Bank Personal Loan Modelling

INTRODUCTION:

In recent years, the number of people applying for loans has risen for a variety of reasons. Bank employees are unable to assess or foresee whether a borrower will be able to repay the loan (good or poor customer) at the specified interest rate. The classification aim is to forecast the probability of a liability consumer purchasing personal loans, we would develop a model that will be used to determine which customers are most likely to approve a personal loan deal based on their relationship with the bank as well as other features in the dataset. We will be using various methods to predict which model, out of Logistic Regression, and Random Forest Algorithm, is the best for this problem. (Choudhary, 2019; Francis Jency, 2018)

DESCRIPTION OF DATASET:

The dataset contains data on 5000 customers and 14 attributes. Customers' demographic details (age, income, etc.), their arrangement with the bank (mortgage, securities account, etc.), and their reaction to the previous personal loan initiative are all included in the results (Personal Loan).

Target Variable is **Personal Loan** which describe whether the person has taken loan or not. This is the variable which we need to predict.

Nominal datatypes:

- ID - Customer ID.
- ZIP Code - Home Address ZIP code of the customer.

Ordinal datatypes:

- Family - Number of family member of the customer.
- Education - Education level of the customer. In our dataset it ranges from 1 to 3 which are Undergraduate, Graduate and Postgraduate respectively.
- Age - Age of the customer

Interval datatypes:

- Experience - Professional experience the customer has.
- Income - Annual Income of the customer in dollars (in thousands of $)
- CCAvg - Avg. spending on credit cards per month (in thousands of $)
- Mortgage - Value of House Mortgage

Binary variables (0 or 1):

- CD Account - Does the customer have CD Account with bank or not?
- Security Account - Does the customer have Security Account with bank or not?
- Online - Does the customer have Online banking facility with bank or not?
- Credit Card - Does the customer have a credit card issued by the Bank or not?
- Personal Loan - This our target variable which we have to predict. This indicates that the customer has taken loan or not

DATA CLEANING:

Certain data cleaning methods have been applied on the dataset. They are discussed as follows.

1) **Reordering of the Personal loan column:**

   Personal loan is our predictor variable, and it was in the 5^{th} column. We have moved the column to the last.

2) **Removing the rows with negative experience:**

   Negative values in the data were checked and it was found out that some of the rows in the Experience column had negative values. In reality, experience cannot be negative. So, these rows have been removed.

3) **Checking for Null Values**

   Null values have been checked in the dataset. It was found out that there were no null values in the dataset.

SUMMARY STATISTICS:

To find the mean values of all the independent variables, summary is the best way to get them. Summary function was used on the dataset. It was found out that the mean values of the variables are as follows.

1) Age: Mean age of the customers of the bank was around 45 years

2) Experience: Mean experience of the customers was around 20 years

3) Income: Mean income of the customers was around $ 73.8k

4) CC Avg: Avg credit card spending of the customers was around $1936 per month

5) Security Account: Number of customers having with the account: 516

6) CD Account: Number of customers having with the account: 302

7) Online Account: Number of customers having online banking: 2954

8) Credit Card: Number of customers having credit card with the bank: 1455

EXPLORATORY ANALYSIS:

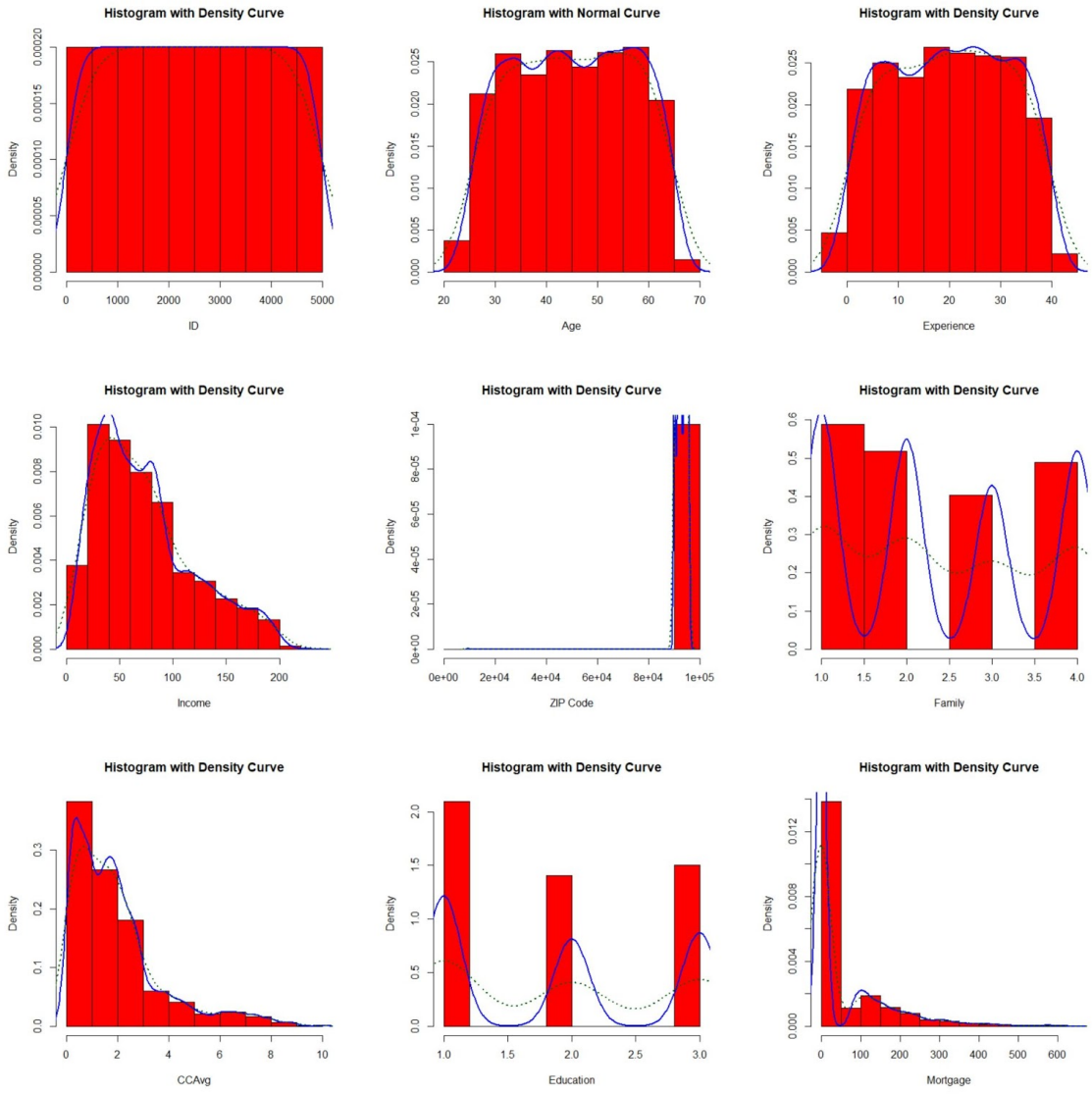


Figure 1
Exploratory Analysis of Dataset.

