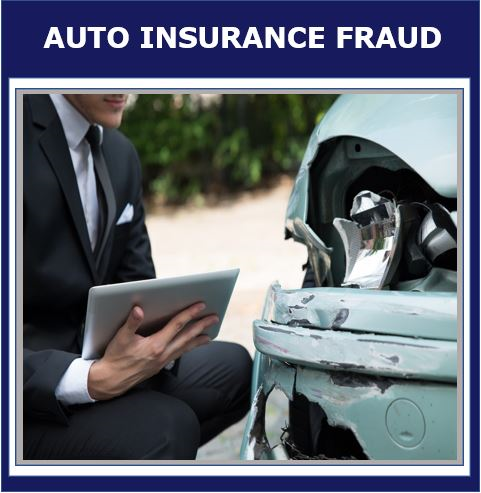
Blog/Article

**Insurance Claim Fraud Detection**



**Created By – Akshay Shingavi**

# **Problem Definition.**

Insurance claim fraud is one of the greatest avoidable losses that impact insurers globally. With car insurance and workers' compensation making up the largest percentage of fraudulent claims that have an annual impact on the insurance industry, the Property, and Casualties (P&C) segment is responsible for the majority of fraudulent insurance claims.

## **What is Insurance Fraud?**

Insurance fraud happens when an insurance company, agent, adjuster, or client knowingly lies to obtain an unfair benefit. It may take place when buying, using, reselling, or underwriting insurance. There are many subcategories of insurance fraud, including fraud committed by customers and insurance companies. Fraud increases expenses for insurance firms while also having a negative financial impact on individuals and businesses.

## **Types of insurance fraud**

* **Hard Fraud** – It refers to the purposeful fabrication of damage or injury to get funds from an insurance provider.
* **Soft Fraud** – It entails inflating an actual accident's cost to maximize the claim payout.

## **Benefits of Insurance Fraud Detection Analytics**

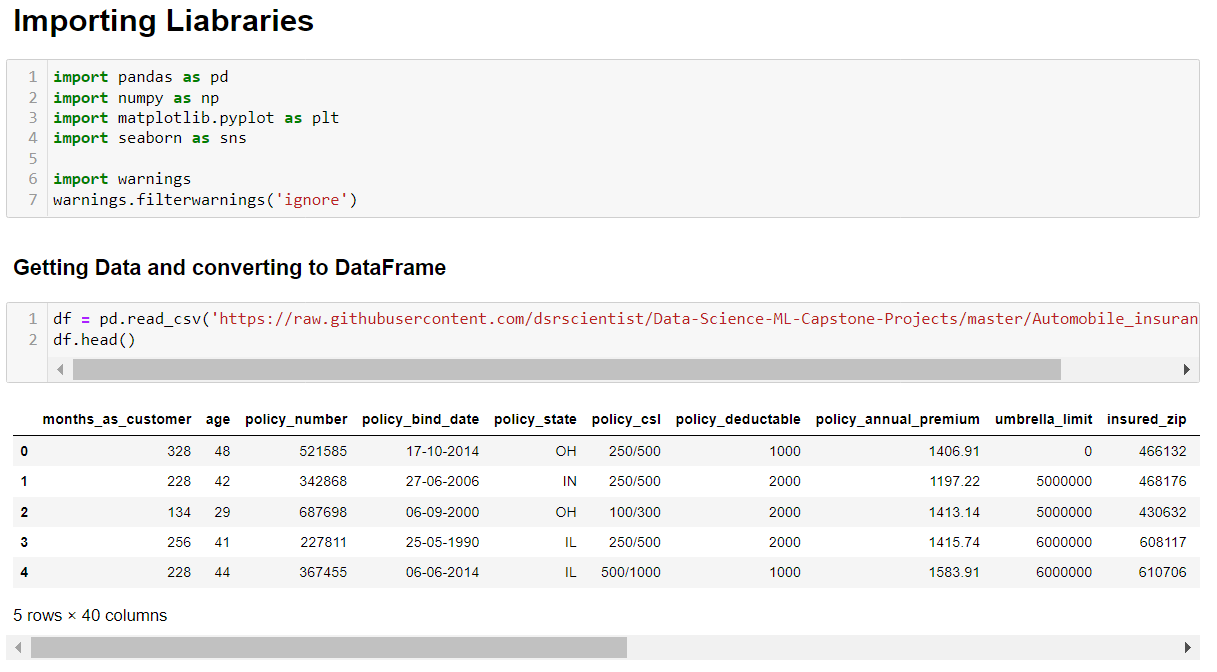
* **Risk evaluation –** A model that analyses insurance fraud is better able to identify risk than any individual. While it would take a person days of manual labor to find the same patterns that AI and predictive modeling systems do, they can examine enormous volumes of data in milliseconds.
* **Heightened fraud detection** - Insurance fraud analytics is more proficient in spotting abnormalities and warning signs that point to probable fraud schemes, much like risk assessment. These indicators support analytics teams in developing top-notch fraud team referrals. A fraud risk assessment should take these high-risk sectors into account, which may also be determined by these algorithms.
* **Fewer human interventions** - Insurers help reduce the number of manual interventions in the claims management process by using technology and data analytics to their fullest potential. This shortens turnaround times and frees up insurance agents, enabling them to concentrate on jobs that are more useful and have a greater effect.

# **Data Analysis.**

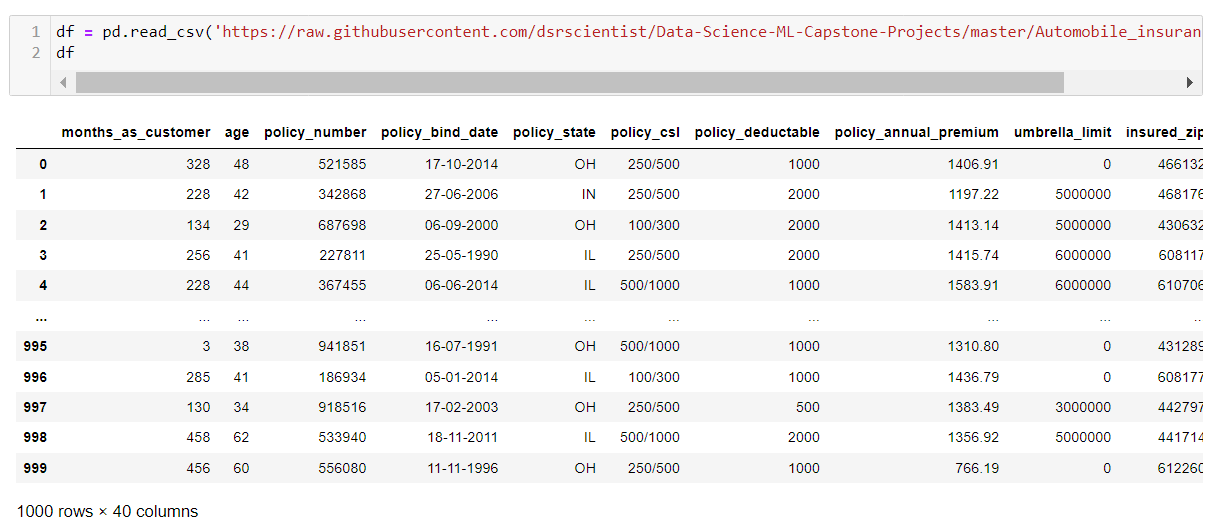
The process of examining, purifying, manipulating, and modelling data to find relevant information, support inferences, and help decision-making. Data analysis is utilized in a range of business, scientific, and social science sectors. It has many dimensions and methodologies, encapsulating several techniques under many titles. Data analysis contributes to more scientific decision-making and more efficient operations of firms in today's business environment.

