Statistics Report

## Analysis of Campaign for selling personal loans

## By Sivasankar

## Problem Definition

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget. The department wants to build a model that will help them identify the potential customers who have higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. The file Bank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Abstract

The main goal of this report is

1. Read the column description and ensure you understand each attribute well
2. Study the data distribution in each attribute, share your findings.
3. Get the target column distribution. Your comments
4. Split the data into training and test set in the ratio of 70:30 respectively
5. Use different classification models (Logistic, K-NN and Naïve Bayes) to predict the likelihood of a liability customer buying personal loans
6. Print the confusion matrix for all the above models
7. Give your reasoning on which is the best model in this case and why it performs better?

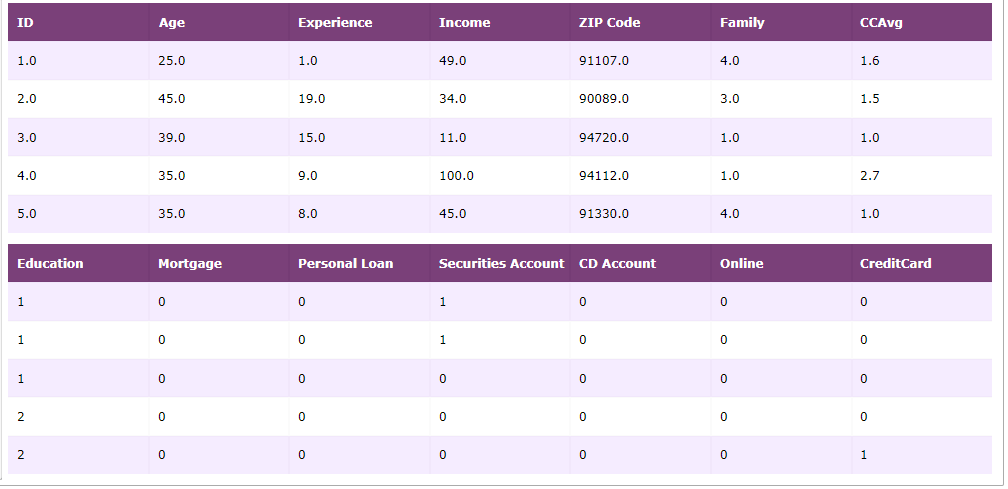
Methodology

1. Description of data
2. Preprocess data
3. Visualize data
4. Build a model
5. Check Assumptions
6. Compare different classification models (Logistic, K-NN and Naïve Bayes)

Description of data

1. Name of the data: Bank Personal Loan Modeling
2. Number of data points: 5000
3. Number of features: 14
4. Target attribute: Personal Loan

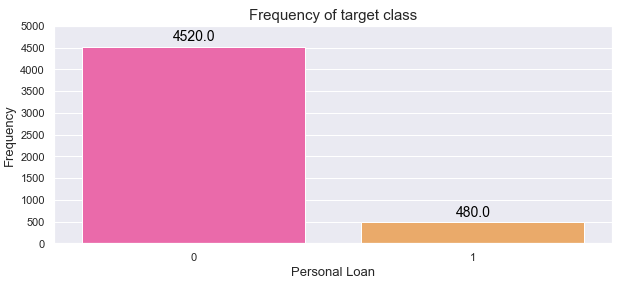
Data



Features

1. ID - Customer ID
2. Age- Customer's age in completed years
3. Experience - #years of professional experience
4. Income - Annual income of the customer ($000)
5. ZIPCode - Home Address ZIP code.
6. Family - Family size of the customer
7. CCAvg - Avg. spending on credit cards per month ($000)
8. Education - Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
9. Mortgage - Value of house mortgage if any. ($000)
10. Personal Loan - Did this customer accept the personal loan offered in the last campaign?
11. Securities - Account Does the customer have a securities account with the bank?
12. CD Account - Does the customer have a certificate of deposit (CD) account with the bank?
13. Online - Does the customer use internet banking facilities?
14. Credit Card - Does the customer uses a credit card

