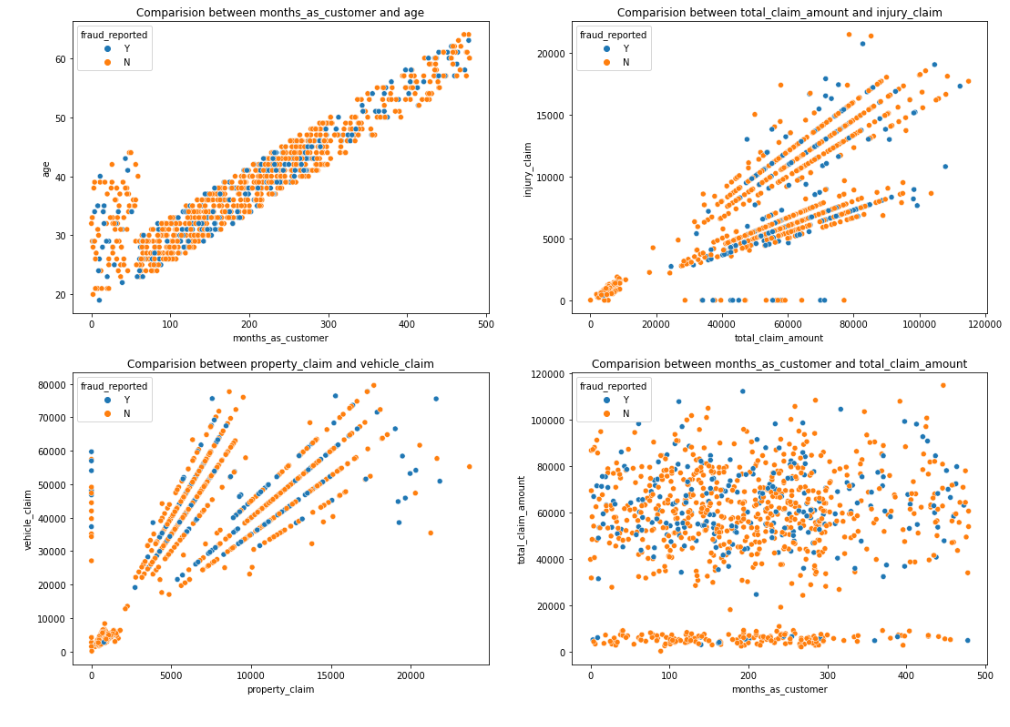




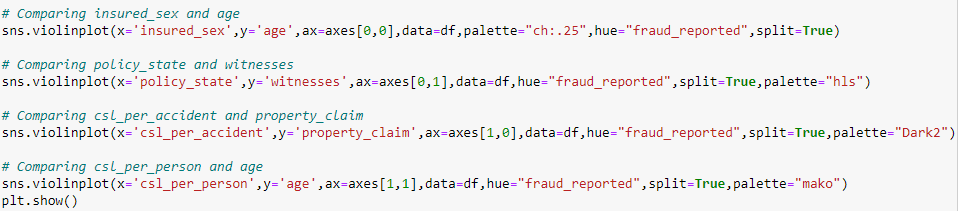


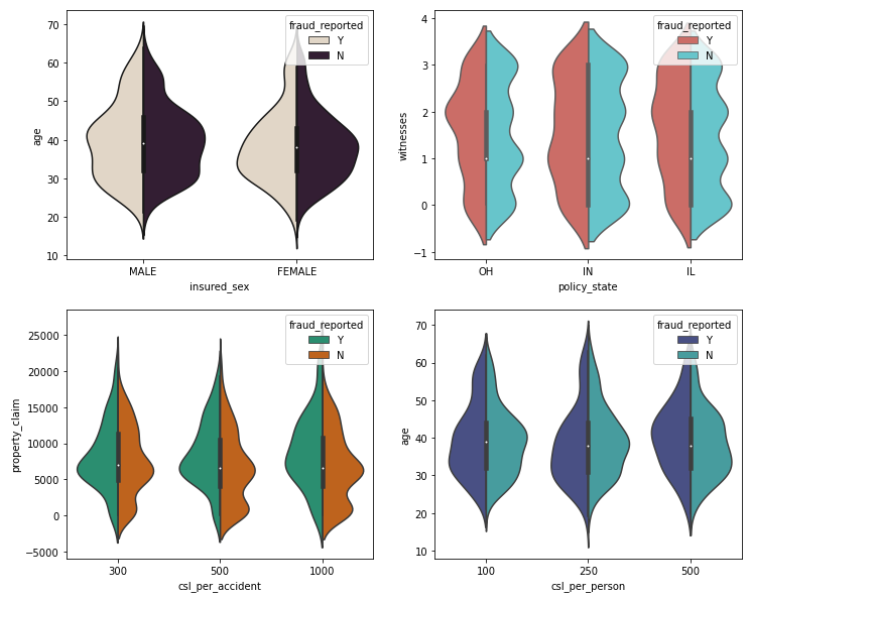
The data is normally distributed in most of the columns. Some of the columns like capital gains and incident months have mean value greater than the median, hence they are skewed to right. The data in the column capital loss is skewed to left since the median is greater than the mean. We will remove the skewness using appropriate methods in the later part.

