Module End (Machine Learning)

Prn no: 230340325014

Name: Harsh Shinde

## BANK FRAUD DETECTION

The given dataset is about bank fraud detection which has 6.3 million rows and 10 columns

Load the libraries and data

Visualizing the Data

As there are common steps in visualizing the data, we apply them through function.

Check for missing values then

There is no null values

## Find types of transactions and number of unique transactions

So the second step would be to check the unique data in step and type because type is a non numeric data so if there aren’t any much unique data we can decide to omit the column

But cash out and payment have higher data sets so we need to keep the column

Cash-out and payment has higher data points. Debit has lowest.

* Cash in: to exchange money for some insurance policy
* Cash Out: To convert non-cash asset into Cash.

To check the collinearity we can apply the correlation to the dataset

If the correlation is 1 it is said to be the perfect collinear

Most of the datapoints are not correlated.

Oldbalance is correlated with new\_balance of person who intitiated transaction.newbalanceDest and oldbalanceDest is correlated.

It is obivous to correlate as substraction and addition of the amount happens and thus these columns are correlated.  
Thus I don’t delete any of the columns as there is no explicit correlation between unrelated variables.

Then skewness of the data is to be checked

We observe that data is largely positively skewed and does not have normal distribution.

After some exploration of feature engineering techniques and methods to deal with skewed data log and boxcox transformation is employed.

For right-skewed data — tail is on the right, positive skew

— common transformations include square root, cube root, and log.

# Variable Transformation

Apply log and boxcox transformation and choose relevant transformations for the respective columns.

‘amount’, ‘oldbalanceOrg’, ‘newbalanceOrig’, ‘oldbalanceDest’, ‘newbalanceDest’ are chosen as it has large skewness. Create seperate set called ‘skew\_cols’ for applying skewness.

0–1 value of skewness containing columns are not chosen.

After observing the skewness after transformation,

Choose boxcox transformation for amount, oldbalanceDest, newbalanceDest

Choose log transformation for oldbalanceOrg, newbalanceOrg

Concat the chosen columns to form the transformed dataframe.

Now the data is transformed

When we view the data for fraud detection, we observe that customer to customer transactions has more frauds than other relationships

So we cant omit the columns, instead we need to preserve the columns and rename the the customer id as ‘C’ and ‘M’

Then we will Explore the type of transactions causing frauds.

It is observed that cash-out and transfer modes have more fraud chances.

Onehot encoding is applied to ‘nameOrig’ and ‘nameDest’ which contains ‘C’ and ‘M’ values. ‘type’ is also encoded.

Now our dataset is processed so we can now start with the modelling process

Now I have worked on 4 models for this dataset

1.Random forest

2.Logistic Regression

3.KNN algorithm

4.Linear Regression

There was no point linear regression because plotting the plot or the corr heatmap we realize that the dataset has no linear regression but it is one of the simplest model to build so why not built it and see what is our accuracy

So now to train the model we need to split our data into training data and testing data

We observe the class distribution in y\_train and y\_test, training and test set has imbalanced classes.

Imbalance in training set will result in poor model.

Thus sampling of training set is employed.

# Sampling

SMOTE is used to oversample the minority class-here ‘Fraud’(1) class

Minority class is sampled with 0.025 times the majority class

Not-Fraud class (0) is undersampled to double the number of ‘Fraud’ class

# Fit the Model

RandomForestClassifier and Logistic Regression is chosen for modelling.

Hyperparameters are chosen after experimentation and prior experience

Then model is saved.

Then lets check the accuracy of the model which is 99.6%

Now to visualize the results we can plot confusion matrix and display it and print the values as well by that we get to know is our model predicting correct result or not

Via confusion matrix I got to know that 22 cases which come in fraud category has been mispredicted as non-fraud

Then we plot the roc curve for the model, to know the performance of the classification model and to visualize the true positive rate and true negative rate

Even though logistic Regression has given higher accuracy, RandomForestClassifier is more reliable.

As false negatives is lesser in RandomForest, ie, ‘Fraud’ detected as ‘non-Fraud’ is very less.

-RandomForestClassifier has given less number of False Negative -22 where as logistic regression has given 1049 False Negatives. Thus RF is more reliable.