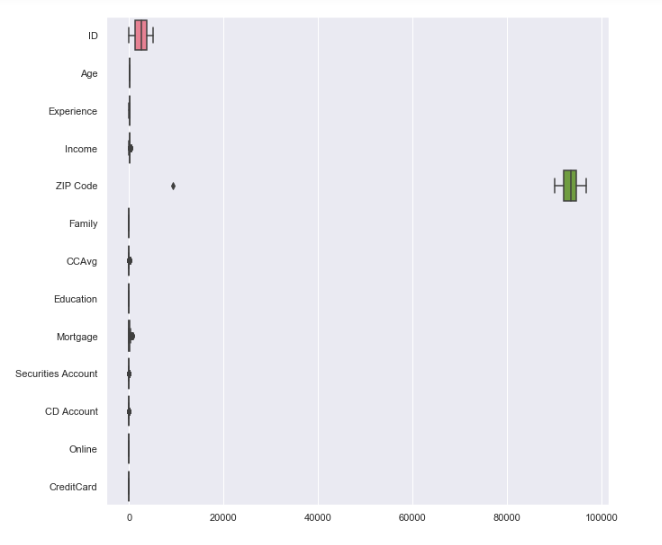
Distribution of Target Attribute



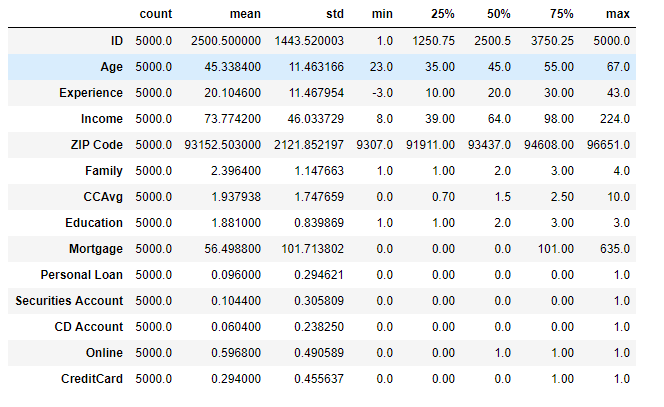
Analysis

The number of data points for customers accept the personal loan offered are very low when compared to customers not accept the personal loan offered. This may affect the model.

Ditribution of featutres



Stats of data features

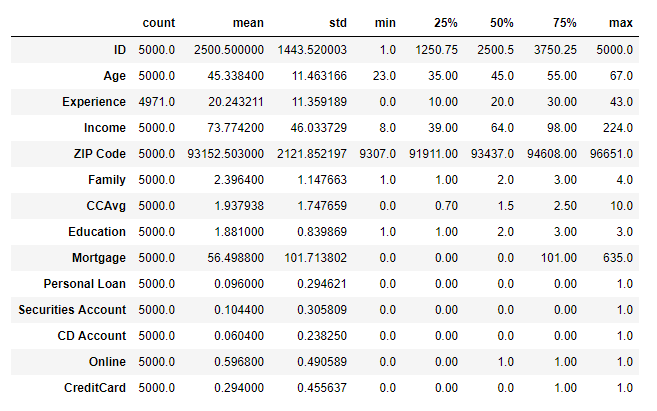


Analysis

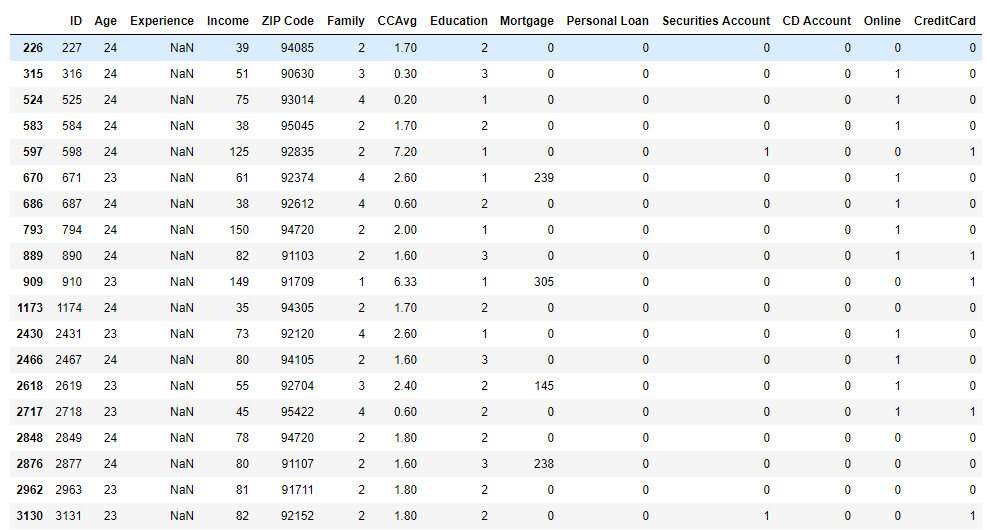
It can be seen that the "Experience" variable has negative values, which could possibly mean some error in data. Hence we need to clean that data. Also the mean of Experience is equals to median.