## **Let’s begin with the process**

1. We started our process by importing some basic required libraries



1. Converting CSV file to Dataframe

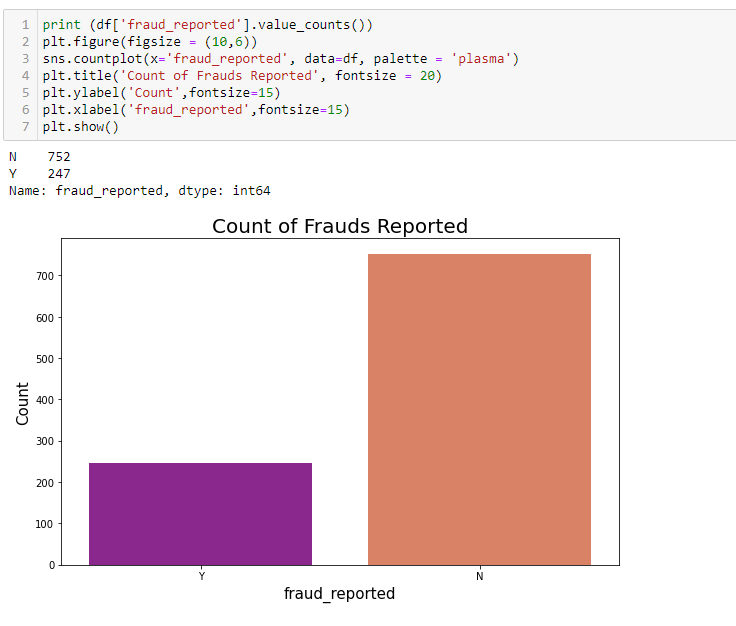


1. After Checking the data, we have identified the below errors which we need to correct before proceeding.
   * There is a question mark in several rows; we will substitute "Some other" for it.
   * As can be seen, there is one row with a negative value. This must be an error or is uncertain, thus the row is dropped.
   * Incident year column will be removed because all of the incidents in the incident data are from 2015 and it won't make much of a difference to omit it.
   * The incident Date is also combined and needs to be separated into 3 different columns.
   * Policy binds date must be divided into three more columns.
   * Additional columns policy number, insured zip, and incident location are optional and can be omitted.
   * We have done basic corrections of data such as filling null values, separating columns, etc, we can now proceed with further action.
2. After taking corrective actions we will being with our data visualization.

## **Data Visualization**

The graphic display of information and data is known as data visualization. Data visualization tools offer an easy approach to observing and analyzing trends, outliers, and patterns in data by utilizing visual components like charts, graphs, and maps. Additionally, it offers a great tool for staff members or business owners to deliver data to non-technical audiences.

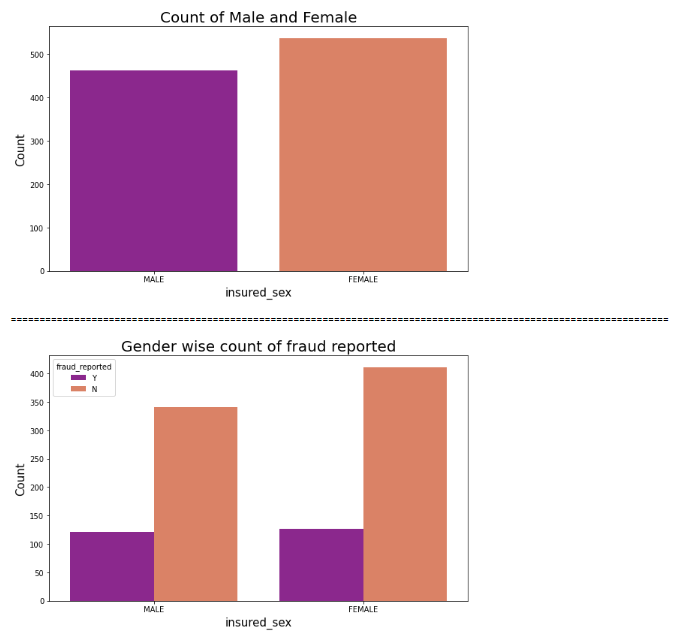
1. **Checking counts of Frauds Reported**

