**Insurance Claims – Fraud Detection**

**Problem Definition**

Fraud is one of the critical issues of the insurance industry. The meaning of fraud in the insurance industry is to knowingly make a fake claim, exaggerating or adding items to it, or in any way being fraudulent to gain more than legal entity. Identifying these fraud claims is difficult which causes huge losses. Considering the huge amount of data, the insurance industry is growing rapidly at a very fast rate. The conventional way of identifying the fraudulent claims will not work as the data size increases. It will become challenging to predict as the new types of claims arise.

Machine Learning is in a unique position to help the Auto insurance industry with this problem. The task is to create a predictive model that predicts if an insurance claim is fraudulent or not and create an auto insurance fraud detective model. Frauds are corrupt due to which companies face massive losses. To cut out these losses for the insurance industries, we will build a Machine Learning model that helps in classifying auto insurance fraud. Machine Learning techniques allow us to enhance predictive accuracy, permitting the loss controls to accomplish higher scope with false positive rates.

**Data Information**

The dataset consists of 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. Here, we are working with some auto insurance data to demonstrate a predictive model that predicts if an insurance claim is fraudulent or not. We have trained and built a Machine Learning model to help the Auto insurance industry with the problem.

The dataset companies 1000 rows and 40 columns in total including the output variable, “Fraud- Reported” is the target variable in this example.

**Data Analysis**

**Data Collection**

The dataset consists of auto insurance claims from Ohio, Illinois and Indiana dated from 01/01/2015 to 01/03/2015. This information has been retrieved from an online data source (github.com). The data received was in the form of a csv file. Before performing any operations or transformations (features engineering), the dataset comprises a total 39 input features and a target Variables. We are not sure of the fact whether the data id of the single company or multiple insurance companies through the con is that the data sample is of a small size. Hence, we being working with small Datasets and create big ones as time goes on to succeed in order to profit a company’s revenue.

**Cleaning and Data Preparation**

Success rate in Machine Learning is achieved depending on the way the data is represented. Features Engineering is a part of data cleaning and preparation which transforms raw data into variables that better the performance of the model to give accurate predictions. We identify the features relevant to finding the prediction as that of the actual output.

In this case, we have featured “Policy Number” as the index variables and split variables “Policy Bind Day” to retrieve “Policy Bind Date”, “Policy Bind Month” and “Policy Bind Year” from the existing data to make the model understand data better. Feature “\_c39” consisted of all missing values that served no value to the Dataset, Hence the variables was deleted. Few uncertain”?” values found in the dataset were replaced with NaN values to fill the missing values using the “Most Frequent” strategy of the SimpleImputer() function, Encoding all strings values using OrdinalEncoder() to convert it to floating values. Therefore , all the feature engineering procedures were used to get rid to the corrupt data and prepare a dataset with useful information.

**Data Analysis Using Visualizations**

Dividing information, understanding its patterns and correlations is the most significant part of analysis and preparing data. After this step, ML algorithms are Applies for training the model to get predictions. For this dataset, we have divided data into two data frames, on the basis of the data types of its value i.e. object or numeric data type.

**Visualization of the String Value**

For the string data, a count plot graph was used as it gives the frequency of the columns. Observation made while analysing data are –:

* Female customers are more than the male customers.
* More than 700 people have been reported with no fraud but around 250 people have been reported with Fraud.
* 300 customer’s Property has been damaged and 700 customer’s property has not been damaged.
* Around 320 customer’s police reports are available and 680 customer’s police reports are not available.
* Incidents are high in states like SC, NY and WV.
* Rear Collision is the max collision type.
* The incident severity is minor damage in most of the cases.
* There are too many categories in the column like “auto model” , “auto make” , “incident location” , “incident date” , “insured occupation” & “insured hobbies”. Hence , it is difficult to conclude any observation out of these columns.

**Visualization of Inf/Float Values**

We have used dis-plot and scatter plot graphs to understand data with Inf/Float values.