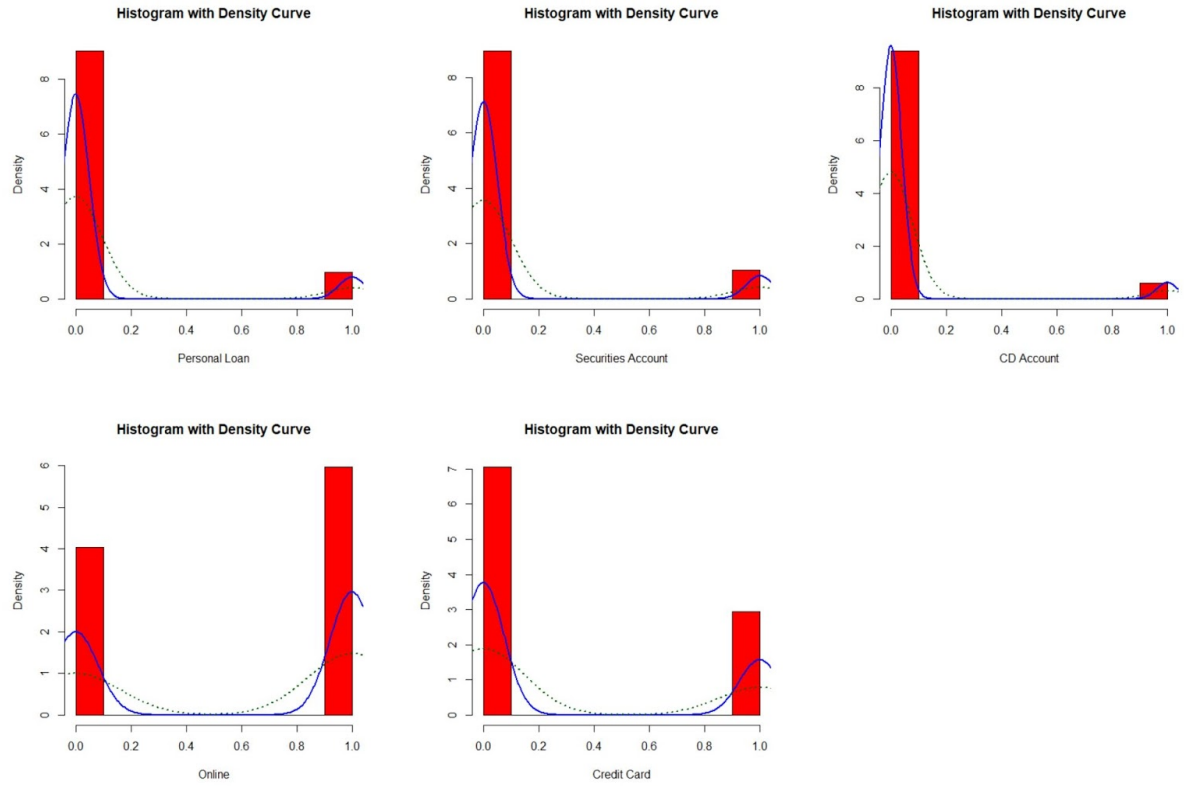


Figure 2
Exploratory Analysis of Dataset

It can be observed from the above Figure 1 and 2, ID is uniformly distributed. Age and Experience is normally distributed. Income distribution is right skewed distribution because the tail goes to the right. Zip is uniformly distributed as datapoint are more with family size 1 and 2. CCAvg is right skewed distribution. In education UG level customers are more than the graduate and advance/professional customers. Also, the mortgage is right skewed distribution. Mortgage is distribution is a right skewed distribution because the tail goes to the right. The online, credit card is Bernoulli Distribution as the number of customers who use these is greater than the number who do not have use them.

STATISTICAL DATA MODELLING:

- By analyzing the dataset, we have decided to use Logistic Regression and Random Forest Algorithm to predict the accuracy of the models we have built.

- The given dataset has been split into train and test datasets. 70% of the dataset has been assigned to training the model and 30% has been assigned to test the models that we have built.

- Normalization of the data has been done. It was necessary because the dataset contained binary variables and numerical variables. Factoring has been done to make the whole dataset uniform and similar.

Logistic Regression: (Bruin, 2006)

When the dependent variable is binary, logistic regression is the best regression analysis to use. The logistic regression, like all regression analyses, is a predictive analysis. To characterize data and illustrate the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is used. Logistic regression model is easier to implement, interpret, and efficient to train.

To begin with, all the variables of the dataset have been taken into consideration and the logistic model was built. The AIC score was 807.99. Assuming the threshold of the p-value as 0.05, the variables that are significantly affecting the dependent variable were identified and final model was built using variables Income, Family, CCAvg, Education, CD Account, and Credit card. AIC for the final model was found out to be 814.77.

