**Insurance Claim Fraud Detection**

**Vinod Adep Satyanarayana**

**1.Problem Definition :-**

I have a dataset of Insurance Claim Fraud Detection, Features are 39 variable and 1 Label totally 40 columns in data, 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, I 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 made.

In this example, I will be working with some auto insurance data to demonstrate ,I create a predictive model that predicts if an insurance claim is fraudulent or not.

This pandemic peoples are applying Mediclaim policy and takes care of health by protecting with Mediclaim policy, but lots of peoples are miss using the claims with guidelines ,for this prevention I created a model to predict claims are fraud or not.

**2.Data Analysis :-**

Insurance Claim is best investment to save our lives in any condition, I have a details of claims

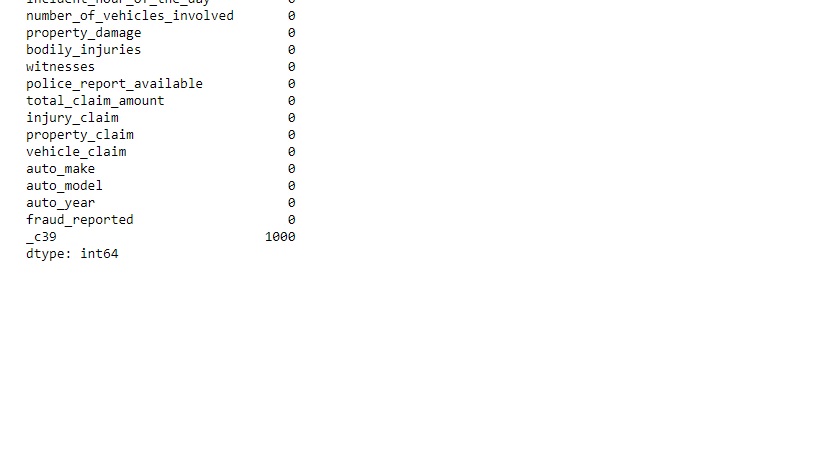
Made, lots of information is available for analyse data which are customers p/month, policy no, age, policy date, policy deductable, policy annual premium, and so many , details of auto accident ,details of insured person ,etc.

Data have lots objects data types because this type of projects consist details of sex, education, occupations, employment ,I have to do more work on this project with features ,Also null values present in only one column so it would be easy for handling data, Shape of data is 1000 rows and 40 columns it’s very difficulty with data cleaning because, when more features are consist we have to more work on the project ,Data Analyse is depend upon youself how you will understood that data .

Claims are made by particular conditions of company with guidelines ,They had already studied about how accidents happens and health risks, But we don’t have any options to play with claim policy ,This project having a deeply information of claims above I didn’t mentioned much more info, but whenever we are going study in deeply lots of details available which are includes police officers, incident severity, authorities contacted, bodily injured , witness, accident collision type, incident state, etc.

**3.Explotary Data Analysis(EDA) :-**

Lots of EDA is observed in this project, \_c39 column have totally null values I’m dropping this columns due to whole columns having null values it is not use full for prediction.

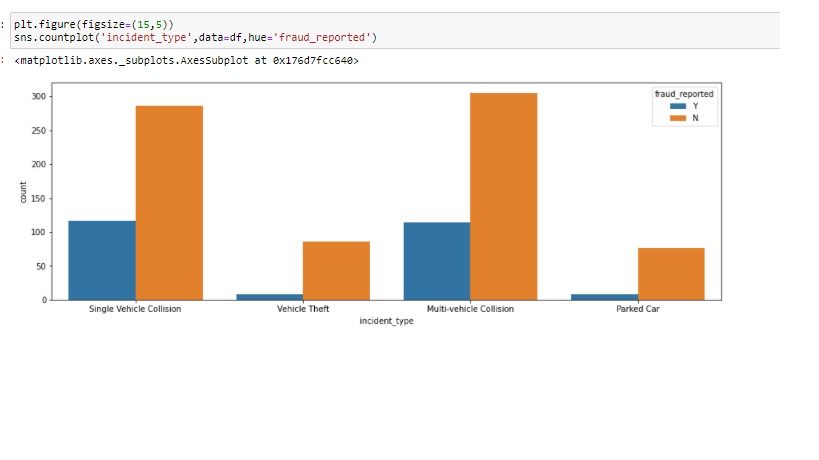


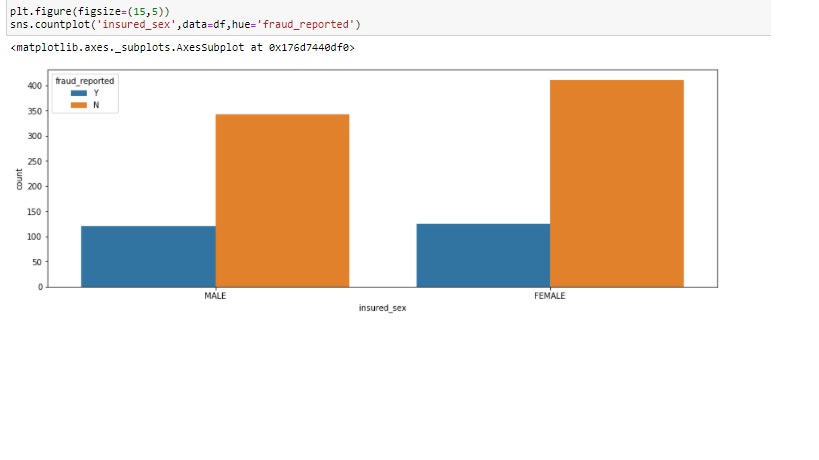
When auto incident is happened two major category is observed incident type and how many vehicles are involved, Incident type having 4 categories for this project which are Multi-vehicle, Single-vehicle, Vehicle theft, Parked car also Number of Vehicle are involved in this accident. I dropped some columns are insured education level, insured hobbies, incident location, incident state, incident city, auto make, auto model, auto year due to use less data for this project .