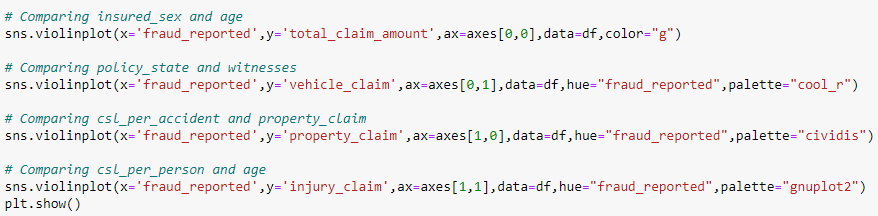


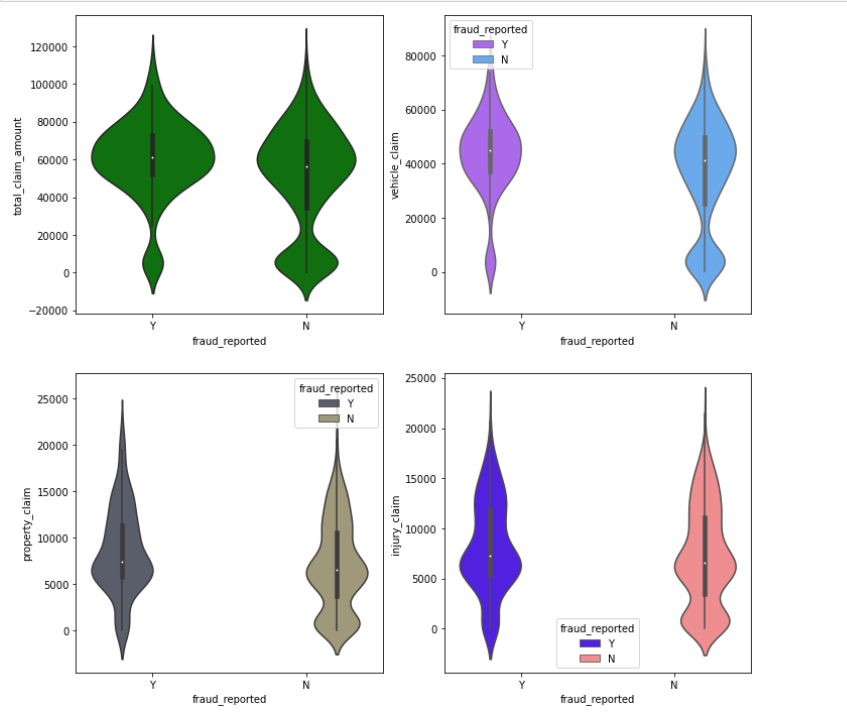


* There is a positive linear relation between age and month\_as\_customer column. As age increases the month\_as customers also increases, also the fraud reported is very less in htis case.
* In the second graph we cna observe the positive linear relation, as total cliam amount increases, injury claim is also increases.
* Third plot is also same as second one that is as the property claim increases, vehicle claim is also increases.
* In the fourth plot we can observe the data is scattered and there is no much relation between the features.