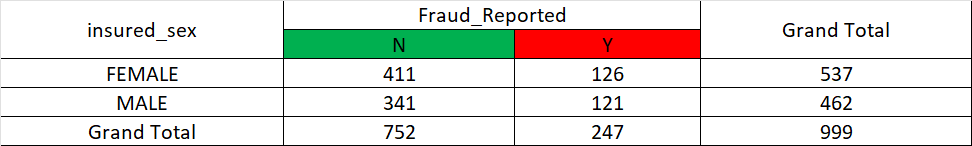
****

**Observation - Out of 999 entries, in 247 entries fraud was not reported, and the remaining 752 were not reported.**

1. **Checking the count of Males and Females and their Gender Wise Fraud Count**

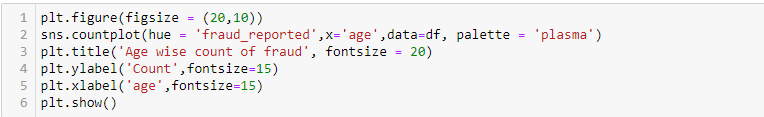




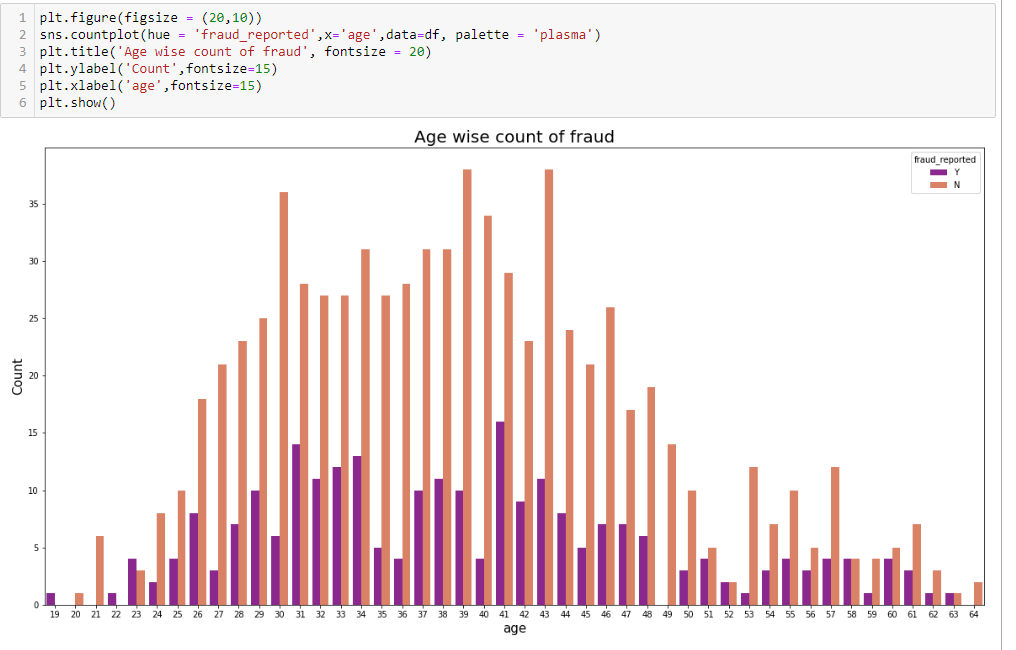


**Observation - Out of 999 people 537 are Female and the remaining 462 are males.**

**3. Plotting Age wise count of fraud**

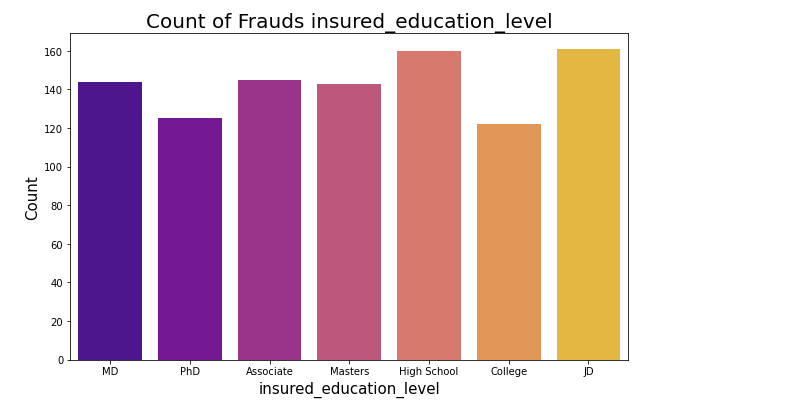
****

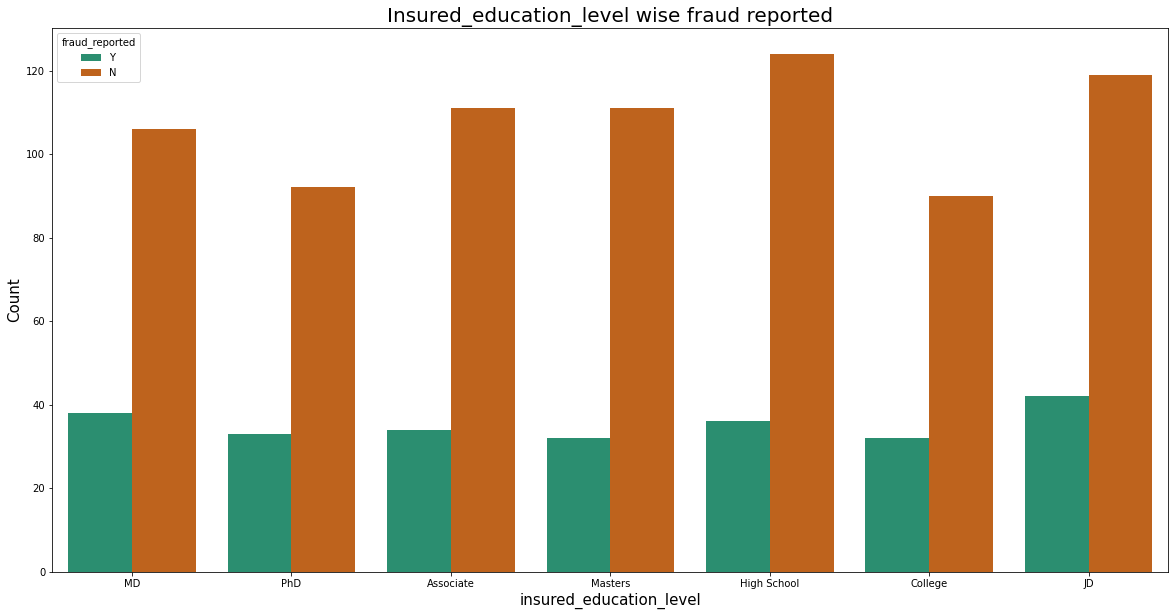
**Observation - Most frauds are of people with 41 years of age followed by 31 years, age 20, and 64 where no frauds were reported.**

