B.S. Abdur Rahman Crescent Institute of Science & Technology
Deemed to be University u/s 3 of the UGC Act, 1956

R-PROGRAMMING PROJECT REPORT

Financial Analytics - Bank Loan Modelling

Submitted by

ABISHEK. A – 210292601152

GEETHIKA KISHAN. K – 210292601160

PREETHAM V C G – 210292601168

Under the guidance of

Dr. Dinesh Babu

Designation: Professor

in partial fulfillment for the award of the degree of

MASTER OF BUSINESS ADMINISTRATION

DECEMBER 2022

CRESCENT SCHOOL OF BUSINESS

Acknowledgement

I would like to express my special thanks of gratitude to the dean **Dr. K. Srinivasan** and Programme coordinator **Dr. S. Panboli** for their unlisted encouragement and timely support and guidance till end of the project

A special thanks to our Professor **Dr. Dinesh Babu** for his guidance and suggestions during the project

I also sincerely thank my teammates and classmates from Crescent school of business for their support and guidance till end of the project.

TABLE OF CONTENTS

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Introduction:

Thera Bank has a majority of liability customers. The number of customers who are also borrowers is small. The management of Thera bank is interested in converting its liability customers to personal loan customers. The department wants to build a model that will help them identify potential customers who have a higher probability of purchasing the loan. They have given a dataset containing 5000 observations with fourteen variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, also securities account, CD account, online banking and credit card. The interval category contains five variables: age, experience, income, CC avg and mortgage. The ordinal category includes the variables family and education. The last category is nominal with ID and Zip code.

1.2 About the dataset provided:

Bank loan

A bank loan is when a bank offers to lend money to consumers for a certain time period. As a condition of the bank loan, the borrower will need to pay a certain amount as interest. It is borrowed for a set period within an agreed repayment schedule.

Personal loan

Most banks offer personal loans to their customers and the money can be used for any expense like paying a bill or purchasing a new television. Generally, these loans are unsecured loans. The lender or the bank needs certain documents like proof of assets, proof of income, etc. before approving the personal loan amount. The borrower must have enough assets or income to repay the loan.

The data provided shows if the customers accepted the personal loan offered by the bank in a campaign done by them.

Mortgage

Mortgage refers to the act of providing collateral or security against which a lender may sanction a loan. The property stays collateral until the borrower pays back the full loan amount to the lender.

The data provided says the value of the house mortgage if any.

Credit card average (CCAvg)

The data provided gives information about the average spending on credit cards by the customers per month.

Securities account

Securities account sometimes known as a brokerage account is an account that holds financial assets such as securities on behalf of an investor with a bank, broker or custodian. Investors and traders typically have a securities account with the broker or bank they use to buy and sell securities.

The data provided says if the customer has a security account with the bank or not.

Certificate of deposits account

A certificate of deposit (CD) is a savings account that holds a fixed amount of money for a fixed period of time, such as six months, one year, or five years, and in exchange, the issuing bank pays interest. When one cash in or redeem their CD, they receive the money they originally invested plus any interest.

The data provided shows if the customer has a certificate of deposit account with the bank or not

Online banking

The data provides information on if the customer uses online banking or not.

Credit Card

The data says if the customer uses a credit card issued by the bank or not.

It also provides information regarding the number of family members of the customers, and the level of education of the customers. If they have completed undergraduate, graduation or professional level of education. Also, information regarding the annual income of the customers completed years of age of the customers, zip code of the customers and customer ID is provided.

Chapter 2

Project problem statement

2.1 Project problem statement:

Thera Bank has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes 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 to better target marketing to increase the success ratio with a minimal budget.

The department wants to build a model that will help them identify potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reducing the cost of the campaign.

Chapter 3

Project objectives

3.1 Project objectives:

- Study and analyze the data given using various charts and draw inferences from it.
- Study and analyze the data given using various statistical methods and techniques and draw inferences from the data.
- Using a classification model to predict the likelihood of a liability customer buying personal loans.

Chapter 4

Tools and techniques used

4. Tools and techniques used

4.1 Tools

4.1.1 R programming

- R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
- It's a platform-independent language. This means it can be applied to all operating systems.
- It's an open-source free language. That means anyone can install it in any organization without purchasing a license.
- R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.

4.1.2 Tableau

Tableau is a Business Intelligence tool for visually analyzing data. Users can create and distribute an interactive and shareable dashboard, which depicts the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, and relational and Big Data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique. It is used by businesses, academic researchers, and many government organizations for visual data analysis.

4.2 Techniques

4.2.1 Bar graph

- A bar graph is a type of graph used in statistics to show the relationship between two variables.
- It is a visual representation of data, in which each bar or rectangle represents the magnitude of one variable.
- The bars are usually arranged in order of magnitude, from highest to lowest.