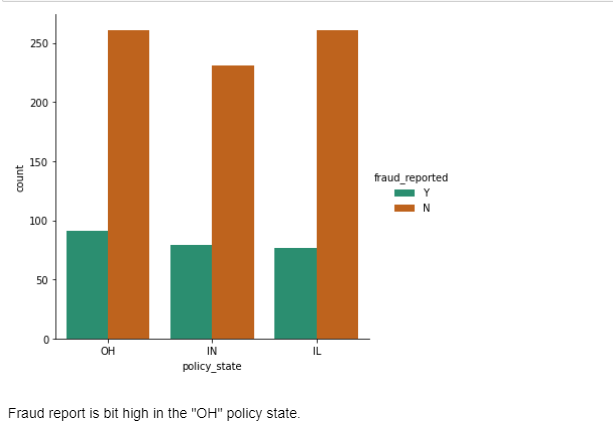




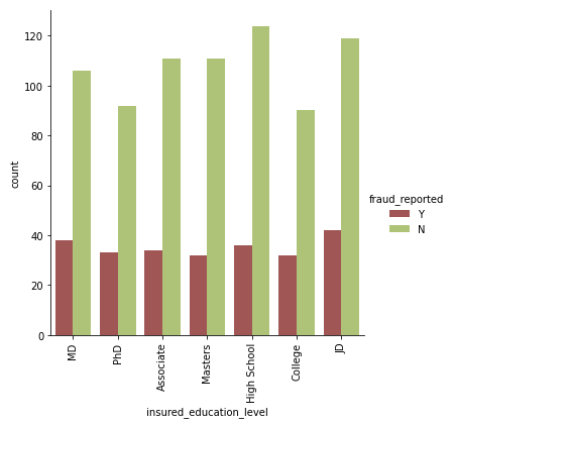


Visualization is a technique where by comparison and plotting the data becomes self explanatory, which we have seen until now. Moving ahead with some more visualization plots before we can proceed to model building.