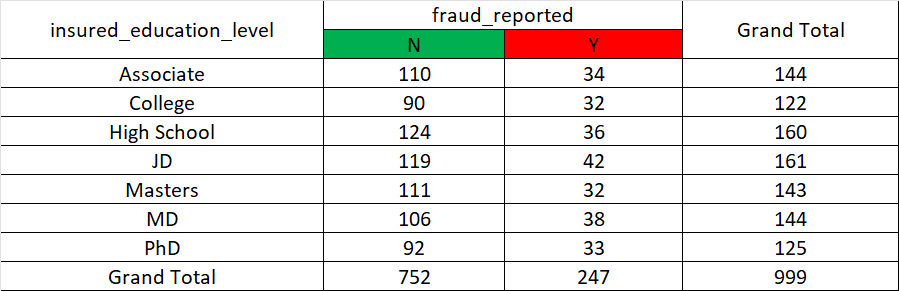
****

**4. Check the count of education level and fraud**

****

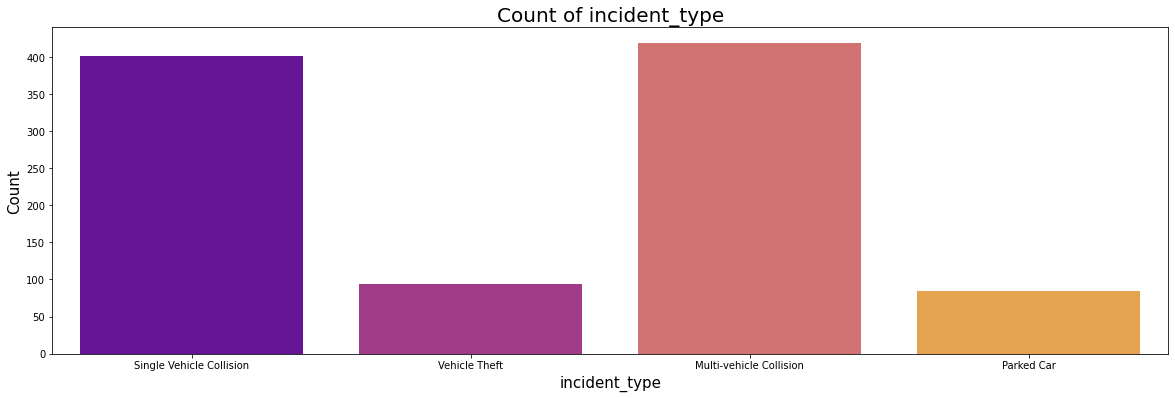
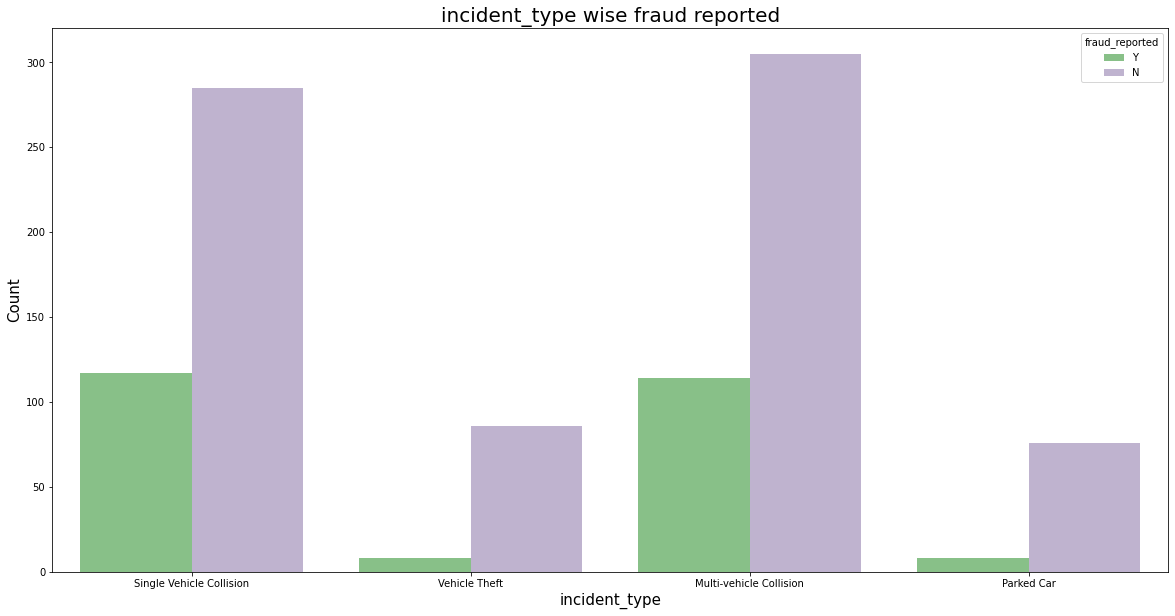
****

