****

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Insurance Claims- Fraud Detection**

Submitted by:

**Shubham Sharad Kathoke**

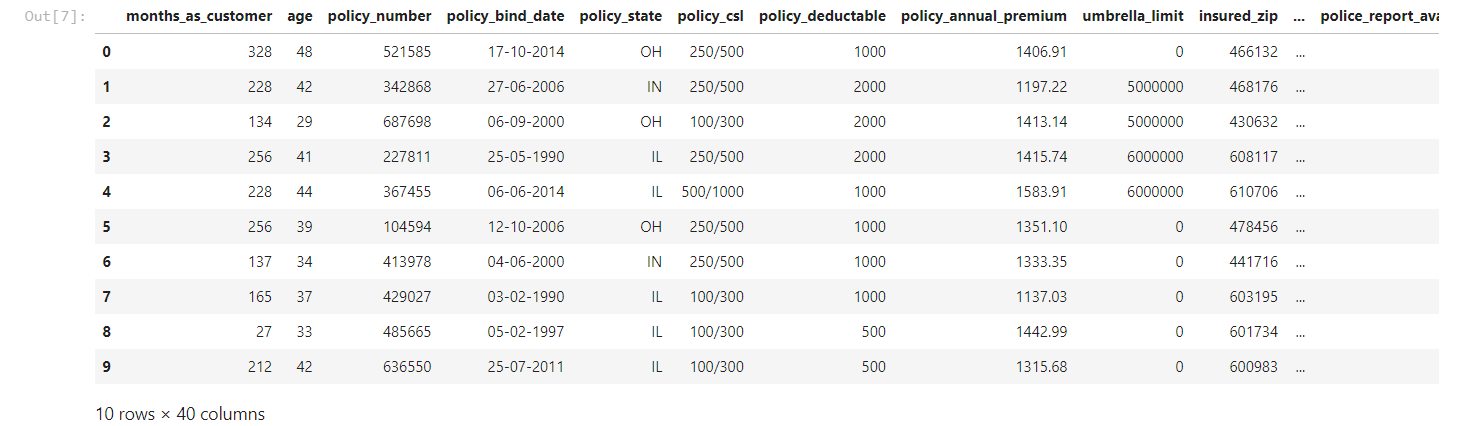
**1. Problem Definition -**

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.

Fraud’s detection using machine learning is less compared to legit insurance claims. While building detection models, the savings from less prevention needs to be balanced. Machine Learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates

**2. Data Analysis –**

In this project, we have a dataset which has auto insurance policy details along with the customer details. The given dataset contains 1000 rows and 40 columns. The column names like months\_as\_customer, age, policy\_number, policy\_bind\_date, policy\_state, policy\_csl, policy\_deductable etc

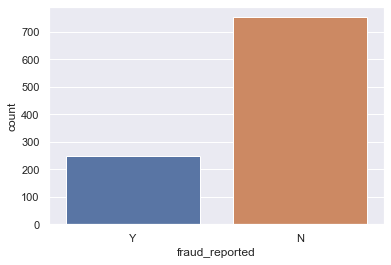
****

Companies which have small datasets is more likely to succeed in machine learning than the companies having huge datasets.

There are some null values present in the given dataset. So I checked it and drop that unnecessary coloumn

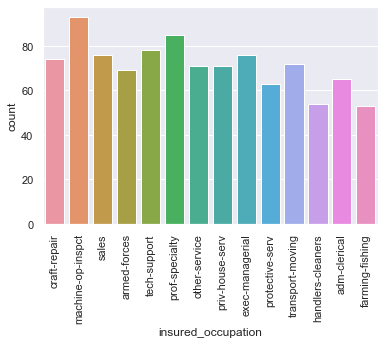
**3. EDA Concluding Remarks –**

**Visualization :**

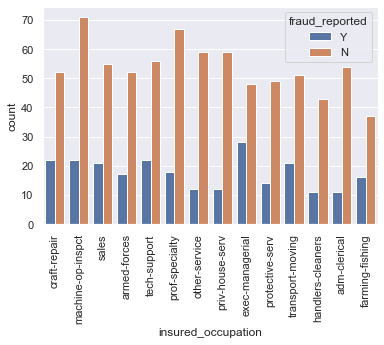


Fraud reported has No count- 700

Fraud reprted has Yes count – 200



In thisinsured\_occupation : machine-op-inspct has highest count and farming-fishing has least count



#### In this insured\_occupation : machine-op-inspct has least fraud cases and exec-managerial has highest fraud cases

**Dependent Variable :** The dependent or output variable in the given dataset is the column ‘fraud\_reported’. There are 247 fraud claims and 753 non-fraud claims. 24.7% of the claims were frauds and 75.3% were non-fraudulent claims.

#### 

**Correlation :** Correlation between dependent and independent variables is plotted using heatmap. Column incident\_severity has maximum correlation with 41% correlation

**4. Pre-Processing Pipelines –**

Data pre-processing is an important step in machine learning to get highly accurate and reliable result. Data pre-processing helps in increasing the quality of data by filling in missing data's(NaN values), removing outliers, scaling the data

Here I use Scatter Plot to check the Correlation and use Label Encoder to convert the dataset to numerical form . Then I checked skewness and the taking range as +0.5 to -0.5.

Then I used Boxplot to check the outliers present in the dataset and find out outliers are present in 4 columns. So, by using using z\_score technique outliers removed. After check for total data loss and 2% of data has been lost after zscore

Checking for Multicollinearity using Variance Inflation Factor

**5. Building Machine Learning Models –**

For building machine learning models there are several models present inside the Sklearn module.

Sklearn module provides two types of models i.e regression and classification.Our datasets target is to predict whether fraud is reported or not. So, for this kind of problems,

Before fitting our dataset to its model first we have to separate the predictor variable and the target variable, then we pass this variable to the train\_test\_split method to create a random test and train subset

Here we use different techniques :

**So For Prediction I used –**

Logistic Regression, Decision Tree Classifier, Random Forest Classifier, SVC and checked the cross validation score and find out Random Forest and its gave me the best accuracy score. So, I hyper tuned the Random Forest Classifier and got the accuracy score of 82.14%

We checked ROC AUC Plot and its use for represents the degree or measure of separability. Higher the AUC, the better the model is at predicting.

**6. Concluding Remarks –**

We got our best model i.e. Random Forest Classifier with the accuracy score of 82.14%. Save the model using pickle function.