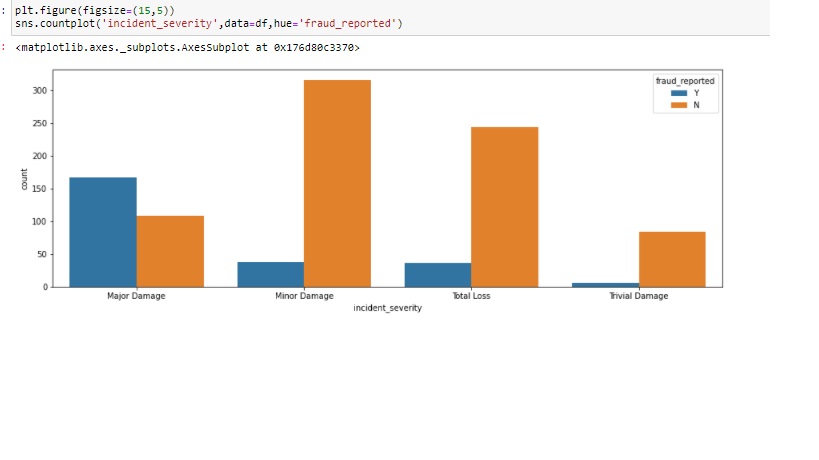
Some features have object data we have to convert them “int” or splitting into “int” form ,The following are in date format policy bind date and incident date , I splitting into date month year category wise ,which is made separately incident date, month, year and policy date, month, year . Also policy csl is available in data, it describes predetermined limit for the combined total of the Bodily Injury Liability coverage and Property Damage Liability coverage per occurrence or accident , I converting into policy csl claim ,policy csl limit .

Fraud Reported label is the our target variable ,fraud report in Female is more than Male, Insured Occupation feature is have more categories I think in this feature we have to apply label encoder for “int” format. Incident type have 4 categories which are Single & Multi vehicle collision is more than other 2, Here we have to analyse on thing on label ,label is balanced data are imbalanced ,fraud reported label having “Y” are 753 values and “N” are 247 values. Also Insurance company will claim basis on how much damage to person due to incident, Incident severity have 4 categories are Major Damage are more than other 3.This is all for done analysing data

I applied pd.get\_dummies to nominal data to from in binary data and Label Encoder to Incident occupation it haves more categories data.







So now finally The whole data made in binary data format we can predict this data with model generation.

**4.Building Machine Learning Models :-**

Splitting of Data into train and test for the model prediction ,we cannot model prediction without splitting of data ,I given values of test size=.30 and random state = 1

Why we have to use machine learning model for prediction ,The answer is machine learning have a capability of read a binary data and gives a accuracy for that particular model. With machine learning we can solve the testical part of AI(Artificial Intelligence).This part is very important for outcome of project or dataset ,with this result anyone can build a project or making business decisions.

To Build a Model what we want, following are requirements splitting of data into train & test ,we have to split data on features and label ,test size which is given to model for prediction results it will depend upon you, how much you want to give but there is certain limit for test size 30 to 25 ,because we have to give more than 70% data to training ,In Model generation training and test split is very important role, So I could consider Machine Learning modeling is big role for any predictions.

I applied classification algorithms on training data. I had given this models for training data Random Forest Classifier gives 80% accuracy, Logistic Regression gives 73 % accuracy, Ada Boost Classifier gives 69 %accuracy, Decision Tree Classifier gives 63 % accuracy, KneighborsClassifier gives 67% accuracy. These all algorithms are given different accuracy because they have different functions and working ,but we need to try all algorithms with different parameters, The problem is based on Classification problem ,it contains metrics such as accuracy score, confusion matrix, classification report ,auc score and so may.

After model applying ,we have to analyse on confusion matrix and classification report this will gives the outcome is good fit to model or not. Also I had done Hyper Parameter Tuning ,I applied DecisionTreeClassifier to [Criterion: “gini”,”entropy”] then model gives best parameter “gini” with 66% accuracy. Also I had tried Cross Validation score with different algorithms , Finally the Best Model for this Dataset is RandomForestCLassifier gives 80% accuracy.

**6.Conclusion :-**

The Best Model for this dataset is RandomForestClassifier gives 80% accuracy. I think this project is used for fraud report of claim ,Insurance Company will easily identify the who’s miss using the claim with scam. In pandemic Insurance Company having lots of claims with covid cases and accidents or heart strokes, so this conditions they’ll hard to handle and they have don’t know who’s one is right or who’s doing scam ,for this prevention I build model with the given data, it will easily identifies the fraud report .