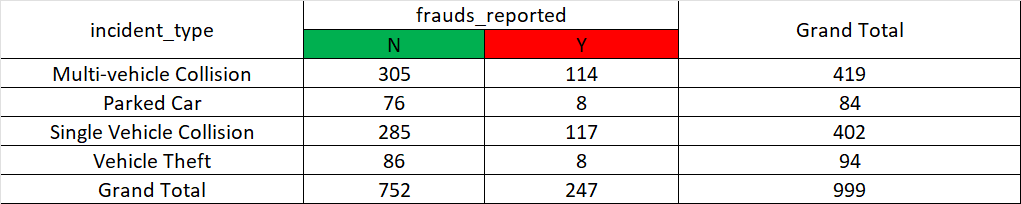
****

****

1. **Checking count of incident type and incident type-wise frauds reported.**



****

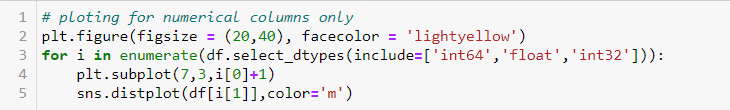


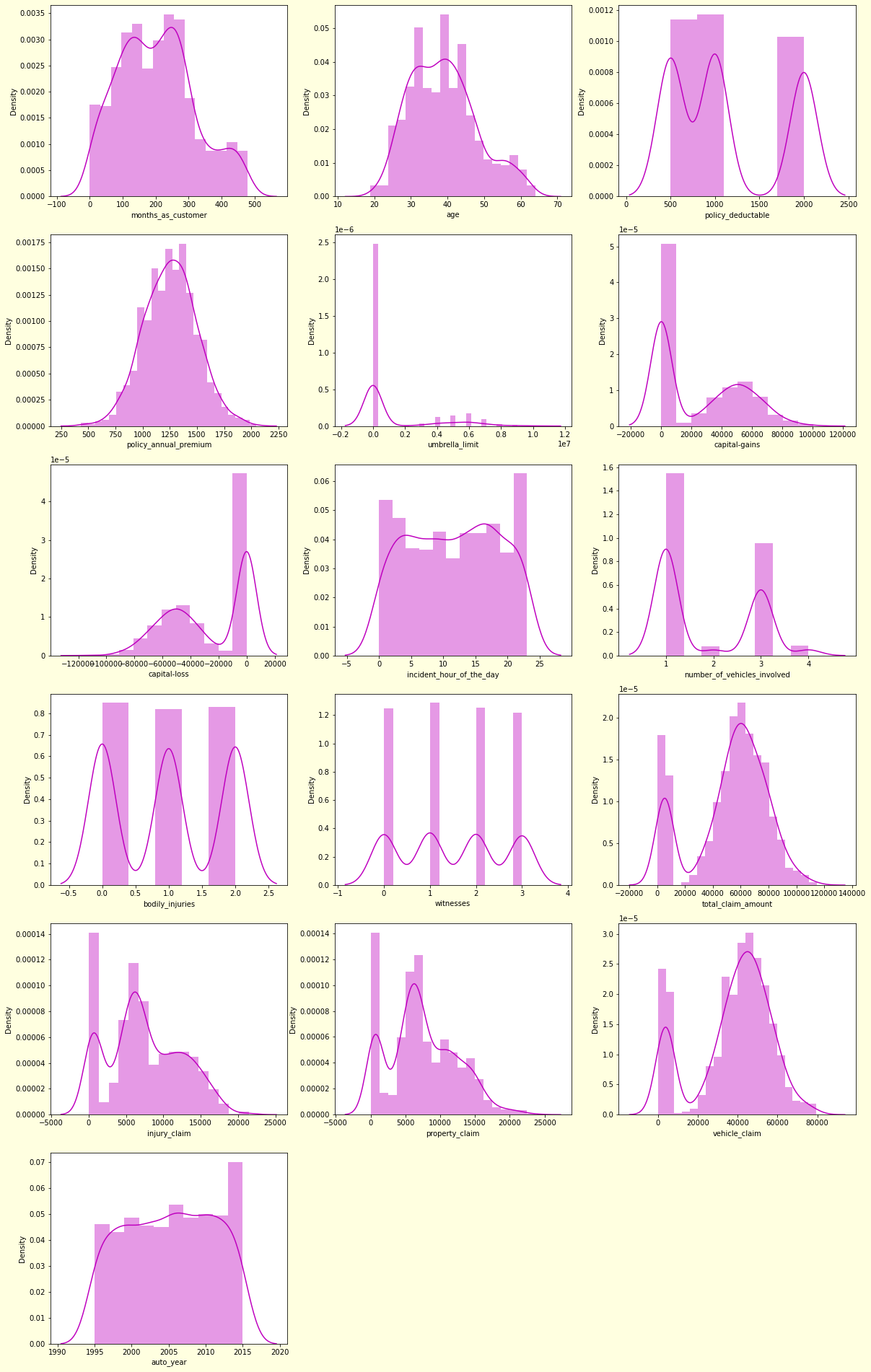
# **3. EDA Concluding Remark.**

Exploratory Data Analysis (EDA) is a method or philosophy for data analysis that makes use of a range of graphical and quantitative approaches to comprehend data better. EDA visuals make it simple to lose one's bearings and the reason for using EDA. EDA seeks to simplify downstream analysis. The steps in data science for EDA are as follows: assemble data, Data preparation, and loading, exploratory data analysis, model construction, model testing, and data visualization and presentation.

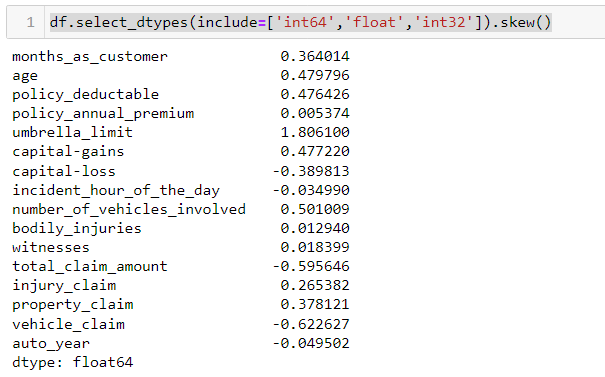
I examined the dataset's null values, but I couldn't find any missing data. I eliminated a few unnecessary columns to get around the multicollinearity issue. To achieve better outcomes without any obstacles, I have extracted several new features from the current features. Then removed the old columns because, if I left them in place, they would function as duplicates and cause a multicollinearity issue. inserted their corresponding mode values instead of the corrupted entries "?" in the columns. When it comes to visualization, we have identified the times and locations when the number of fraud complaints is highest. I have utilized bar plots, box plots, count plots, and distribution plots to gain a better understanding of the characteristics.

