DATA SCIENCE PROJECT ON CUSTOMER TRANSACTION

PROJECT-ID = PRCP-1003-CustTranPred

BUISNESS CASE: BASED ON GIVEN FEATURE OF DATASET WE NEED TO PREDICT WHICH CUSTOMER MAKE TRANSACTION IN THE FEATURE IRRESPTIVE OF THE AMOUNT OF MONEY TRANSACTED

TASK: BINARY CLASSIFICATION TASK

INTRODUCTION

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, the dataset contain 200000 observation with 202 columns with 200 columns having values for var\_1 to var\_200, one column for ID code and one column for target

NOTE:

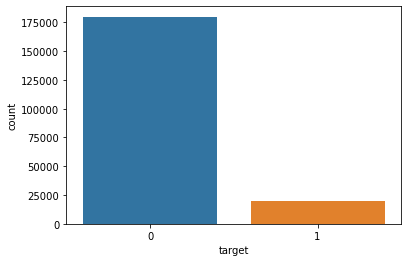
* In this data we not do any domain analysis and eda because the data is contain private information of bank customer
* In this data we only check the distribution of feature and target column

DOMAIN ANALYSIS and eda:

TARGET COLUMN == TARGET [0, 1]

\* 0 Represent -- CUSTOMER DID NOT DO A TRANSACTION

\* 1 Represent -- CUSTOMER DID DO THE TRANSACTION



OBSERVATION:

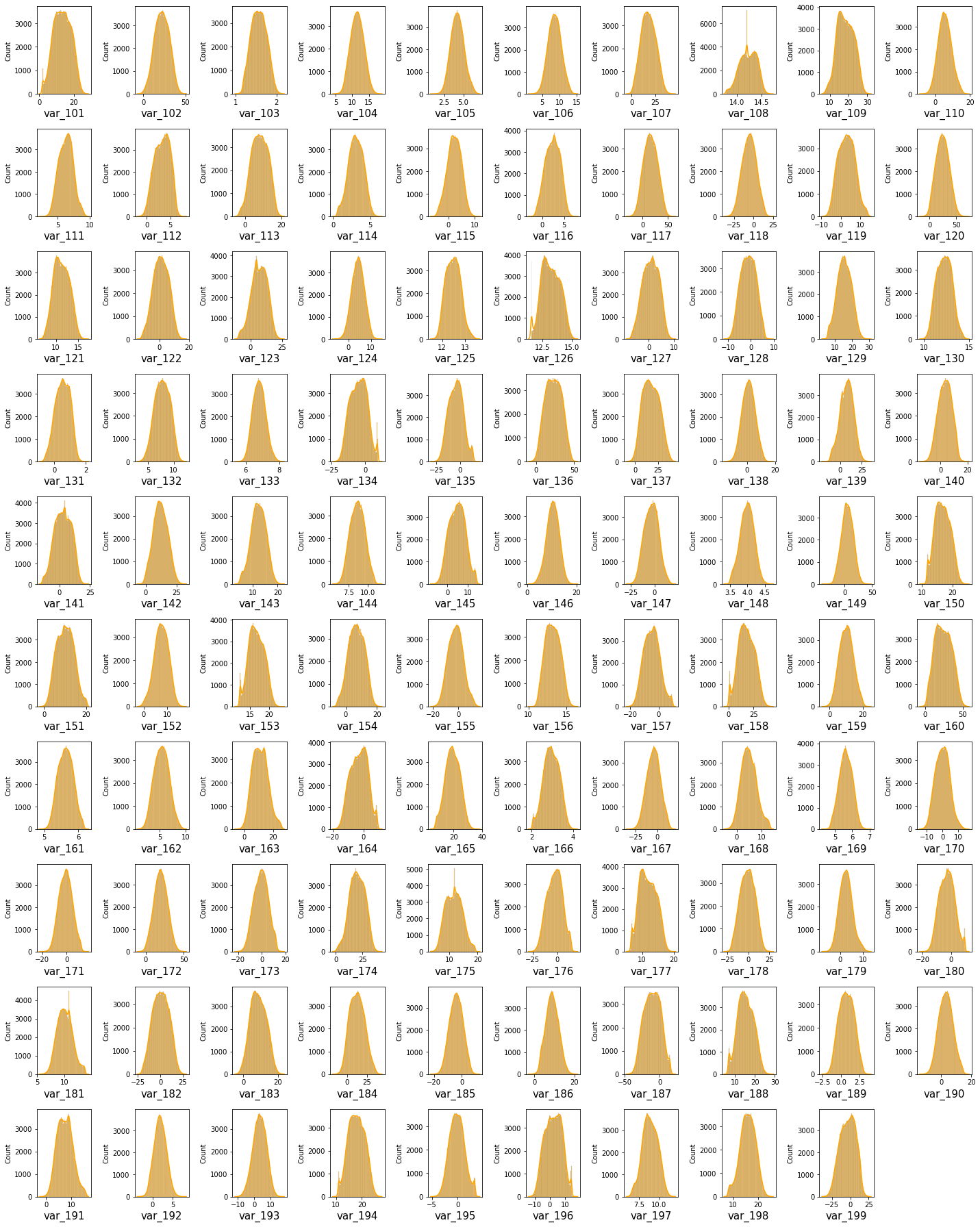
\* In this plot we are clearly seen the 90% customer are did not do a transaction and 10% customer are did do the transaction

\* This target feature is imbalance so we need to balance the data with the help of oversampling.