- Bar graphs are often used to compare the distribution of different types of data, such as age, gender, or occupation.
- They can also show how data evolves over time.

4.2.2 Line graph

- A line graph is a type of graphical representation of data that uses points connected by straight lines to visualize a quantitative variable.
- In statistics, line graphs are used to represent the relationship between two or more variables.
- This can be used to show trends in data over time or to compare the values of multiple data sets.
- Line graphs can be used to represent data from surveys, experiments, and other types of data sources.

4.2.3 Histogram

- A histogram is a type of graph used in statistics to represent the frequency of a given set of data.
- A histogram is a graphical representation of the distribution of numerical data.
- It is a type of bar chart that shows the frequency of occurrence of different values of a particular variable.
- It is used to summarize large amounts of data, in a visual format, making it easier to interpret the data.

4.2.4 Box plot

- A boxplot is a graphical representation of numerical data in which the data is divided into percentiles.
- It is a statistical graphic that shows the dispersion and distribution of data.

- Box Plots are useful for displaying a set of data's median, quartiles, range, and outliers.
- It is a useful tool to analyze variance in a set of data and for comparing different samples.

4.2.5 Mean

- The mean is a measure of central tendency, which is the average of a given set of data. It is calculated by taking the sum of all the values in the set and dividing it by the total number of values.
- The mean is typically used to compare the differences between groups of data, such as the difference between the average scores of two different groups.

4.2.6 Median

- The Median is a measure of central tendency used in statistics and is defined as the middle value of a dataset when the values are arranged in numerical order.
- The median can be calculated using the median() function which takes a numeric vector as an argument and returns the median of the vector.

4.2.7 Mode

- The most commonly used modes in R programming are interactive mode, batch mode and script mode.
- In interactive mode, the user can type commands directly into the console.
- In batch mode, users can run a series of commands from a script file.
- In script mode, users can create a script file containing a series of commands and then run it as a single unit.

4.2.8 Minimum

In statistics, the lowest value in a series of data is referred to as the minimum. It is determined by finding the lowest value in a data collection.

4.2.9 Maximum

In statistics, the maximum is used to find the biggest value in a set of data. It may also be used to find outliers and compare data points.

4.2.10 Correlation

- The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its values can range from -1 to 1.
- A correlation coefficient of -1 describes a perfect negative, or inverse, correlation, with values in one series rising as those in the other decline, and vice versa.
- A coefficient of 1 shows a perfect positive correlation or a direct relationship.
- A correlation coefficient of 0 means there is no linear relationship.

4.2.11 ANOVA

ANOVA (Analysis of Variance) is a statistical technique for comparing the means of two or more groups. It is used to test for mean differences between groups and determine whether or not there is a significant difference between them.

4.2.12 ANCOVA

- ANCOVA (Analysis of Covariance) is a statistical technique for comparing the means of a dependent variable between two or more groups.

- When the independent variable is a categorical variable and the dependent variable is a continuous variable, this method is used.

4.2.13 Logistic regression

- Logistic regression is a statistical method for analyzing data sets that contain one or more independent factors that affect the result.
- The outcome is frequently a binary variable (i.e. where there are only two possible outcomes).
- Logistic regression is used to calculate the probability of a situation occurring given a set of independent factors.
- It's used in a variety of professions, like medicine, finance, and sociology etc.
- Logistic regression can be used to forecast whether or not a customer will pay on a loan, whether or not a patient will respond to a specific medical treatment, and so on.

Chapter 5

Project data analysis using Tableau and R Software

5.1 Descriptive statistics

5.1.1 Installing packages

Code:

```
install.packages("xlsx")

library("xlsx")

dataset<-read.xlsx("Bank_Personal_Loan_Modelling.xlsx",sheetIndex =2)

attach(dataset)

head(dataset)

myvar<-c('Age','Income','CCAvg','Experience','Education')
```

5.1.2 Mean

Code:

```
sapply(dataset[myvar],mean)
```

Output:

Age	Experience	Income	CCAvg
45.338400	20.104600	73.774200	1.937913

5.1.2.1 Interpretation

The average age, income, experience , and CCAvg of the customers of Thera bank are as mentioned above.

5.1.3 Median

Code:

```
sapply(dataset[myvar],median)
```

Output:

Age	Experience	Income	CCAvg
45.0	20.0	64.0	1.5

5.1.3.1 Interpretation

The median age income, experience, and CCAvg of the customers of Thera bank are as mentioned above.