## **Checking and Visualising Skewness**

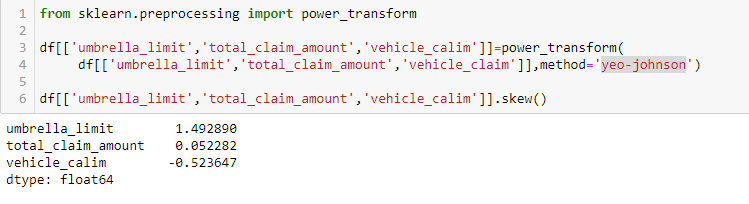




**There is some Skewness in 3 columns (umbrella\_limit, total\_claim\_amount, vehicle\_calim)**



**We have removed skewness using the yeo-Johnson method.**

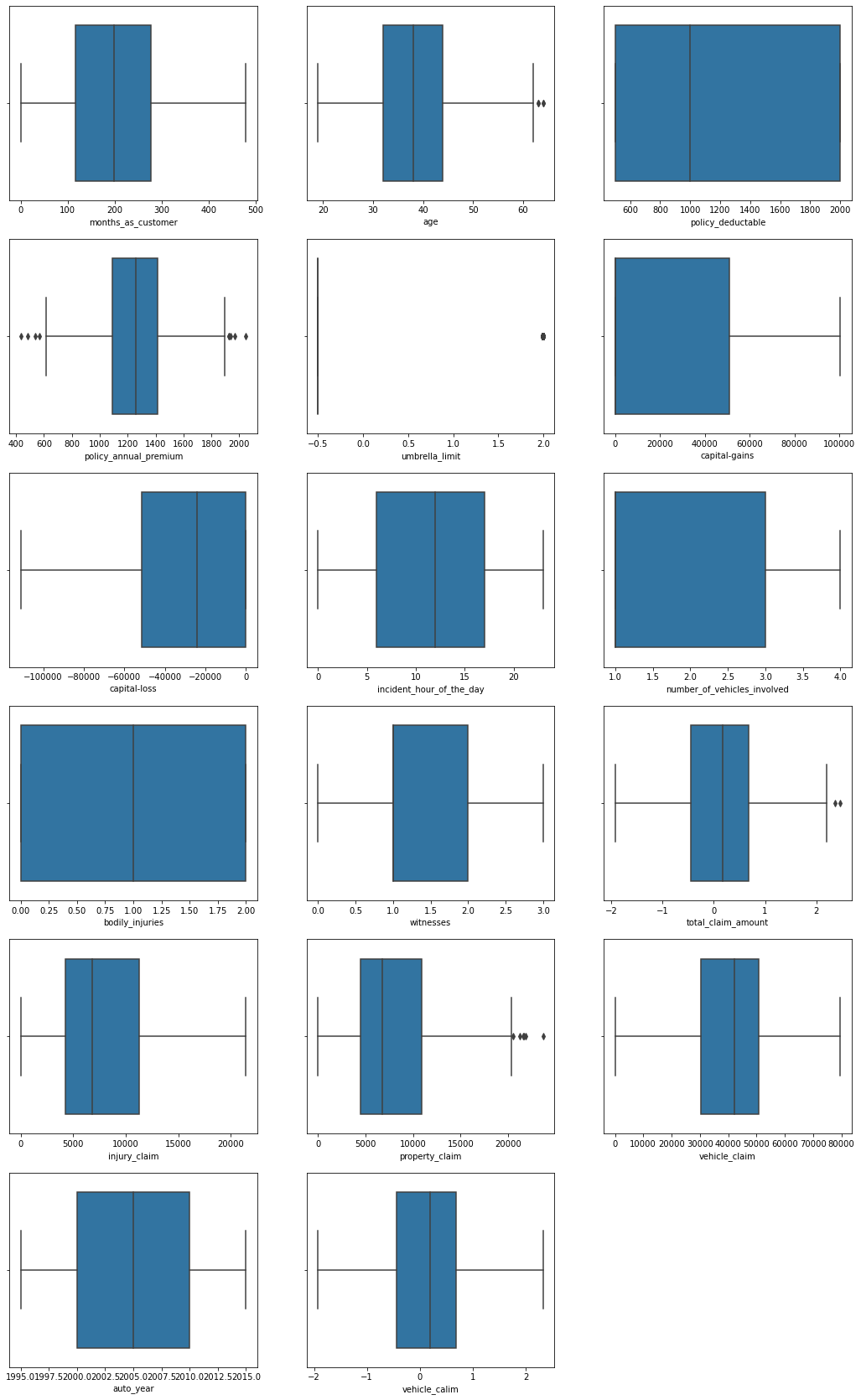


## **Handling Outliers**

**We will further check if our data has any outliers, and we will use a box bot to view the same.**

****

**There are a few columns where outliers are present, we need to remove these.**

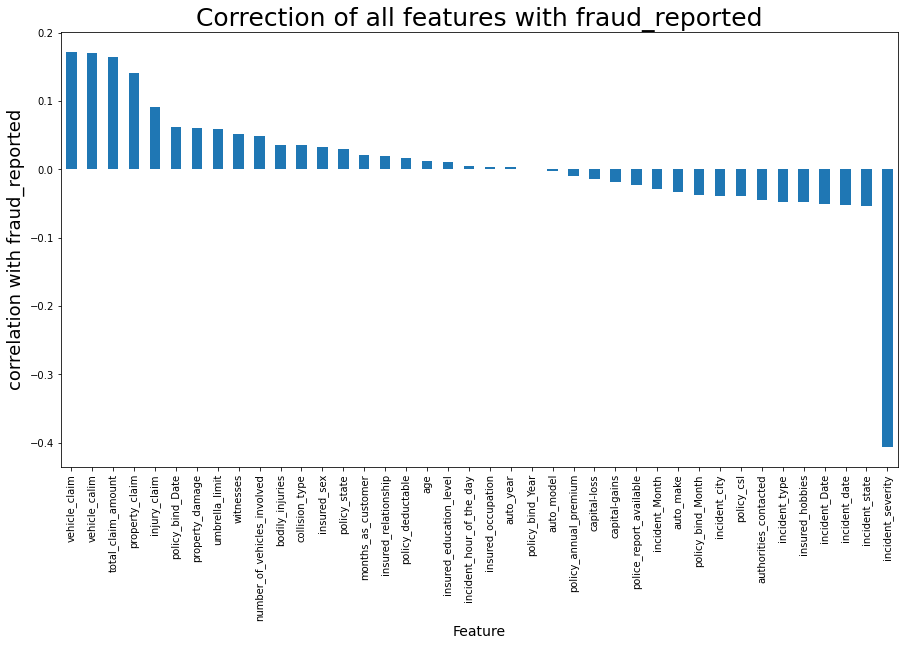
****

## **Removing Outliers – Z Score**

**We will remove outliers in the dataset using the Z- Score Method. There is little data loss using the Z score method, if we would have used the IQR method then there would be much data loss as compared to this method.**



**Correlation of Features with Target Column**

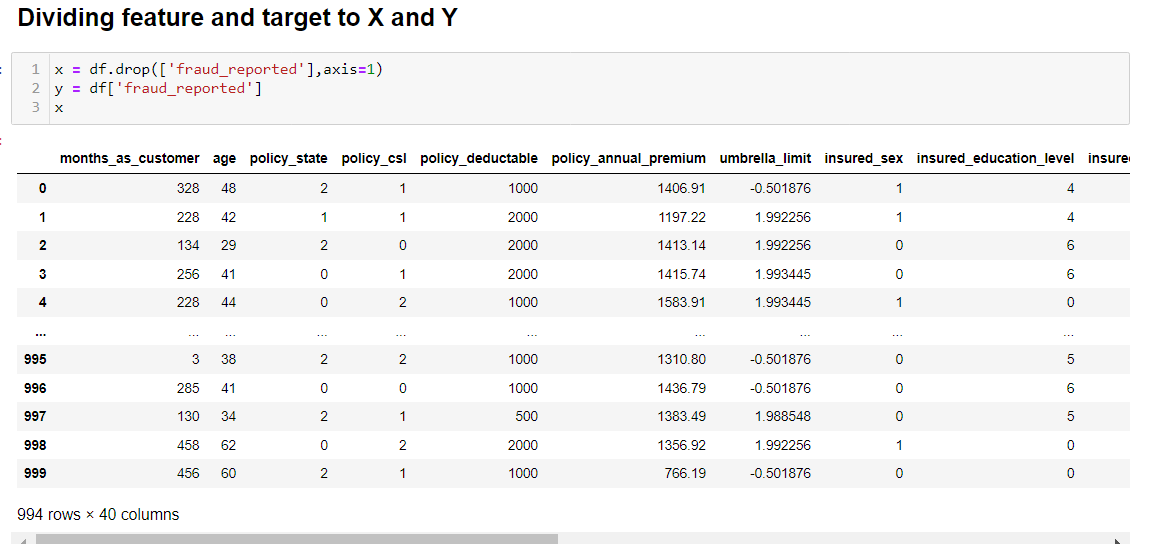


# **4. Pre-Processing Pipeline.**

Important Data Pre-processing Techniques are Data Cleaning, Dimensionality Reduction, Feature Engineering, Sampling Data, Data Transformation, and Imbalanced Data. We will first Divide features and columns, to begin with, data pre-processing.