The feature ID does not add any interesting information. There is no association between a person's customer ID and loan, also it does not provide any general conclusion for future potential loan customers. We can neglect this information for our model prediction.

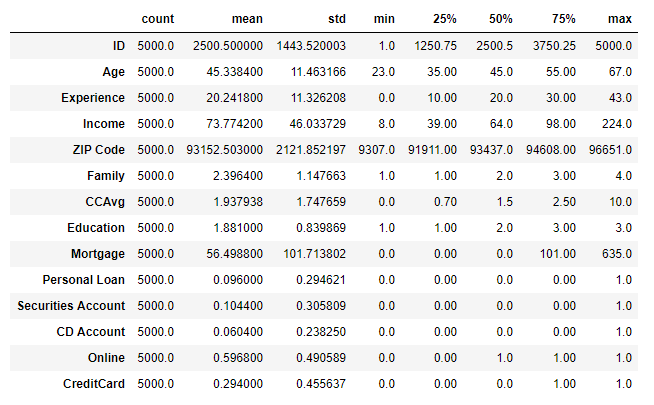
Data Description after cleanup



correction of NAN Value

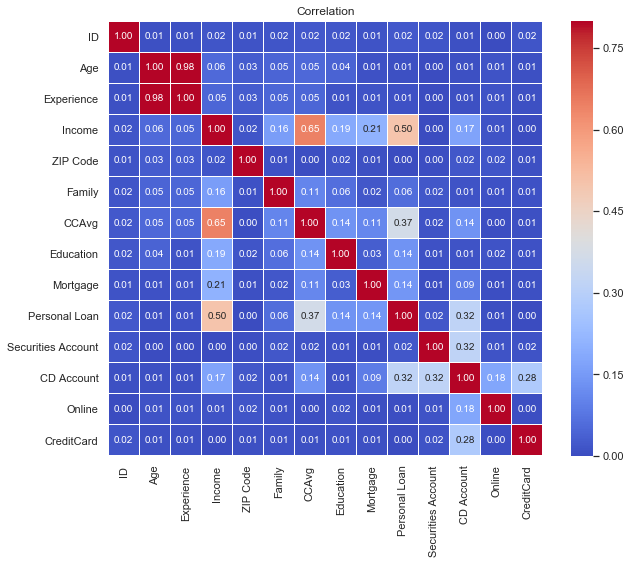


Data Describe after corrections



Visualize data

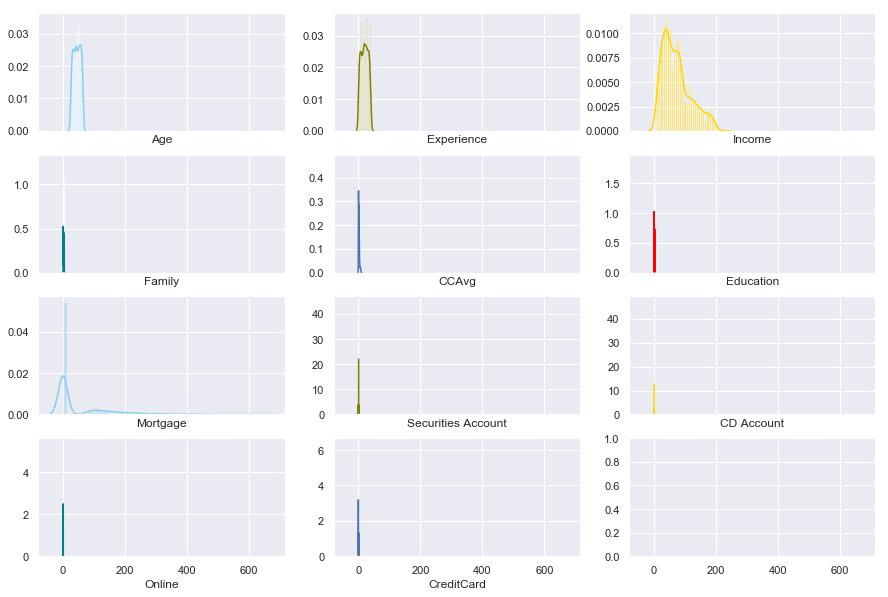
Correlation between features

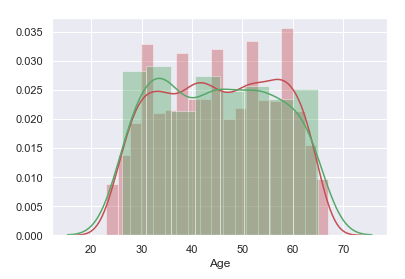


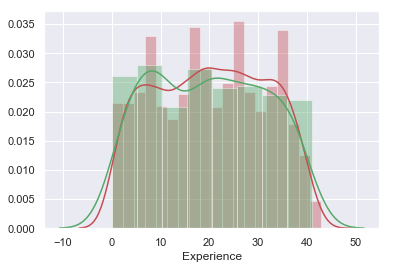
From the above correlation chart, we can observe that

