**Insurance Claims- Fraud Detection**

**1-> Problem Statement**

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.

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.

**2-> Data Analysis**

**Data analysis** is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

#Importing necessary libraries

**Graphical user interface, text, application

Description automatically generated**

# Loading csv file

Table

Description automatically generated

# **First five rows of the dataframe by default**

Table

Description automatically generated

# **Columns of the dataset**

**Text, letter

Description automatically generated**

# **Checking the null value**

**Shape, rectangle

Description automatically generated**

# **Dimensions of the array**

**A picture containing graphical user interface

Description automatically generated**

# **Datatypes**

**Table

Description automatically generated**

# **Representation of the data set (Heat map)**

**Graphical user interface, application

Description automatically generated**

# **Statistical representation**

**Table

Description automatically generated**

# **Replacing ‘?’ with NULL and then checking the dataset**

**Table

Description automatically generated**

# **Importing datetime library**

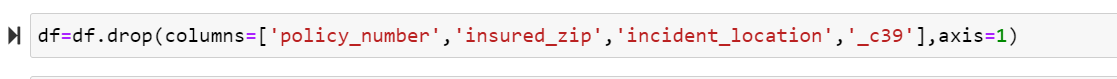
****

# **Coverting the date into datetime**

**Text

Description automatically generated**

**# Dropping the unnecessary column**

****

**# Ploting the dataset**

**A picture containing bar chart

Description automatically generated**

**# Checking the correlation**

**Table

Description automatically generated**

**# Again checking the null value in the dataset**

**A picture containing table

Description automatically generated**

**#Checking the plot figure of fraud report**

**Chart, bar chart

Description automatically generated**

**#Checking the numbers of fraud report**

**Graphical user interface, text, application, email

Description automatically generated**

**#Visualizing the dataset through heatmap**

**Chart

Description automatically generated**

**#Checking the number of vehicle which are fraud reported using the count plot.**

**Chart, bar chart

Description automatically generated**

**#Checking the number of vehicle which are fraud reported using the count plot.**

**Chart, bar chart

Description automatically generated**

**#Checking the number of capital plot on graph.**

**Graphical user interface, application

Description automatically generated**

**#Checking the capital loss with fraud report**

**Chart, radar chart

Description automatically generated**

**#Checking the injury claim with fraud reported on plot graph**

**Chart

Description automatically generated with medium confidence**

**#Checking the witnesses with fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the incident hour of ther day with fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the policy deductable with fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the number of age with fraud reported on plot graph Chart, bar chart, histogram

Description automatically generated**

**#Checking the insured education levelwith fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the insured occupation with fraud reported on plot graph Chart, bar chart

Description automatically generated**

**#Checking the incident state with fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the incident state with fraud reported on plot graphChart, bar chart

Description automatically generated**

**#Checking the policy state with fraud reported on plot graphChart, bar chart

Description automatically generated**

**#Checking the property claim with fraud reported on plot graph**

**Graphical user interface, chart

Description automatically generated**

**#Checking the insured hobbies with fraud reported on plot graph**

**Chart, bar chart

Description automatically generated**

**#Checking the insured sex with fraud reported on plot graphChart, bar chart

Description automatically generated**

**#ploting the dataset**

**Table

Description automatically generated**

**#To plot multiple pairwise bivariate distributions in a dataset**

**Table

Description automatically generated**

**Chart

Description automatically generated with medium confidenceTable

Description automatically generated**

**3-> EDA concluding remark**

In this model project, we did EDA in which we have checked the null values using heat map and converted the categorical data into numerical forms and dropped the irrelevant columns from the data set , we have also counted and plotted the data required. We have also used pair plot feature . In this dataset we have replaced “?” with NULL.

**# To check the Outliers**

Outliers are uncommon values to your dataset, and they could distort statistical analyses and violate their assumptions. Unfortunately, all analysts will confront outliers and be pressured to make choices approximately what to do with them. Given the issues they could cause, you may suppose that it’s first-rate to get rid of them out of your data. But, that’s now no longer usually the case. Removing outliers is valid most effective for precise reasons.

**Table

Description automatically generated**

**We did this to check the outliers for the dataset of the insurance and further on we will do the removal of the outliers**

**#Removal of the outliers**

**Table

Description automatically generated**

**Checking the deficit of the data and loss in the dataset after removal of the ouliers .**

**Background pattern, table

Description automatically generated**

**Removal of data is 1.9 %**

**#Skewness**

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean

**Table

Description automatically generated**

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

**#Removal of Skewness using YEO Johnson**

**Graphical user interface

Description automatically generated with medium confidence**

**#Correlation**

Correlation **measures the relationship between two variables**. We mentioned that a function has a purpose to predict a value, by converting input (x) to output (f(x)). We can say also say that a function uses the relationship between two variables for prediction.

**Table

Description automatically generated**

**#Converting categorical data into numerical**

**Machine Learning Models can not work on categorical variables in the form of strings**, so we need to change it into numerical form. We can assign numbers for each categories but it may not be that effective when difference between the categories can not be measured.

**Table

Description automatically generated**

**4-> Pre-processing Pipeline**

Pipelines are a easy manner to hold your information preprocessing and modeling code organized. Specifically, a pipeline bundles preprocessing and modeling steps so that you can use the entire package as though it have been a unmarried step. Many information scientists hack collectively fashions with out pipelines, however pipelines have a few essential benefits. Those include: Cleaner Code: Accounting for information at every step of preprocessing can get messy. With a pipeline, you might not want to manually hold tune of your education and validation information at every step. Fewer Bugs: There are fewer possibilities to misapply a step or neglect about a preprocessing step. Easier to Productionize: It may be notably tough to transition a version from a prototype to some thing deployable at scale. We might not pass into the various associated issues here, however pipelines can help. More Options for Model Validation: You will see an instance withinside the subsequent tutorial, which covers cross-validation.