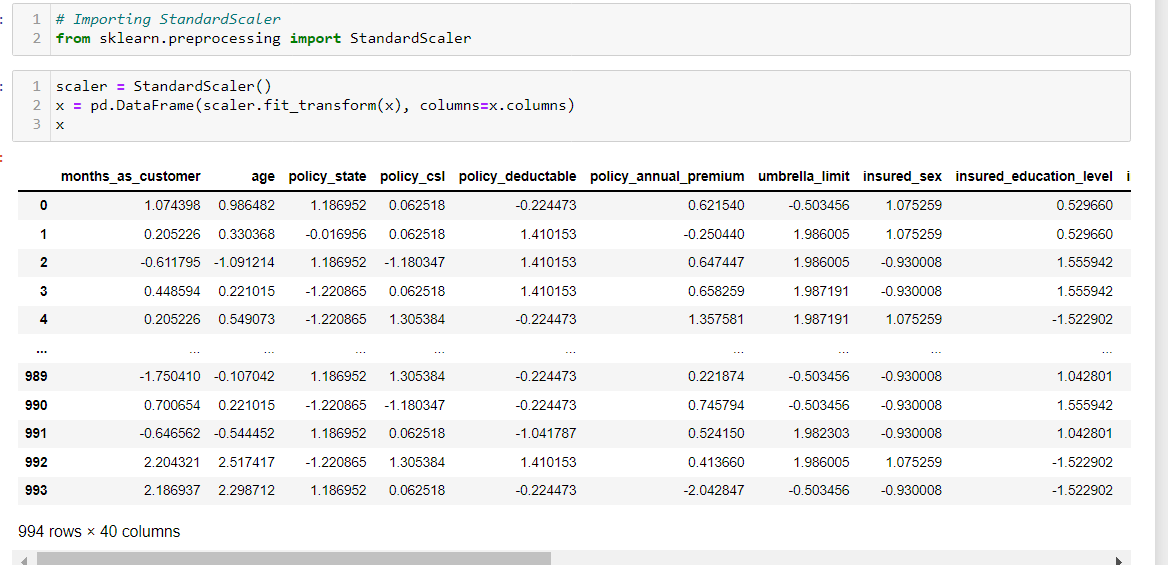
1. **Dividing features and targets and assigning them to X and Y**
2. **Scaling Data – StandardScaler**
3. **VIF – Variance Inflation Factor**
4. **Oversampling Data**

## **Dividing features and targets and assigning them to X and Y**

`

## **Scaling Data – StandardScaler**

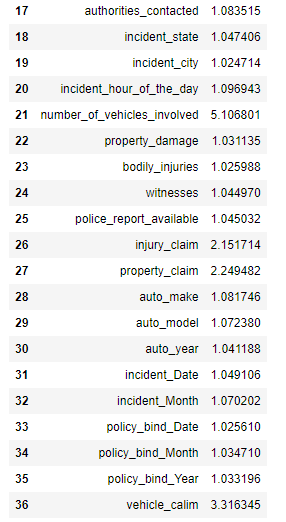
**The Standard Normal Distribution is followed by StandardScaler (SND). As a result, it adjusts the data to unit variance and sets mean = 0. If there are negative values in the datasets, MinMaxScaler scales all of the data features in the range [-1, 1] otherwise. All of the inliers in the restricted range [0, 0.005] are compressed by this scale.**



## **VIF – Variance Inflation Factor**

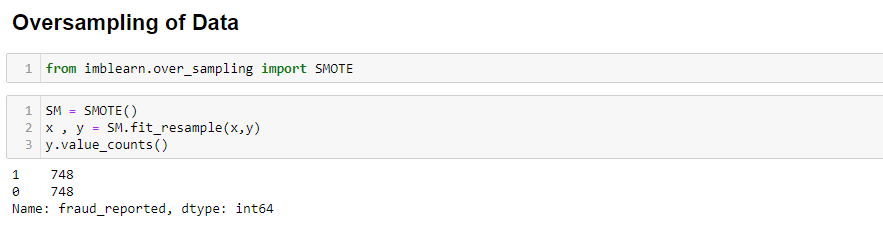
The variance inflation factor (VIF), which is used in statistics, is calculated by dividing the variance of estimating a particular parameter in a model that contains several other terms (parameters) by the variance of a model that is built using just one term. It rates how multicollinear a regression study using ordinary least squares is. It offers an index that gauges how much collinearity raises the variance (the square of the predicted standard deviation) of a regression coefficient.



## **Oversampling Data**

As our data is not balanced, we will use Smote technique from imblearn to balance our data.



Data is now balanced, and both results have the same numbers.

# **5. Building Machine Learning Models.**

We will now be being on building our Machine Learning Model. This is a classification problem.

## **What is the Classification problem?**

A classification model tries to infer some meaning from the values that were seen. A categorization model will attempt to forecast the value of one or more outputs given one or more inputs. Labels that may be used on a dataset are outcomes.

We will follow the below steps to Build Models.

1. Finding the best random state and Splitting Data into Train and Test using the Best random State
2. Building different classification models
3. Checking cross-validation score
4. Hyperparameter tunning using Grid Search Cv
5. ROC Curve
6. **Finding the best random state and Splitting Data into Train and Test using the Best random State**

Finding the best random state helps get the best accuracy score, so finding it is the best option before we start building our machine learning model.



We have got random state 67 with the best accuracy of 92%, we will further proceed with the next step.