5.1.4 Minimum

Code:

```
sapply(dataset[myvar],min)
```

Output:

Age	Experience	Income	CCAvg
23	-3	8	0

5.1.4.1 Interpretation

The minimum age, income, experience, and CCAvg of the customers of Thera bank are as mentioned above.

5.1.5 Maximum

Code:

```
sapply(dataset[myvar],max)
```

Output:

Age	Experience	Income	CCAvg
67	43	224	10

5.1.5.1 Interpretation

The maximum age, income, experience, and CCAvg of the customers of Thera bank are as mentioned above.

5.1.6 Standard deviation(sd)

Code:

```
sapply(dataset[myvar], sd)
```

Output:

Age	Experience	Income	CCAvg
11.4631656	11.4679537	46.0337293	1.7476662

5.1.6.1 Interpretation

The standard deviation of age, income, experience, and CCAvg of the customers of Thera bank are as mentioned above.

5.1.7 Variance

Code:

```
sapply(dataset[myvar],var)
```

Output:

Age	Experience	Income	CCAvg
131.404166	131.513962	2119.104235	3.054337

5.1.7.1 Interpretation

The variance of age, income, experience and CCAvg of the customers of Thera bank are as mentioned above.

5.1.8 Quantile:

Code:

```
sapply(dataset[myvar], quantile)
```

Output:

	Age	Experience	Income	CCAvg
0%	23	-3	8	0.0
25%	35	10	39	0.7
50%	45	20	64	1.5
75%	55	30	98	2.5
100%	67	43	224	10.0

5.1.8.1 Interpretation

The quantile of age, income, experience, and CCAvg of the customers of Thera bank are as mentioned above.

5.1.9 Range

Code:

```
sapply(dataset[myvar], range)
```

Output:

	Age	Experience	Income	CCAvg
[1,]	23	-3	8	0
[2,]	67	43	224	10

5.1.9.1 Interpretation

The range of age, income, experience and CCAvg of the customers of Thera bank are as mentioned above.

5.1.10 Summary

Code:

```
summary(dataset[myvar])
```

Output:

	Age	Experience	Income	CCAvg
Min.:	23.00	Min: -3.0	Min.:8.00	Min.: 0.000
1st Qu.:	35.00	1st Qu.:10.0	1st Qu.: 39.00	1st Qu.: 0.700
Median :	45.00	Median :20.0	Median: 64.00	Median: 1.500
Mean :	45.34	Mean :20.1	Mean: 73.77	Mean. : 1.938
3rd Qu.:	55.00	3rd Qu.:30.0	3rd Qu.: 98.00	3rd Qu.: 2.500
Max. :	67.00	Max. :43.0	Max. :224.00	Max. :10.000

5.1.11 Scatterplot:

Code:

```
plot(Income,CCAvg,col="red",pch=2, cex=0.5)
```

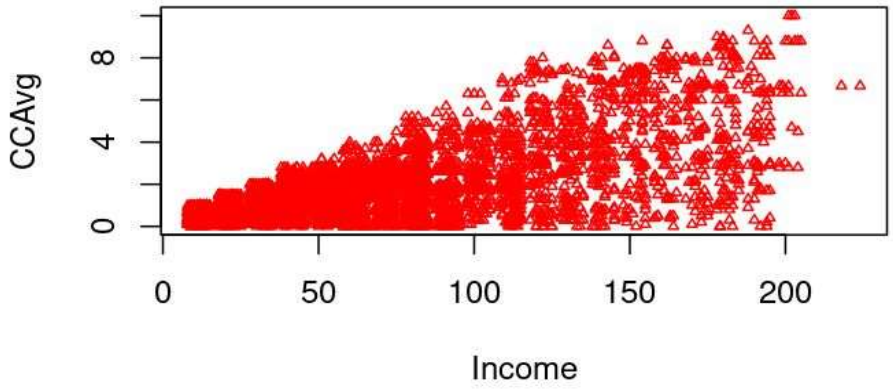

Fig:5.1 Scatter plot

5.1.11.1 Interpretation

The relationship between Income and CCAvg is linear.

