

DATA SCIENCE PROJECT ON CUSTOMER TRANSACTION PREDICTION -PRCP-1003

PROJECT-ID = PTID-CDS-DEC-25-3471

BUSINESS CASE: Based on given feature of dataset we need to predict which customers will make a specific transaction in the future, irrespective 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.

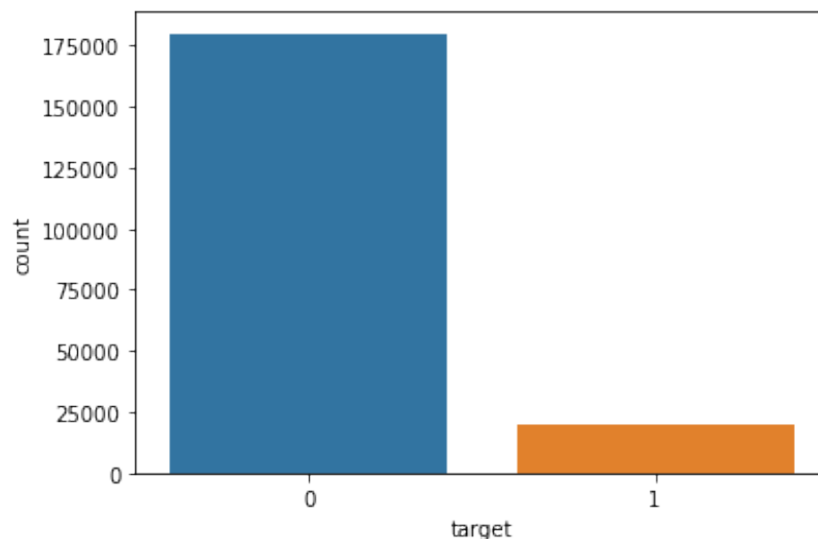
NOTE:

- In this data we not do any domain analysis and eda because the data contains 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]

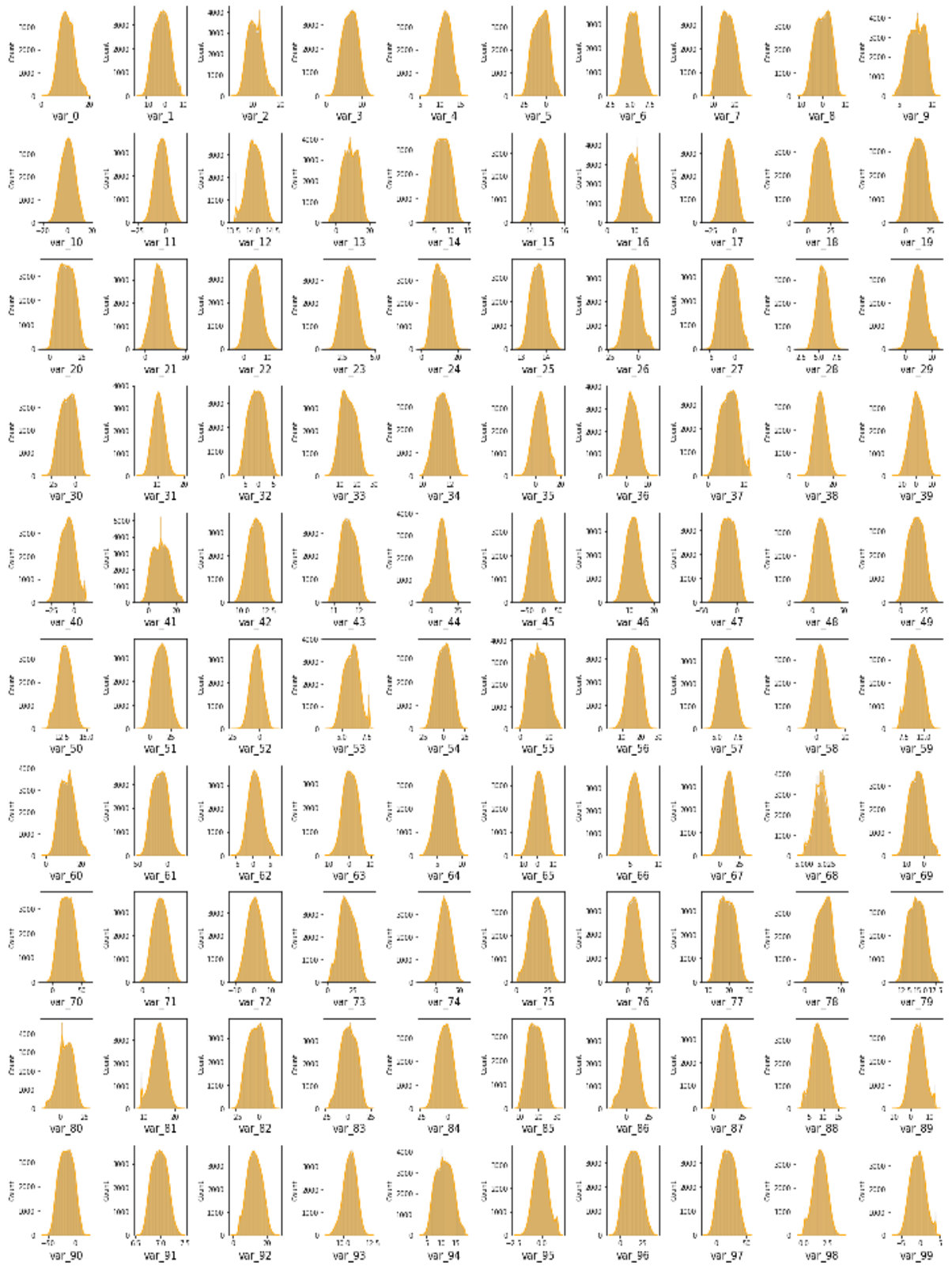
- 0 Represent -- CUSTOMER DID NOT DO A TRANSACTION
- 1 Represent -- CUSTOMER DID DO THE TRANSACTION



