**CUSTOMER TRANSACTION PREDICTION**

PROJECT-ID-PRCP-1003-CustTranPred

Business use case: Using a dataset with a specific feature, we must forecast which client will make a transaction in the feature regardless of the amount of money transacted.

Task: Binary Classification Task

Project Brief

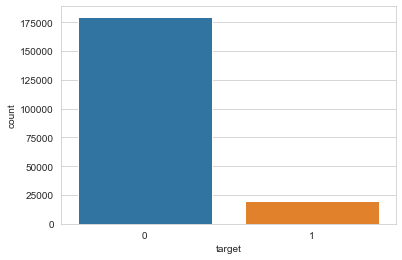
The dataset contains 200000 observations with 202 columns with 200 columns having values for var 1 to var 200, one column for ID code, and one column for target to address the problem of identifying customers who will make a transaction with the bank in the future, regardless of the amount of money transacted previously with the bank.

Exploratory Data Analysis

1. MISSING VALUE:

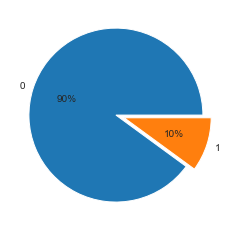
No Missing Values Present in the Dataset

1. Target Variable Distribution



From The figure above we can get to know that there is imbalance in data

Where 90% of data has value 0 and 10% of data has value 1



1. All the Variables are distributed Normally

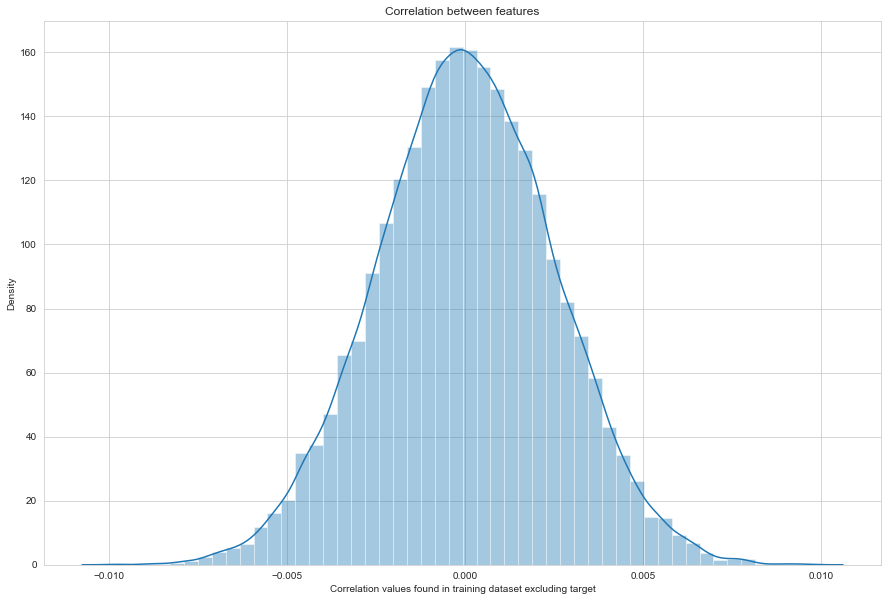




We can see from the above figures that all the variables are distributed normally.

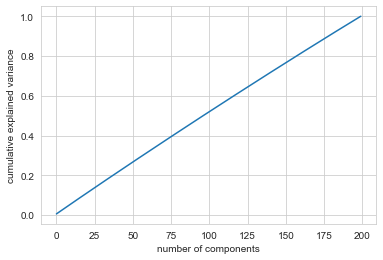
1. Correlation between variables

The max correlation is 0.0667 and min is -0.0809. Let's dive deeper into the correlations



Important thing to note is that these correlations are all very small and the correlations also follow normal distribution. This is for the correlations between the 200 features and the target which is mostly between +0.05 and -0.05. For the correlations between the features themselves they are also weak being mostly between +0.005 and -0.005. There are NO strong correlation hence, feature reduction either by combining features or dropping features will be difficult.

Let’s apply PCA and check the important variables



All 100 variables needed to explain 100% of variance, hence the data must be PCA of some other larger dataset and we are more sure that feature reduction would not be a good approach

MODEL CREATION AND SELECTION:

OBSERVATION:

* logistic regression model gives the AUC score of 0.7685
* Naïve Bayes model gives the AUC score of 0.8046
* Now since the target variable is not balanced, we will balance the dataset using down sampling and try to apply various algorithms
* Post down sampling below are the results
* logistic regression model gives the AUC score of 0.7723
* Naive Bayes model gives the AUC score of 0.8124
* We can see that down sampling slightly improves our metric
* Random Forest gives the AUC score of 0.7786
* XGB Classifier model gives the AUC score 0.7973
* From above all model, we are select Naive Bayes because it has a better AUC score.