* Correlation between Age and Experience is 0.98.
* Correlation between Income and CCAvg is 0.65.
* These are the pairs of features having high correlation i.e (>0.5)
* Any ways we will not consider ID column.

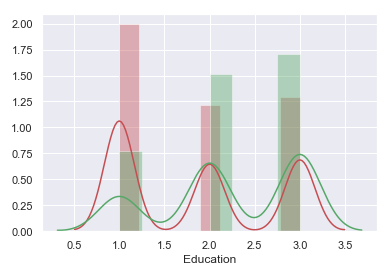
Distribution of each feature

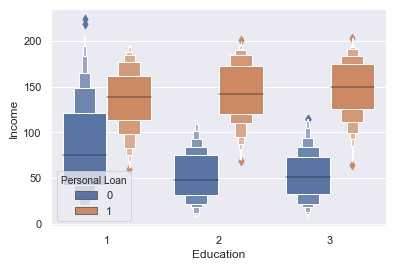




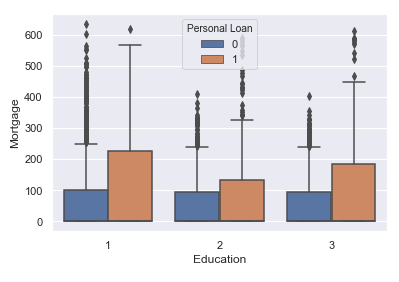




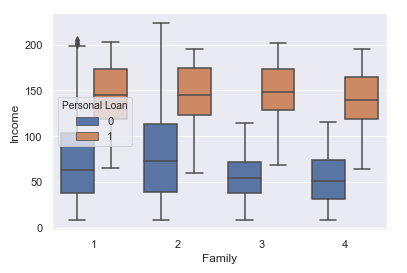


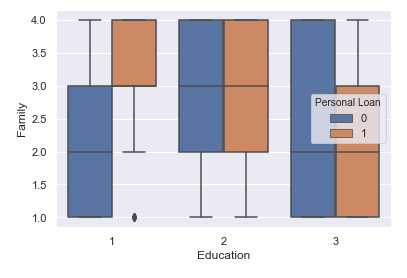


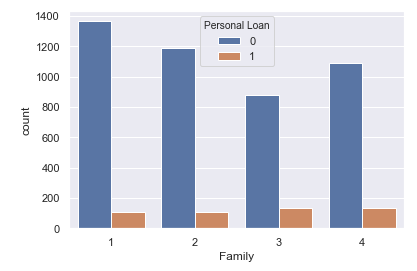
As per above plot we can say that customers whose Education level 1 has more income levels, however customers who taken the Personal Loan almost has same income levels.