OBSERVATION:

- In this plot we are clearly seen that 90% customer did not do a transaction and 10% customer did do the transaction
- This target feature is imbalanced so we need to balance the data with the help of oversampling.

1. CHECKING DISTRIBUTION OF FIRST 100 FEATURE:



2. CHECKING DISTRIBUTION OF REMAINING FEATURE:

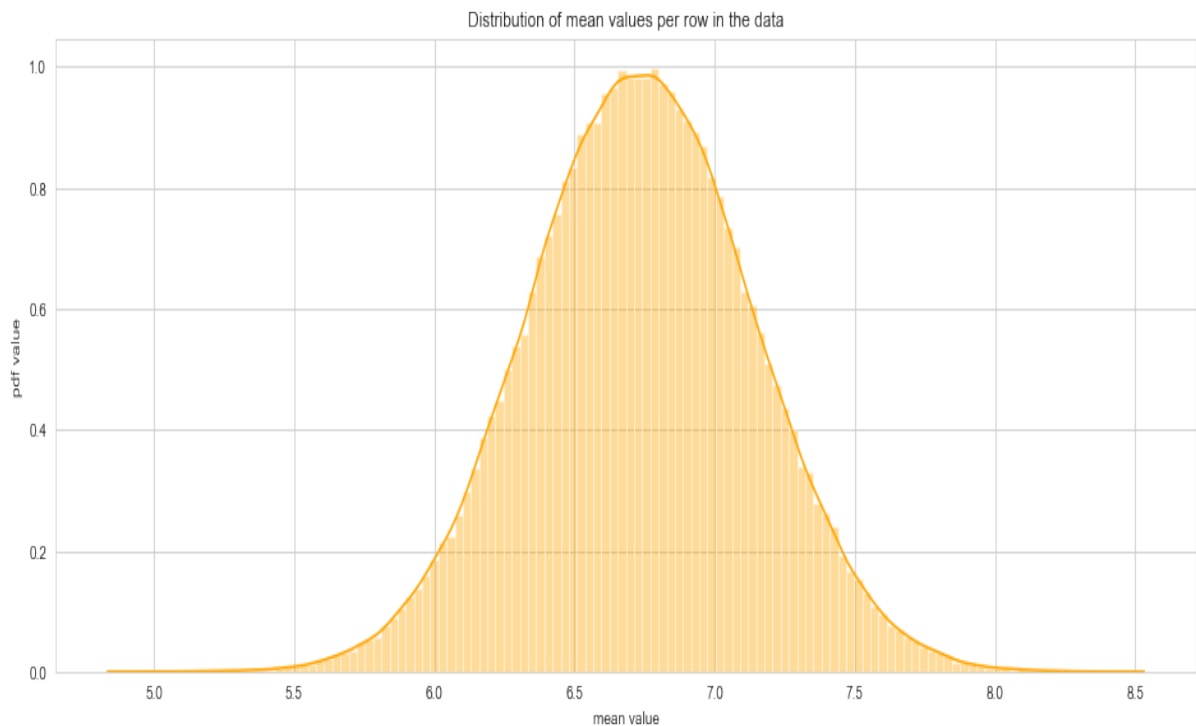


OBSERVATION:

- From above both plots we can clearly see that most of the feature follows a normal distribution and some features are very close to the normal distribution so we do not need to use feature transformation techniques.