The model was built using testing data. The model built is as follows.

```
Call:
glm(formula = Personal.Loan ~ Income + Family + Education + CD.Account +
    CreditCard + CCAvg, family = "binomial", data = bank.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1729  -0.1902  -0.0644  -0.0188   4.1349

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.298904   0.688286 -19.322  < 2e-16 ***
Income        0.064839   0.003781  17.147  < 2e-16 ***
Family2      -0.260498   0.283904  -0.918  0.35885
Family3       2.152007   0.307388   7.001 2.54e-12 ***
Family4       1.810532   0.289412   6.256 3.95e-10 ***
Education2    3.758544   0.327071  11.492  < 2e-16 ***
Education3    3.921747   0.325531  12.047  < 2e-16 ***
CD.Account1   2.913264   0.351404   8.290  < 2e-16 ***
CreditCard1  -0.846918   0.261993  -3.233  0.00123 **
CCAvg         0.178689   0.055442   3.223  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3

The expression for logistic regression was built. The expression that was derived was as follows.

```
> exp(coef(model3))
 (Intercept)      Income     Family2     Family3     Family4   Education2   Education3  CD.Account1  CreditCard1
1.824800e-06 1.066267e+00 8.419411e-01 5.294919e+00 4.219668e+00 6.278295e+01 6.414170e+01 2.888722e+01 3.563988e-01
       CCAvg
1.198611e+00
```

Figure 4

It can be seen from the Figure 4 that, for one unit increase in the value of Income, the dependent variable increases by a factor of 1.066 times, when other independent variables are constant. Similarly, it applies to other independent variables too.

Now that the model is ready, we have used this model to predict the accuracy of the training set as well as test set. The accuracy of the training set was found out to be 95.84%. The accuracy of the test set was found out to be around 95.48%. With this it can be concluded that the accuracy of the Logistic regression model was 95.48%.

ROC graph was also plotted for the model that was built to check if the accuracy we got is correct.

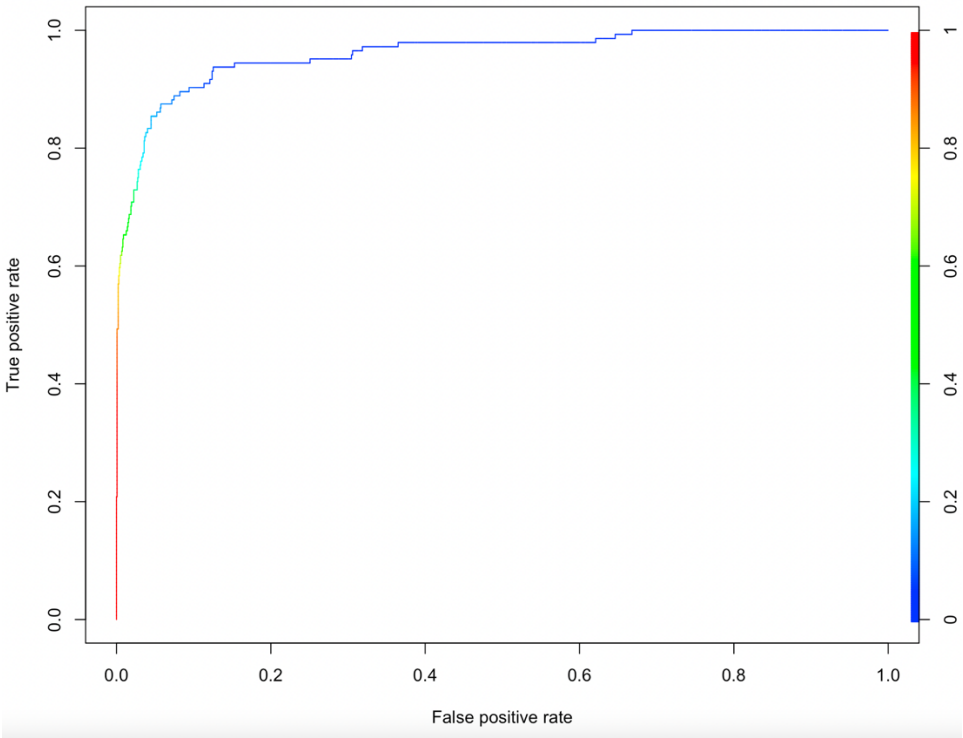


Figure 5

ROC curve for logistic regression

In the above Figure 5, False positive rate is considered as specificity and True positive rate is considered as Sensitivity. Classifiers that give curves closer to the top-left corner indicate a better performance. The bend in the curve is close to the top-left corner in this case. So, the accuracy of the model is also tested using ROC curve and found out to be high.

Random Forest:

Random Forest is the advanced version of Decision Tree. The Decision Tree algorithm is a method for solving both regression and classification problems. The level of understanding of the decision trees algorithm is much easier than the other classification algorithms. Few of the advantages of random forest are the missing values will be handled by the random forest classifier, which will ensure the precision of a significant portion of the data, there will be no over-fitting trees in the model if there are more trees, it can accommodate a broad data set with a higher dimensionality. Apart from these reasons the most important reason for us to choose the model is It provides higher accuracy through cross validation. (Anurag, 2018)

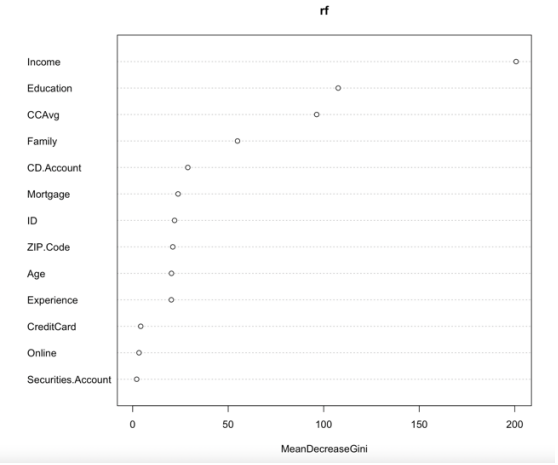
In our dataset binary category have been converted to factor vectors so that the model can understand the all the data is in one format, this is done using a lapply() function. The model was built using the training data. It is shown as follows.