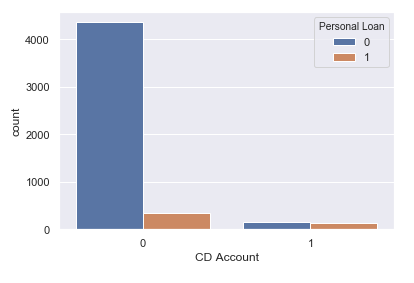


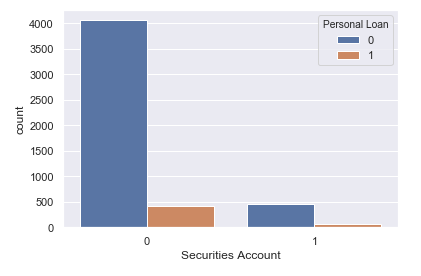
As per above plot we can say that customers who have Personal loans and the who does not opted for personal loans have high mortagae

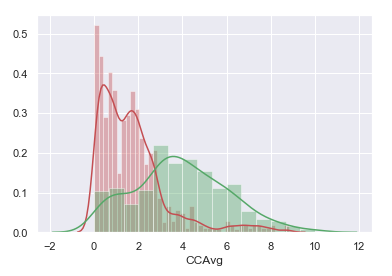




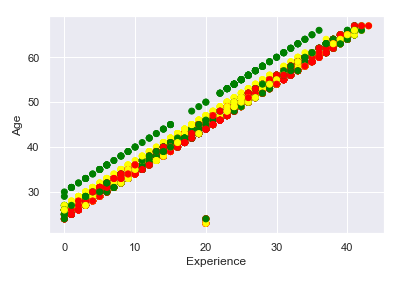








Customers who are opted for personal loans are having high Cerdit card Avg spends per month.



Above plot shows that experinece and age are having positive correlation, as increase age experinece also incresed.

Pair plot between features

- This is to understand the relation between features.

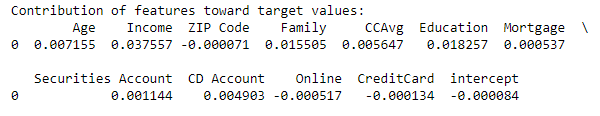


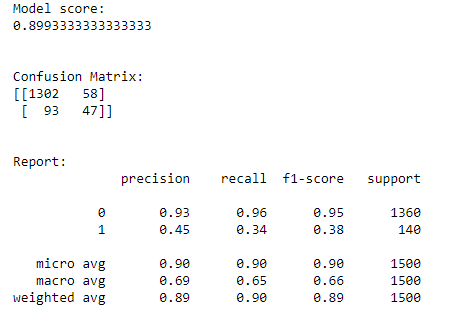
* Age feature is normally distributed with majority of customers falling between 30 years and 60 years of age. We can confirm this by looking at the describe statement above, which shows mean is almost equal to median
* Experience is normally distributed with more customer having experience starting from 8 years. Here the mean is equal to median.
* Income is positively skewed. Majority of the customers have income between 45K and 55K. We can confirm this by saying the mean is greater than the median
* CCAvg is also a positively skewed variable and average spending is between 0K to 10K and majority spends less than 2.5K
* Mortgage 70% of the individuals have a mortgage of less than 40K. However, the max value is 635K
* The variables family and education are ordinal variables. The distribution of families is evenly distributing.

Build models

Logistic Regression:

* The method of Logistic Regression that finds the coefficients of different
* how independent variables are contributing to the dependent variable.
* The below report shows the coefficients of features(contribution).

Model Score and Confusion Matrix



Model Accuracy is 0.899 = 0.90

True Positive = 47

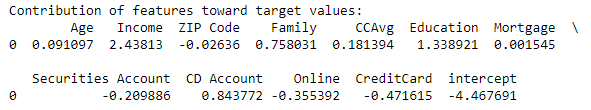
True Negative = 1302

False Positive = 58

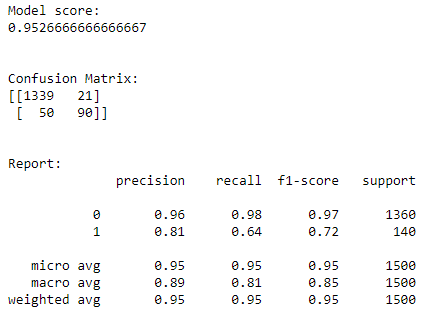
False Negative = 93

Lets Apply scaled data to the Logestic training model

* The below report shows the coefficients of features(contribution) on scaled data.