STATISTICAL ANALYSIS OF DATA:

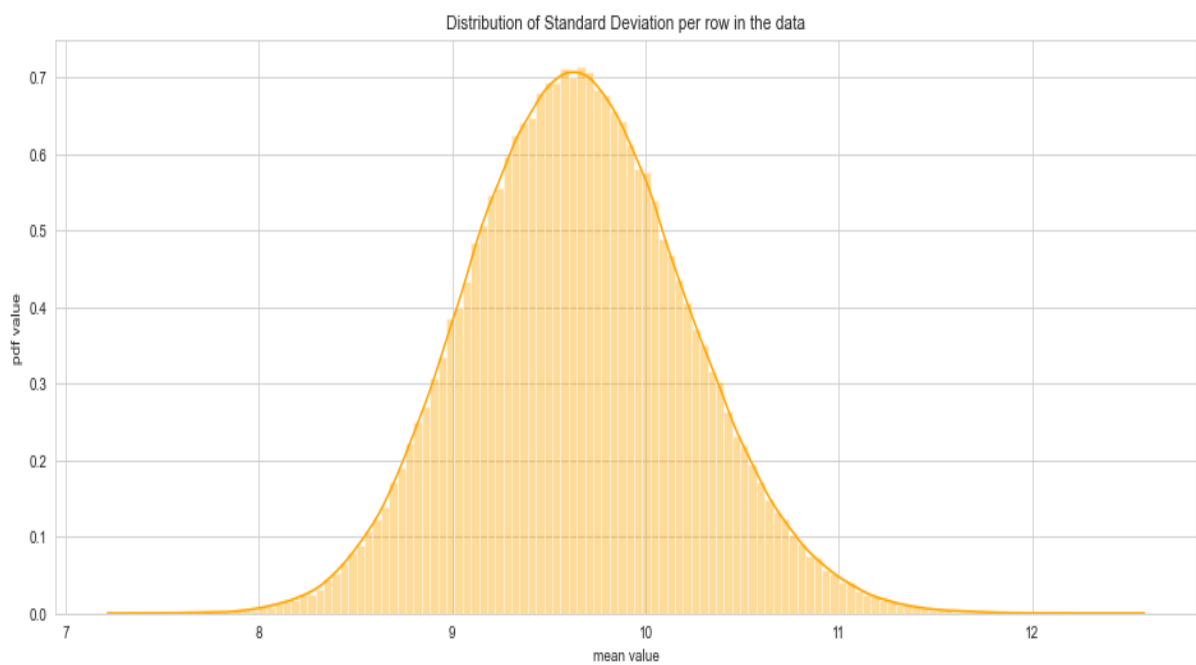
1. DISTRIBUTION OF MEAN AND STANDARD DEVIATION OF THE DATA:



OBSERVATION:

- From the above graph looks like gaussian distribution with mean value 6.73
- From the above graph we can say that around 80% feature mean lies between 6.5 to 7.0

2. 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 ENGINEERING:

1. MISSING VALUE:

In this data no missing value is present.

2. CATEGORICAL DATA:

This data does not contain any categorical feature.

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 that all features are very close to the normal distribution and distance between two point is very less so we use robust scaling.

4. 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

2. CHECK THE CORRELATION:

In this data no highly correlated feature is present

3. CHECK DUPLICATES:

There is no duplicates present in data

4. PRINCIPLE COMPONENT ANALYSIS:

In PCA we are select 175 feature because of less variance loss

MODEL CREATION AND SELECTION:

OBSERVATION:

- logistic regression model training and testing score is not good, even after applying bagging score is still lagging.
- ANN MLP Classifier model is very well working on training data as well as testing data, the score of training is 93.74 and testing score is 90.17 with f1 score is 90.26
- XGB Classifier model is also working well on training and testing side, the score of training 93.69 and testing score is 90.59 with f1 score is 90.66.
- From above all model we select ANN MLP Classifier because we use certain parameter like max_iter.

NOTE:

Since the dataset is very large, performing hyperparameter tuning on the XGBoost and MLP classifier models was extremely time-consuming. Even running the models on Jupiter did not solve the timing issue. Therefore, I decided to remove all hyperparameter tuning from the process.