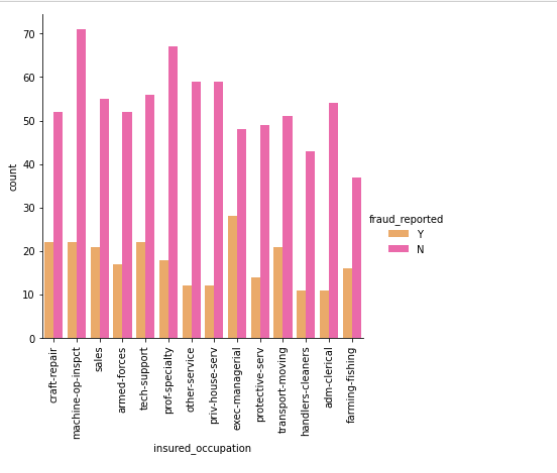


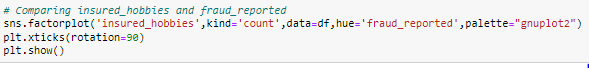


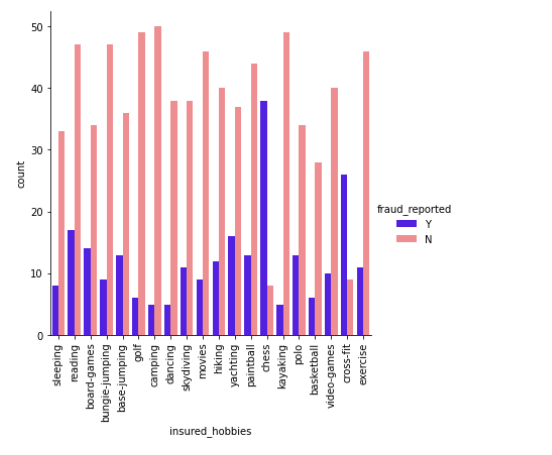


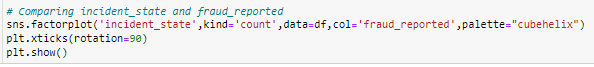


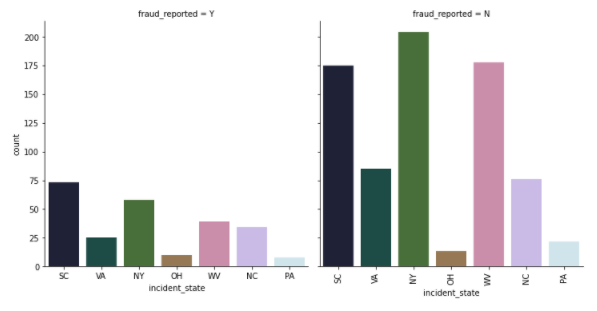




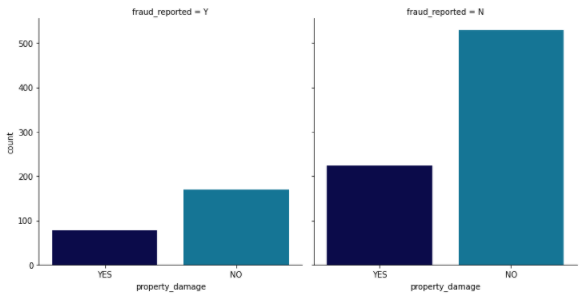






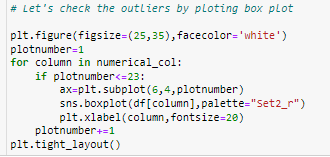


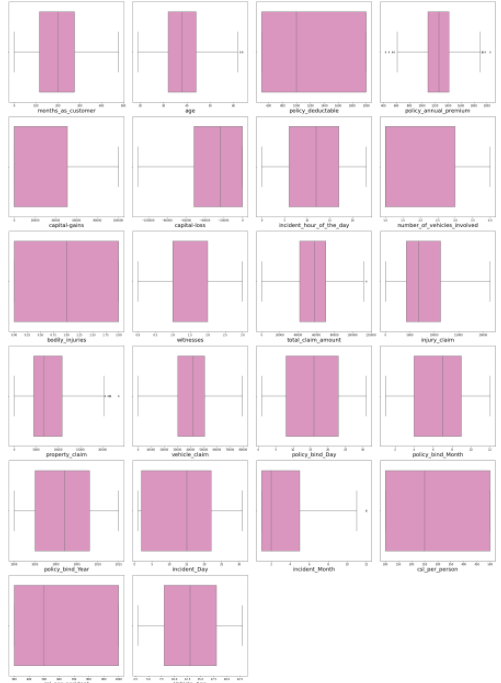




Now we have done with the visualization in order to analyze and understand the data. So in this EDA part, we have looked into various aspect of the dataset, like looking for the null values and imputing, extracting date time, observing the value counts and doing the feature extraction etc. Now we will be performing another analysis by identifying the outliers and removing them. Along with it we will also look for the skewness of the dataset and remove the skewness.

**Identifying the Outliers and Skewness**

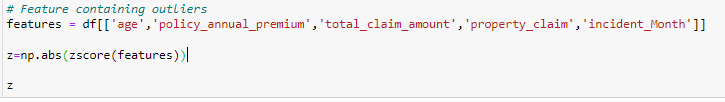
****

****

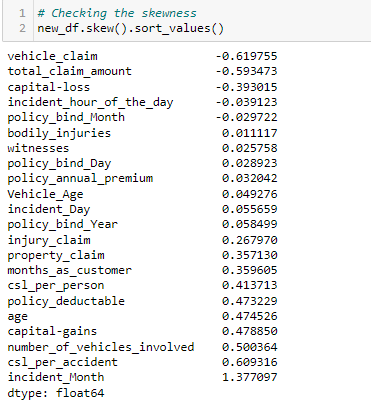
As we can see, I have used box plot to identify the outliers and we can find the outliers in the following columns:

* Age
* policy\_annual\_premium
* total\_claim\_amount
* property\_claim
* incident\_month

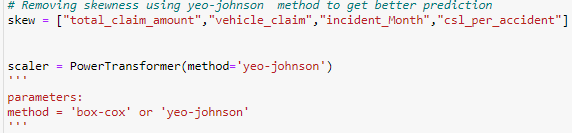
These are the numerical columns which contains outliers hence removing the outliers in these columns using Zscore method.



