Data Science Project On

Customer Transaction

PROJECT-ID = PRCP-1003-CustTranPred

Business case:- Based on the given dataset we have to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

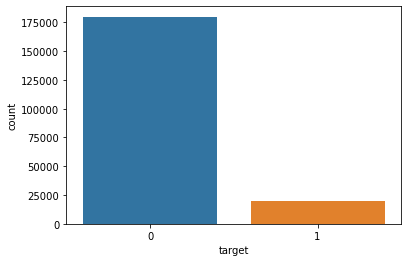
Introduction:

The dataset contains 200000 observations with 202 columns with 200 columns having values for var\_1 to var\_200, one column for ID code and one column for target with the problem of identification of the customers who will make a transaction with the bank in future, irrespective of the amount of money transacted previously with the bank.

Domain Analysis & EDA:

Target column contain(0,1)

* 0 represent -- customer did not do a transaction
* 1 represent -- customer did do a transaction



* In this plot we are clearly seen the 90% customer are did not do a transaction and 10% customer are do the transaction
* This target feature is imbalance so we need to balance the data with the help of oversampling.
* After checking all plotting of columns we are clearly see that most of the feature is follow a normal distribution and some feature are very close to the normal distribution so we not need to use feature transformation technique.

Data preprocessing & Feature Engineering:

**1.Missing value:**

The dataset doesn’t contain missing value.

**2.Categorical data:**

The dataset doesn’t contain categorical value.

**3.Outlier Handling:**

In this data outlier is present in all feature, but we not handle outlier because the from above EDA part we understand the all feature is very close to the normal distribution and distance between two point is very less so we use robust scaling.

**4.Scaling:**

* Here we use robust scaling because the dataset contains outliers but follow the normal distribution
* Robust scaling makes the data look similar to standardization in term of range, it brings the median to 0 and scales based on the interquartile range instead of the standard deviation.

Feature Selection:

**1.Drop unique & Constant column:**

The dataset contains only one constant column.

**2.Checking correlation:**

* In this data no high correlated feature is present

**3.Check duplicates:**

* There are no duplicates present in data

4.**Priciple Component Analysis:**

* In PCA we are select 175 feature because the less variance loss

Model Creation and Selection:

* Logistic regression model and Decision tree score is not good.
* XGB Classifier and MLP model works very well with testing accuracy up to 96.25 and f1\_score is 96.29
* From the above we prefer the MLP model for better result of our model.

**NOTE**:

Decision tree training take too much time. After hyperparameter tunning the model accuracy also not increases and take much time so I removed the hyperparameter tunning. Smote is not working properly in jupyter notebook so I run my code on google Collaboratory where everything works clearly.