1. **Building different classification models.**

**Classification algorithms using the below methods**

1. Logistic Regression

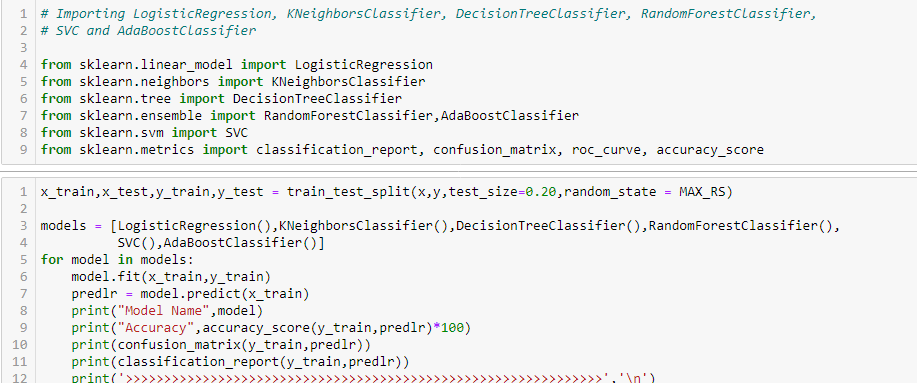
2. KNeighborsClassifier

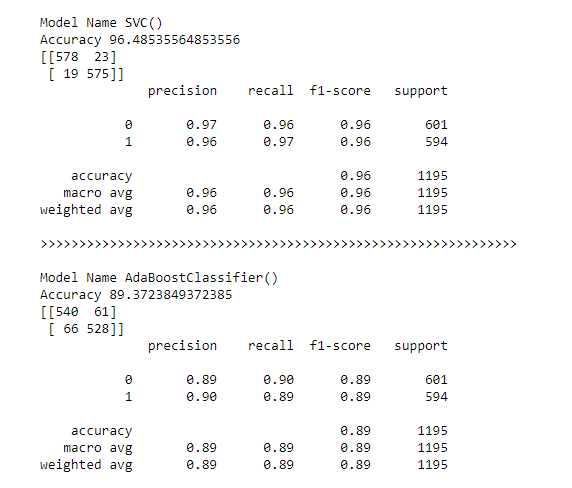
3. DecisionTreeClassifier

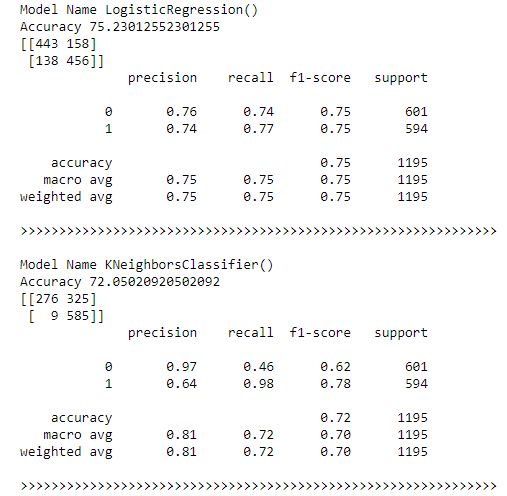
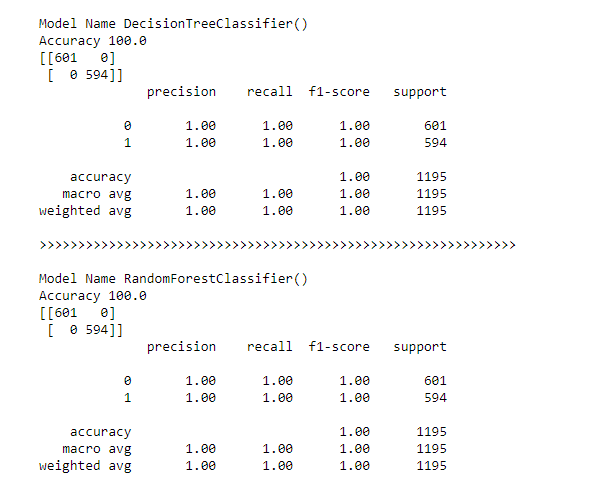
4. RandomForestClassifier

5. SVC

6. AdaBoostClassifier

****



From the above models, Random Forest Classifier has the highest accuracy score, we will further use the cross-validation technique to get an accurate score

1. **Checking cross-validation score**

Employing the complementary subset of the data set to test our model after it has been trained using the subset of the original data set, this approach is known as cross-validation.

The cross-validation score is also not final we will have to hyper-tune this and get a new score which will help us get the exact and accurate score.

Random Forest Classifiers have achieved the highest accuracy with approximately 87%. We will further use Hyper Parameter tunning to get the final accuracy score.

****

****

**Hyper Parameter Tunning**

Selecting the best collection of hyperparameters for a learning algorithm is known as hyperparameter tuning. A model input called a hyperparameter has its value predetermined before the learning process even starts. Hyperparameter weakening is essential for machine learning algorithms to work.

The accuracy score after hyperparameter tuning is 89.29%. It is approximately increased to 2-3% which is great and we can use this as a final accuracy score.

****

We have found the best parameters as above, this will be further used to train the model.

{'criterion': 'Gini',

'max\_depth': 8,

'max\_features': 'auto',

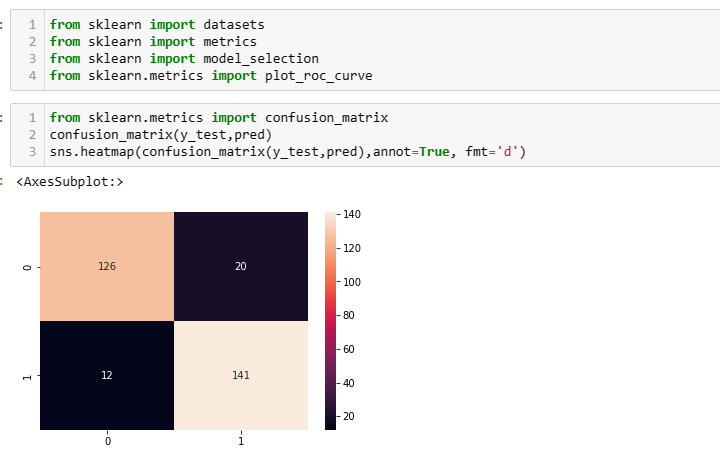
'n\_estimators': 200}

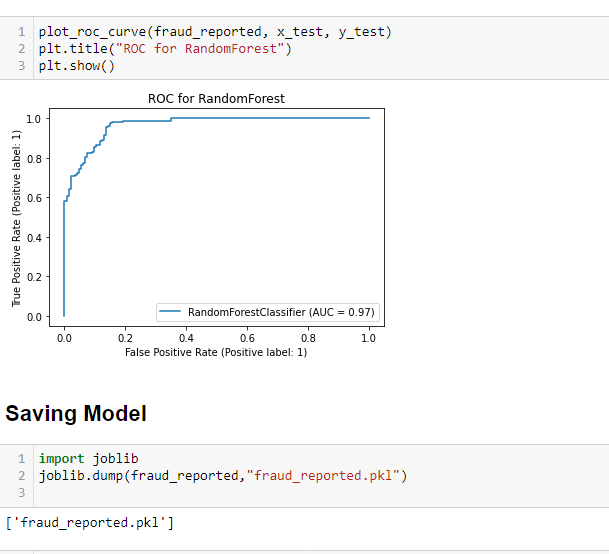
****

The final accuracy score after tunning is 89%. We have used Random Forest Classifier to achieve this accuracy.

**5. Roc Curve**

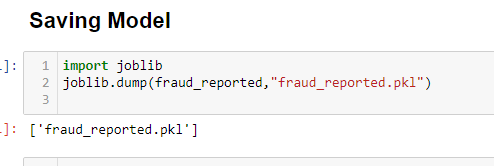
ROC stands for Receiver Operating Characteristic. ROC curves are a convenient visual tool for analyzing two classification models. ROC curves appear from signal detection theory that was produced during World War II for the search of radar images.

****

****

**Saving Model**

Saving model for future use.

****

# **6. Concluding Remarks.**

The most important step in this project was feature engineering, where we eliminated skewness and outliers. Additionally, categorical columns were handled by encoding the data, scaling the data, and finally, building various classification models to determine whether an insurance claim is fraudulent or not. Hyperparameter tuning was also carried out to enhance the model using various parameters.

Insurance fraud poses a serious financial danger to insurance companies. They ought to act proactively to enhance their fraud prevention methods. In the insurance industry, predictive analytics, AI, and machine learning are now gaining traction quickly and helping insurers restructure their whole operation. The insurance claim settlement procedure will be significantly improved by artificial intelligence, making it less tedious and more engaging. The insurance claims procedure may be automated to free up human resources from paperwork-reading duties. These new technologies offer a large return on investment in fraud analytics since they assist insurance firms in stopping fraud leaks in the claims sector.