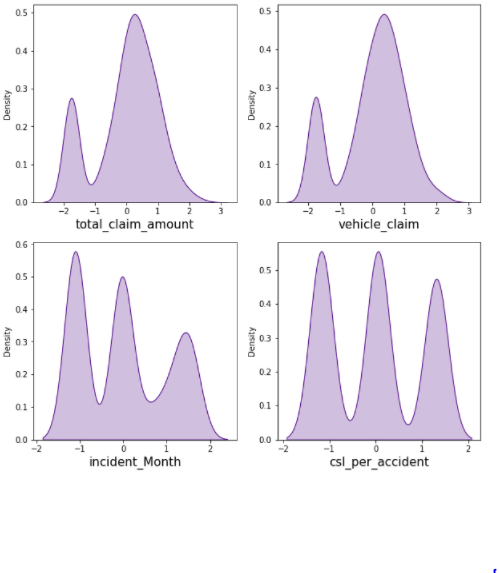
Now that we have removed the outliers, I will proceed to look into the skewness of the data and then remove it.



As we can see that skewness is present in the dataset, hence I am using the yeo-johnson method to remove the skewness.

****

Now we have removed the skewness and the data looks normally distributed.

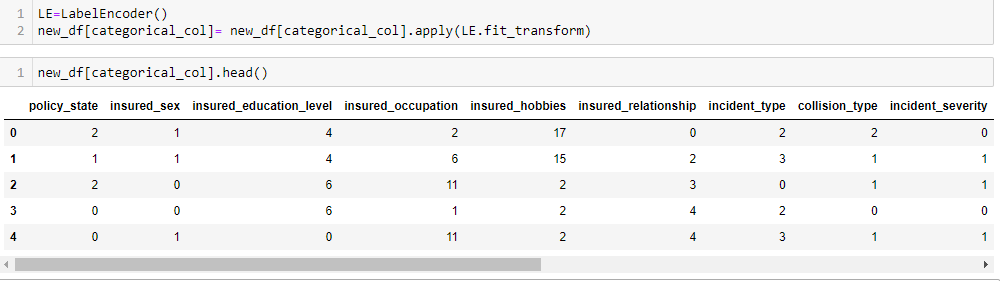
****

Now we have completed our analysis of the dataset and also cleaned the dataset so that we can build a good model….

However we are not yet ready. We have seen above that there are some problems that still exist in the dataset. We have seen that the dataset has both numerical and categorical data. The model only understand numerical data, hence we will encode the data. Also we have seen that there can be some multi-colinearity, that we will see through a heatmap and also further remove it. Again we have also seen that the target variable is imbalance, hence will fix it by oversampling. And finally we will scale the data so that it is ready to be trained and tested.

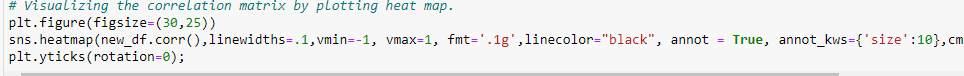
Let’s begin.

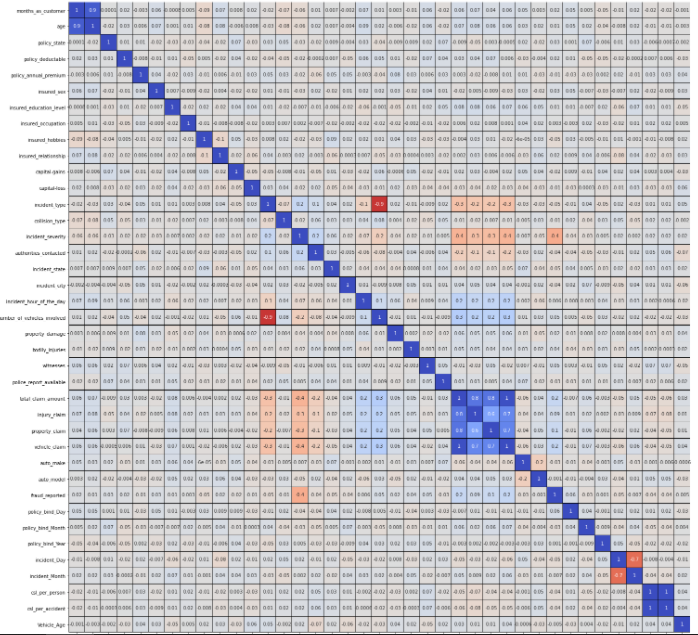
**Encoding the data**

****

Now we have encoded the dataset using label encoder and the dataset looks like this.

Moving forward, to check the co relation between the feature and target and also the relation between the features using the heatmap.





This heatmap shows the correlation matrix by visualizing the data. We can observe the relation between one feature to another.