Observation made are –

* Variables “Policy Bind Year” , “Policy Bind Month” , “Policy Bind Day”, “Auto Year” & “Incident Hour of the Day” are showing values at a constant rate.
* Variable “Umbrella Limit” consists of maximum number of 0 values.
* We could observe negative values in the “capital-loss” feature.
* “Capital-Gains” consists of values ranging till positive 120000.
* “Capital-Loss” consists of values ranging till negative 130000.
* Variables like “Number of Vehicles Involved” , “Bodily injuries” & “Wintenesses” have a nominal/categorical type of data.

**EDA Concluding Remarks**

Performing complete EDA on the data (cleaning, integrating and transforming of data), we get a dataset with 1000 rows and 40 columns.

Concluding Observations are-

* The standard deviation of the variables in the dataset is very huge which means that the values in these columns are largely scattered and are not near means values. They are very far from their mean values.
* The values inside the dataset ranges from high negative values to high positive values . The value ranges are very high within he dataset.
* The min & max values in every feature have huge range differences.
* Understanding data properly is difficult due to the huge number of columns.
* The most negatively correlated variable to the target variable is “Incident Severity”.
* The most positively correlated variable to the target variable is “Vehicle claim”.
* High positive correlation was not observed in the data with respect to the target variable.
* There are 22 features which are positively correlated with the target Variable.
* There are 17 features which are negatively correlated with the target variable.

**Pre-Processing Pipeline**

**Skewness Correction**

Considering threshold value as+/0.5 as the range for skewness , we could see skewness in features such as “umbrella limit” , “insured zip” , “police report available” , “total claim amount” , “vehicle claim” & “fraud reported”. The skewness for the required column were resolved using power transform function.

**Outliers Detection And Corrections**

Maximum outliers were observed in the “umbrella limit” features and in few other variables as well. These outliers were resolved using z- score technique. The information after the removal outliers was found to be 2% which is not much of a huge information loss. The shape of the dataset before outliers removal was 1000 rows and 40 columns and that after removal came out to be 980 rows and 40 columns.

**Normalization**

As the values in the data set have high ranges , it becomes complex for an ML model to understand and read data , hence data training becomes difficult which is not a proper way to deal with data to achieve good accuracy and get accurate predictions. Therefore , it is very important to normalize/standardize data which means getting data within a certain range to have proper understanding of data. In this example, we have used standardscaler() techniques to normalize the data which brings data between the range of 0 to1.

**Decision Tree Classifier**

Visualisation view of the Decision tree is in the form of the graphs. Decisions tree divides the main data set into a subset of trees that consists of choice and results. Node of each tree depicts a choice and the edges depicts the decision. The main data data set is categorized into 80% training data and 20% testing data. A model is built with a training data set which predicts the accuracy. This model is applied on the test data and the predicted accuracy is validated. The model gives us a prediction in the form of a Yes/No for this data.

**Cross Validation** technique was applied on both Logistic Regression and decision Tree Classifier algorithms to overcome the issues of overfitting and underfitting. The mean cross validation score was obtained to compare both the algorithms and check which algorithm works best and gives highest accuracy for the given data. Since both algorithms predicted 100% accuracy , Logistic Regression was used for further hyper tuning of data.

Used of  **“Grid Search CV”** techniques was made to hyper-tune the algorithm with the best found parameters i.e. “Max Iter” at 1 and “Penalty” as l1 to achieve the best accuracy.

**Concluding Remarks**

* The f1 score , precision and recall for both Logistic Regression and Decision Tree Classifier is 100%. Hence, accuracy score for both the algorithms is also 100%.
* Comparing the CV score and accuracy score both algorithms , the accuracy achieved was 100% . Therefore , Logistic Regression was selected as the model.
* Performing hyperparameter tuning on the selected algorithm with the best found parameters using GridSearchCV , the predicted accuracy comes out to be 100%.
* Hence , we have achieved 100% accuracy for the model after trying different algorithms and different testing techniques on the data.