Model Score and Confusion Matrix on scaled data



Model Accuracy is 0.95

True Positive = 90

True Negative = 1339

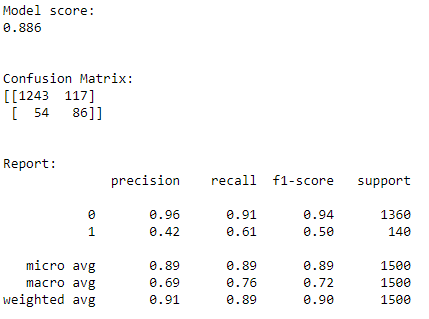
False Positive = 21

False Negative = 50

Model accuracy increased with scaled data compared to original data.

Naive Bayes Model

Model Score and Confusion Matrix



Model accuracy is 0.886

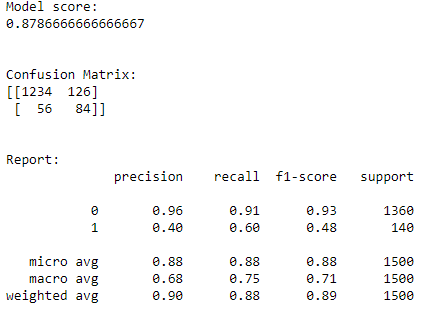
True Positives = 86

True Negatives = 1243

False Positives= 117

False Negatives= 54

Model Score and Confusion Matrix on scaled data



Model accuracy decreases to 0.878

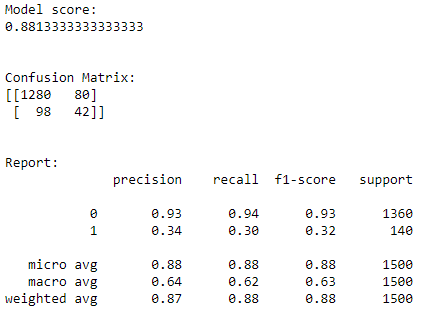
True Positives = 84

True Negatives = 1234

False Positives= 126

False Negatives= 56

k-Nearest Neighbors (KNN)

Model Score and Confusion Matrix

Model accuracy is 0.881

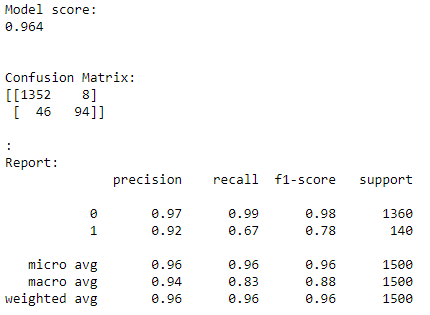
True Positives = 42

True Negatives = 1280

False Positives= 80

False Negatives= 98

Model Score and Confusion Matrix on scaled data



Model accuracy increases to 0.964

True Positives = 94

True Negatives = 1352

False Positives= 8

False Negatives= 46

Model Comparision

Overall model accurecy before scaling:

Logistic Regression: 0.8993333333333333

Naive Bayes: 0.886

K-Nearest Neighbors: 0.8813333333333333

Overall Model Accuracy After scaling:

Logistic Regression: 0.9526666666666667

Naive Bayes: 0.8786666666666667

K-Nearest Neighbors: 0.964

Conclusion

* We have visualized dataset in all possible ways and they are

shown in the form of plots.

* Logistic Regression, Naive Bayes and KNN models are built to predict the target variable.
* Some improvements have been done on the model by removing some features that are not contributing.
* The tests of assumptions for Logistic Regression, Naive Bayes and KNN are also checked and they are analyzed properly.
* In the end, all model compared
* We can see clearly K-Nearest Neighbors model with scaled data gives better accuracy which is 96.4%, hence K-Nearest Neighbors is best choice to predict the customers who will accept the personal loan.