There is very less correlation between the target and the label. We can observe the most of the columns are highly correlated with each other which lead to the multicollinearity problem. We will check the VIF value to overcome with this multicollinearity problem.

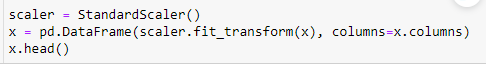
**Preprocessing Pipelines**

## Separating the features and label variables into x and y

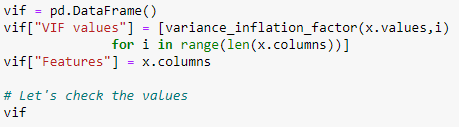
****

**Scaling the DataSet**

## Feature Scaling using Standard Scalarization



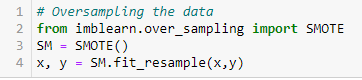
**Checking Multi-colinearity using VIF**

****

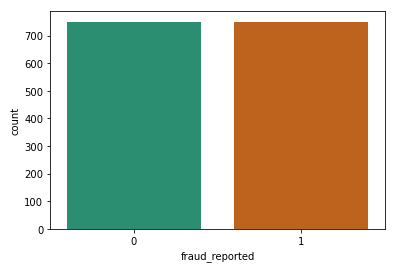
It is observed that some columns have VIF above 10 that mean they are causing multicollinearity problem. Let's drop the feature having high VIF value amongst all the columns.

I have dropped the total\_claim\_amount and csl\_per\_accident features with colinearity more than 10, and now we have removed the problem.

We had earlier identified another problem of imbalance data in the target variable, lets treat it.



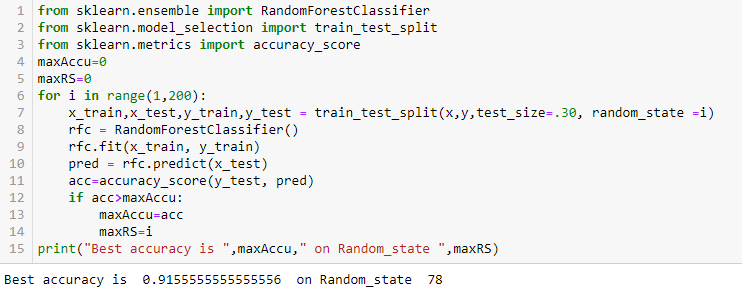
As we have treated the oversampling issue using SMOTE, now the data looks good.



Finally we have got into the position where we will start building the model.

At first let’s find the best random state in which we can build the model*.*

*(Random state ensures that the splits that you generate are reproducible. Scikit-learn use random permutations to generate the splits. The random state that you provide is used as a seed to the random number generator. This ensures that the random numbers are generated in the same order.)*

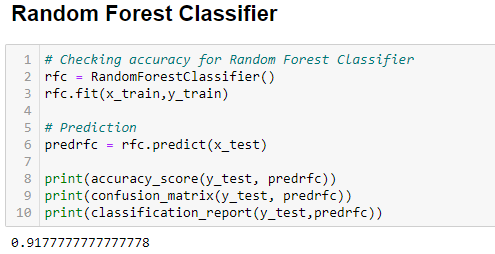
****

Here we have used the RandomForestClassifier to find the best random state, as we have got an accuracy score of 91% (pretty good), at the random state of 78. Let’s use this random state to build our models.

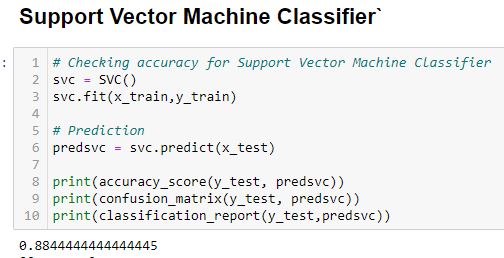
Before doing that, let us split the dataset into train and test using train\_test\_split.