1. CHECKING DISTRIBUTION OF FIRST 100 FEATURE:



1. CHECKING DISTRIBUTION OF REMAINING FEATURE:

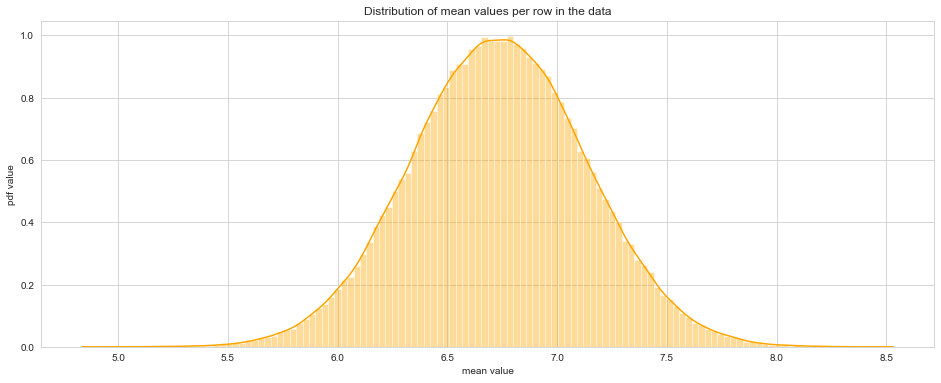


OBSERVATION:

\* From above both plots 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.

STATISTICAL ANALYSIS OF DATA:

1. DISTRIBUTION OF MEAN AND STANDARD DEVIATION OF THE DATA:

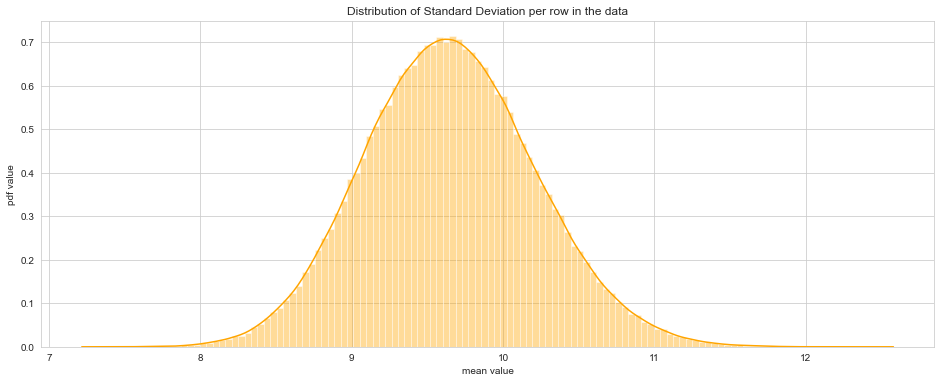


OBSERVATION:

\* From the above graph looks like guassian distribution with mean value 6.73

\* we can say that around 80% feature mean lies between 6.5 to 7.0

1. DISTRIBUTION OF STANDARD DEVIATION OF DATA:



OBSERVATION:

\* This graph also look like guassian distribution around 60% of feature standard deviation around the range of 9 to 10.

DATA PREPROCESSING AND FEATURE ENGINNERING:

1. MISSING VALUE:

* In this data no missing value is present.

1. CATEGORICAL DATA:

* This data is not contain any categorical feature.

1. 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.

1. FEATURE SCALING:

* Here we are use Robust scalar to scale the feature between IQR ,

Because the all feature contain outlier and close to the normal distribution

FEATURE SELECTION:

1. DROP UNIQUE AND CONSTANT COLUMN:

* In this data only one unique column is present

1. CHECK THE CORRELATION:

* In this data no highly correlated feature is present

1. CHECK DUPLICATES:

* There is no duplicates present in data

1. PRINCIPLE COMPONENT ANALYSIS:

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

MODEL CREATION AND SELECTION:

OBSERVATION:

* logistic regression model training and testing score is not good after applying bagging score is still lagging.
* ANN MLP Classifeir model is very well work on training data as well as testing data the score of training is 93.87 and testing score is 90.47 with f1 score is 90.58
* XGB Classifier model is also well work on training and testing side the score of training 93.77 and testing score is 90.60 with f1 score is 90.67.
* From above all model we are select ANN MLP Classifier because we use certain parameter like max\_iter.

NOTE:

FROM ABOVE XGBOOST AND MLP CLASSIFIER MODEL IM USE HYPERPARAMETER TUNNING BUT THE DATA IS VERY HUGE IT TAKES MORE TIME TO FIT THE DATA ALSO IM TRYING ON GOOGLE COLAB AS WELL BUT NOT SOLVE THE TIMINING PROBLEM THAT WHY I'M REMOVE THE ALL HYPERPARAMETER TUNNING.