Problem Statement:

The main aim of this project is to build a model to detect the fraud taking place in insurance sector. Here we are particularly focusing on auto insurance and this would be more challenging. This data set is imbalanced and we would need to classify whether fraud has happened during the claim.

Frauds are unethical and not appreciated by insurance companies as they are loss to the company. By building a model that can classify whether fraud has happened or not in an auto insurance company we can reduces the losses to the company and increase its earnings.

Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for the purpose of financial gain. Fraud may be committed at different points in the transaction by applicants, policyholders, third-party claimants, or professionals who provide services to claimants.Auto insurance fraud ranges from misrepresenting facts on insurance applications and inflating insurance claims to staging accidents and submitting claim forms for injuries or damage that never occurred, to false reports of stolen vehicles.

Data Analysis:

We see that imbalance data at the target variable in most of the data set from different industries. Many times we are interested in a minority class against majority class or classes. This model should be able to classify if a claim is fraud or not on a data set which is not seen accurately. This is measured by the F1 score and AUC ROC curve if the model is able to classify between fraud and legit claims.

In the data set provided to us we can see that there are numerous columns present in the data set and we can see that policy number would be unique to every user and this can be set as our index column as it would help us reduce the columns and the user can be tracked with the policy number. We need to check if null values are present in the data and try to remove if the majority of the rows in a columns have null values else replace those with either mean or mode of the data present in the column depending on the type of the data present in the columns.

Since the machine learning models can only be trained on integral/float data we need to convert the object data type to integral data type using Label Encoder so that model can predict the output with better accuracy.

EDA Concluding Remarks:

After converting the data into integral data type we can perform EDA process and find the insights of data. We can apply describe function to find the mean, median, percentile , min and max values for each columns and based on this we can make decisions on the data. If the mean of the data is more than median then the data would be skewed or if the there is more difference between the 75th percentile and the max value in the column then there would be chances of outliers present in the data. Since insurance data contain mostly categorical data we can find these observations only in columns having integral values and we see this in few of the columns.

We can perform **bivariate analysis** between any of the input variable with the output variable to see how each input columns is behaving with output variable so that it would be easy for us to make certain decisions looking at the graphical representations. We can also check value counts in the categorical data columns to find which of these are maximum in the column and how they would influence the output variable.

We can later find the **correlation** among the columns to find if any columns are highly correlated with each other or if they have negative correlation. We can also check correlation of the columns with the output variable to see how each column is correlated with the output. We can also remove certain columns which do not provide any information in predicting the outputs. Using heat maps we can plot the graphical representation of the correlations in the matrix form which it would be easy for us to analyse the data.

We need to check the **skewness** in the data setting a threshold limit as the model performs well for normalised data. So looking at the skewness in the data if we find high skewness and if they are not providing any information in predicting the output we can remove those but if they are useful columns then we can remove skewness using different method. Here in this insurance data set we find skewness in umbrella limit column since its is good correlation with the output variable we need to retain this column. Using distplot we can plot the skewness of the columns.

We need to check **outliers** in the dataset as these are non uniform values present in the data. Outliers are mostly present in the integral data type and we can remove outliers present in the data using z score technique. But while removing outliers if there is more loss of data then we need to retain them depending on the circumstances as more loss of data can lead to misconceptions from the model. So we need to consider all these aspects while removing outliers in the data. In the insurance data we find only certain columns having outliers and removing these outliers causes only 2% of loss in data which is acceptable.

So EDA process involves clear observation and cleansing of data so that it can be helpful for the model to provide good predictions.

 Pre-Processing Pipeline:

After cleaning the data we need to check the data if its normalised as the model provides good prediction for normalised data. We need to remove the skewness in the data and this can be done using boxcox or yeo jhonson method.This will remove the skeness in the data. Once skewness is removed from the data we need to scale the data so that its normalised. We can scale the data using Standard Scales or Min max Scaler. We have used standard scaler here as it can scale the data to mean=0 and standard deviation=+/-1. So once the data is scaled it would be normalised and it now ready to be trained to the machine.

We need to split the data into independent and dependent variable before loading it into the machine for training. Since here fraud detection column is our output variable we would consider it as our target variable.

Building Machine Learning Models:

After analysing the data , performing EDA and pre-processing it we need to now train the model with our data set. Since in this Insurance data we are trying to classify between fraud an legit we would be using calssifiaction model.We would different classification model and train the data to check who can provide us the best accuracy and good f1 score in classifying it. We can see that among all the models Decision tree Classifier is providing highest accuracy of 77% but even this model is failing in providing good f1 score.

Looking at the dataset we can see that the target variable is imbalanced and this is the reason for low f1 score. Hence we will use SMOTE technique to balance the dataset and now the new dataset will contain equal amount of values in target variable and it would also increase the number of rows.

After applying Smote we would again train the data with different classification model and now we find SVC is providing highest accuracy of 87% with a good f1 score as well. So after applying smote we have got positive changes in the models output.

We would now perform Cross Validation for this data to check if the accuracy is same for every part of data taking it in folds. This would help us check if the model is working very accurately taking any rows in random. We find that there is no much difference between both accuracies and hence the model is working fine.

We will now perform Grid search CV with hyper parameter tuning so that it can enhance the model and can help the model work more accurately. Since we have got good accuracy using SVM we will use certain parameters and tune the model so that it works at the best accuracy. We can see that after tuning the model its working at 88% accuracy.

We may further use ensemble method which would further enhance the model but it may over fit the model which can lead to wrong predictions so we have not been using ensemble method here.

Concluding Remarks:

We would now plot the AUC ROC curve for the model and we can see that we are getting 95% and this depicts that the model is able to differentiate between fraud and legit at 95% accuracy which is good sign.

The model is able to differentiate between false positive and true positive at a good rate and hence we can conclude that the model is 95% accurate in helping insurance companies classifying the fraud.

This can help insurance companies save their money against these fraud claims and help them build a robust detection model.