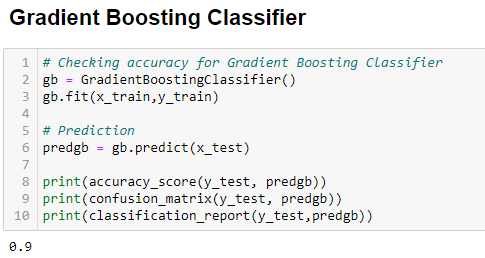
**Model Building**



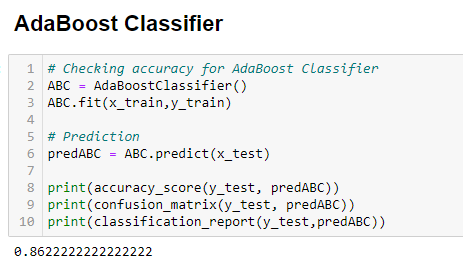
The first model was built using RandomForestClassifier, which gave an accuracy score of 91%, however we are hungry data scientist and will not be satisfied with only one model. We will try various models and see what accuracy score we get.



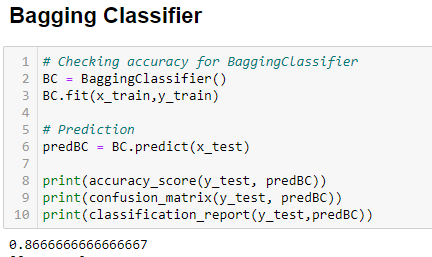
With SupportVectorClassifier we got an accuracy score of 88%.



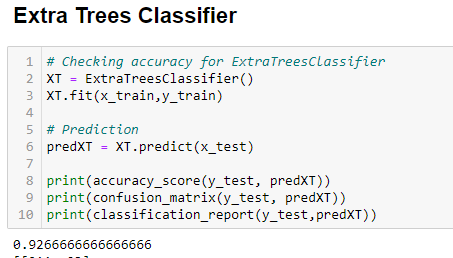
With GradientBostingClassifier we got an accuracy score of 90%.



With AdaBoostClassifier we got an accuracy score of 86%.

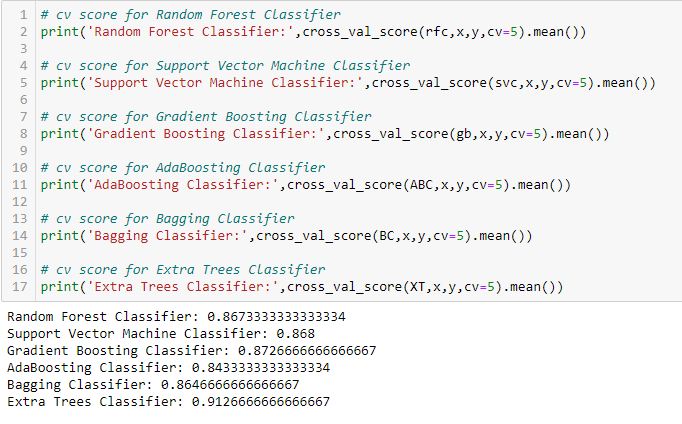


With BaggingClassifier we got an accuracy score of 88%.



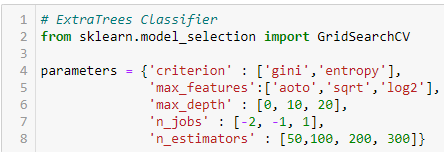
With this model, ExtraTreesClassifier we have got accuracy score of 92%, which is better than RandomForestClassifier.

Before we can announce the best model, we always have to make sure that the model is not over fitting; hence we will perform cross validation of all the models built.



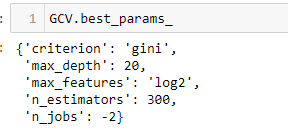
After the cross validation we can clearly see that ExtraTreesClassification is the best fit model.

Now, that we have found the best fit model, lets perform some HyperParameterTuning to improve the performance of the model.

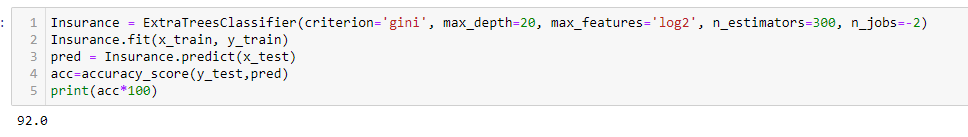




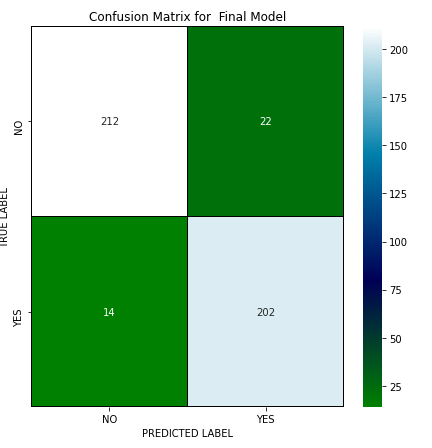




Here we have got the best parameters, and we will build our final model using these parameters.