```
Call:
 randomForest(formula = Personal.Loan ~ ., data = bank.train,      ntree = 500, mtry = 6, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6
```

Figure 6

In figure 7, we have used important () function to check the importance of the variables The variables which have the highest importance and influencing the model are: **Income, Education, CCAvg, Family and CD.Account.**

We have taken mtry as 6 which is sq.root(6) that indicates number of variables randomly sampled as candidates at each split. The variables Income and education has high mean decrease gini coefiicient which means that they seem to result in higher purity and contribute to uniformity of the nodes and leaves in random forest.




```
> importance(rf)
                    MeanDecreaseGini
ID                        21.931074
Age                       20.310001
Experience                20.234653
Income                   200.715125
ZIP.Code                  20.992773
Family                    54.870853
CCAvg                     96.319471
Education                107.588799
Mortgage                  23.760922
CD.Account                28.868768
Online                     3.330686
CreditCard                 4.220396
Securities.Account         2.100304
>
```

Figure 7
Importance of the variable and Mean Decrease Gini graph

Accuracy was tested on the test dataset and it was found to be 98.81%. It can be concluded that the accuracy of the Random Forest model is 98.81%.

ROC graph was also plotted for the model that was built to check if the accuracy we got is correct. As, I have mentioned earlier the classifiers that give curves closer to the top-left corner indicate a better performance. The bend in the curve is at the top-left corner in this case. So, the accuracy of the model is also tested using ROC curve and found out to be high.

Conclusion:

- The accuracy of the Logistic regression was found out to be 95.48%
- The accuracy of the Random Forest was found out to be 98.81%
- The most important variable for predicting if the customer accepts the loan are the same in both the models which are Income, Education, CCAvg, Family and CD.Account with an exception of Credit Card in the Logistic Regression.

REFERENCES:

Anurag. (2018, Aug 17). *Random forest analysis in ML*. Retrieved from Newsgen App: https://www.newgenapps.com/blog/random-forest-analysis-in-ml-and-when-to-use-it/

Bruin. (2006, 01 01). *Institute for Digital research and education*. Retrieved from UCLA: https://stats.idre.ucla.edu/r/dae/logit-regression/

Choudhary, P. (2019, Sep 01). *Kaggle*. Retrieved from Kaggle: https://www.kaggle.com/pritech/bank-personal-loan-modelling

Francis Jency, S. (2018, Nov 01). *Exploratory Data Analysis for loan prediction* . Retrieved from https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E2026017519.pdf