Spliting data into x and y axis

Graphical user interface, text, application

Description automatically generated

Checking the x-y values

Graphical user interface, application, table

Description automatically generated

* Standard Scaler

Graphical user interface, text, application, email

Description automatically generated

StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature.

#**Over sampling Technique**

Z rating is an crucial idea in statistics. Z rating is likewise referred to as preferred rating. This rating facilitates to recognize if a facts cost is extra or smaller than imply and the way some distance away it's far from the imply. More specifically, Z rating tells what number of preferred deviations away a facts factor is from the imply.

Graphical user interface, text, application, email

Description automatically generated

**5>Building Machine Learning Models**

A system mastering version is constructed via way of means of mastering and generalizing from education information, then making use of that obtained information to new information it has in no way visible earlier than to make predictions and satisfy its purpose. Lack of information will save you you from constructing the version, and get right of entry to to information is not enough.

Graphical user interface, text, application

Description automatically generated

Importing important libraries for building of the data machine models .

\*\*Score\*\*

Text

Description automatically generated

Gradient Boosting classifier

It is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Graphical user interface, text, application, email

Description automatically generated

AdaBoostClassifier

A meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

Graphical user interface, text, application, email

Description automatically generated

BaggingClassifier

Bagging classifier is an ensemble meta-estimator that suits base classifiers every on random subsets of the authentic dataset after which combination their man or woman predictions (both via way of means of balloting or via way of means of averaging) to shape a very last prediction. Such a meta-estimator can normally be used as a manner to lessen the variance of a black-field estimator (e.g., a choice tree), via way of means of introducing randomization into its production technique after which making an ensemble out of it.

Graphical user interface, text, application, email

Description automatically generated

KNeighborsClassifier

K-Nearest Neighbour is one of the handiest Machine Learning algorithms primarily based totally on Supervised Learning technique. K-NN set of rules assumes the similarity among the brand new case/statistics and to be had instances and placed the brand new case into the class this is maximum just like the to be had categories. K-NN set of rules shops all of the to be had statistics and classifies a brand new statistics factor primarily based totally at the similarity. This way whilst new statistics seems then it could be without difficulty categorised right into a nicely suite class via way of means of the usageof K- NN set of rules. K-NN set of rules may be used for Regression in addition to for Classification however in the main it's far used for the Classification problems. K-NN is a non-parametric set of rules, this means that it does now no longer make any assumption on underlying statistics. It is likewise known as a lazy learner set of rules as it does now no longer examine from the education set straight away as an alternative it shops the dataset and on the time of classification, it plays an movement at the dataset. KNN set of rules on the education section simply shops the dataset and whilst it receives new statistics, then it classifies that statistics into a class this is a good deal just like the brand new statistics.

Graphical user interface, text, application, email

Description automatically generated

DecisionTreeClassifier

Decision Tree is one of the maximum effective and famous set of rules. Decision-tree set of rules falls below the class of supervised gaining knowledge of algorithms. It works for each non-stop in addition to express output variables.

Graphical user interface, text, application, email

Description automatically generated

RandomForest Classifier

A random wooded area is a meta estimator that suits some of selection tree classifiers on diverse sub-samples of the dataset and makes use of averaging to enhance the predictive accuracy and manage over-fitting. The sub-pattern length is managed with the max\_samples parameter if bootstrap=True (default), in any other case the entire dataset is used to construct every tree.

Graphical user interface, text, application, email

Description automatically generated

ROC curve

ROC curve, additionally called Receiver Operating Characteristics Curve, is a metric used to degree the overall performance of a classifier model. The ROC curve depicts the charge of real positives with appreciate to the charge of fake positives, consequently highlighting the sensitivity of the classifier model.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidenceA picture containing line chart

Description automatically generatedChart

Description automatically generated with medium confidenceA picture containing graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Joblib is a set of tools to provide lightweight pipelining in Python. In particular: transparent disk-caching of functions and lazy re-evaluation (memoize pattern) easy simple parallel computing

**6-> Concluding Remarks**

Insurance fraud is a planned deception perpetrated in opposition to or via way of means of an coverage organisation or agent for the motive of economic gain. Fraud can be dedicated at exceptional factors via way of means of applicants, policyholders, third-celebration claimants, or experts who offer offerings to claimants. Insurance retailers and organisation personnel may devote coverage fraud. Common frauds include “padding,” or inflating claims; misrepresenting records on an coverage application; filing claims for accidents or harm that in no way occurred; and staging accidents.

In this project, we have seen that data is visualized using the python and juypter notebook. We have converted the categorical data into numerical and we have corrected the data set and balanced the data using the machine learning model and also we have used the plot to check the data on graph.

We have also checked the skewness of the data and removed the skewness , and also we have checked the outliers and removed the outliers using the zscore method

By checking through various process such as LogisticRegression, DecisionTreeClassifier, DecisionTreeClassifier, GradientBoostingClassifier, AdaBoostClassifier , BaggingClassifier We have checked the best score is given by the Random Forest and also accuracy score didn’t increased.

Also we have checked that the actual and the predicted value is similar so our model is working good ..

Also at the end we have find the ROC and RUC to analyse the data .

By using this model we can prevent the insurance company from wrong fraudulent claims and also we can save the insurance company from loss.