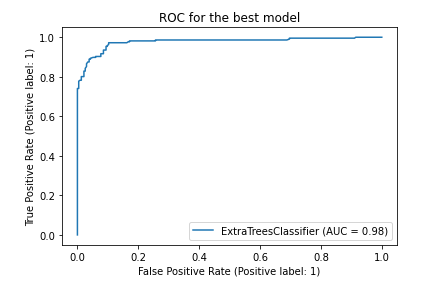


We have built our final model and we can see that the accuracy scores has increased by 1% from the cross validation score.



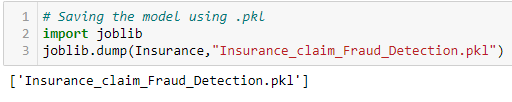
This is the confusion matrix for the model.

Plotting and AUCROC curve for the final model.

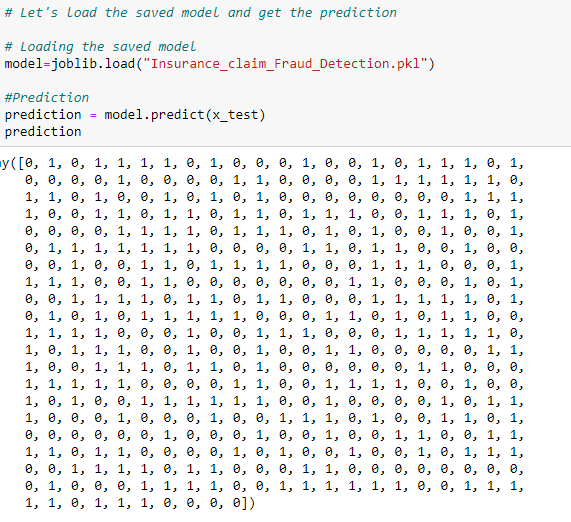


So here we can see that the area under curve is quite good for this model.

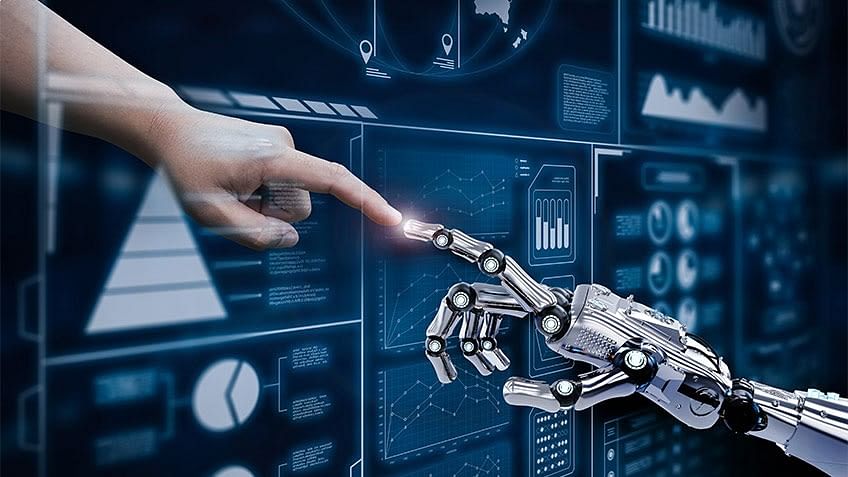
**Saving the model**

****

**Predicting the model**

****

**Concluding Remarks**

****

In the beginning of the blog we have discussed about the lifecycle of a Machine Learning Model, you can see how we have touched based on each point and finally reached up to the model building and made the model ready for deployment.

This industry area needs a good vision on data, and in every model building problem Data Analysis and Feature Engineering is the most crucial part.

You can see how we have handled numerical and categorical data and also how we build different machine learning models on the same dataset.

Using hyper parameter tuning we can improve our model accuracy, for instance in this model the accuracy remained same.

Using this machine Learning Model we people can easily predict the insurance claim is fraudulent or not and we could reject those application which will be considered as fraud claims.