5.1.12 Barplot:

Code:

```
table_1<-sapply(dataset[chart],mean)

barplot(table_1,main="Average",col="red")

chart<-c('Age','Experience')
```

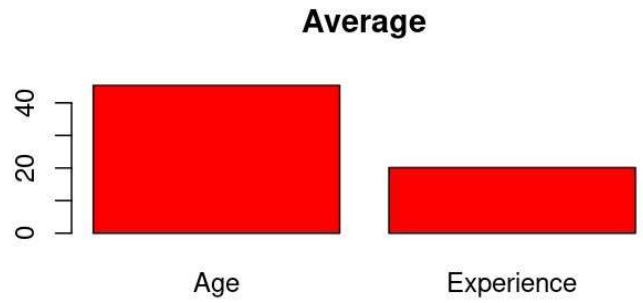

Fig:5.2 Bar chart

5.1.13 Histogram:

Code:

```
table bar<-sapply(dataset[myvar], mean)

hist(dataset$Income)
```

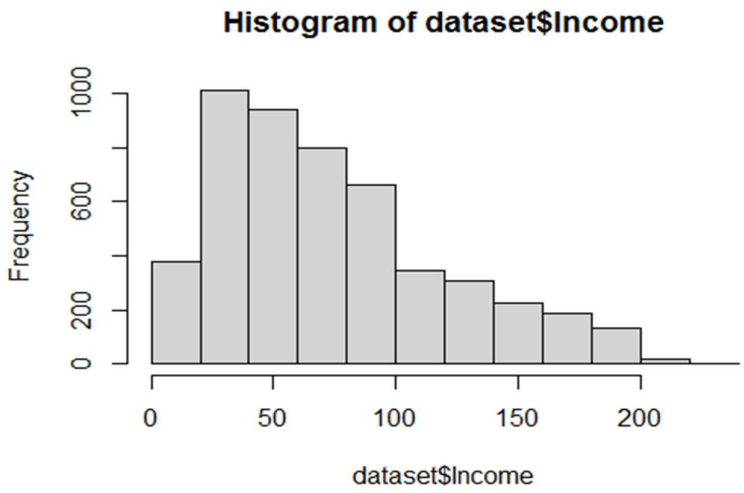

Fig:5.3 Histogram

5.1.13.1 Interpretation

The histogram is positively skewed. The majority of the people have less income than the mean income.

5.1.14. Boxplot:

Code:

```
boxplot(Experience)
```

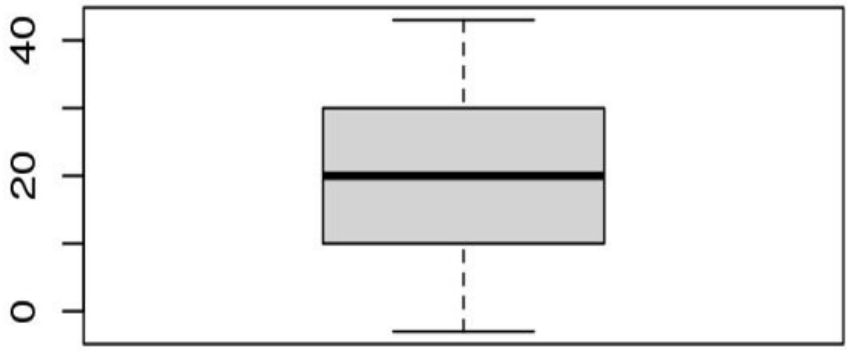

Fig:5.4 Box chart

5.1.14.1 Interpretation

In our data set, we found some outliers using a boxplot on the experience of the customers. This showed us that the data needs to be cleaned.

5.2 Inferential statistics

5.2.1 Chi-Squared tests

The chi-Squared test is used to find whether the two variables are related to each other. If two variables are independent (unrelated), the probability of belonging to a certain group of one variable isn't affected by the other variable.

Hypothesis:

H0 = There is no relationship between the number of people in the family and getting accepted for a personal loan

H1 = There is a relationship between the number of family members and getting a personal loan accepted.

Code:

```
ch<-chisq.test(Family,PersonalLoan)
ch
```

Output:

```
Pearson's Chi-squared test


data:  Family and PersonalLoan
X-squared = 29.676, df = 3, p-value = 1.614e-06
```

Conclusion:

As the significance level is more than 0.05. The null hypothesis is accepted. There is a significant relationship between the number of members in the family and personal loans.

5.2.2 Correlation:

Correlation is a statistical measure that describes the strength and direction of a linear relationship between two variables.

Code:

```
cor (CC Avg, Income)
```

Output:

```
0.6459926
```

Conclusion:

The value lies between 0 to 1 so it is a positive correlation. We have a moderate positive correlation for this dataset.

5.2.3 Anova

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Hypothesis:

H0 = The average income of the 3 education groups is not different.

H1 = The average income of the 3 education groups are different.

5.2.3.1 1 Way ANOVA

Code:

```
anova<-aov(Income~Education)
summary(anova)
```

Output:

```
Df        Sum Sq    Mean Sq     F value Pr(>F)
Education      1          372521       372521   182.2 <2e-16 ***
Residuals    4998       10220881      2045
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

As the significance level is less than 0.05. The null hypothesis is Rejected. It means that the mean income is different for each education group.

5.2.3.2 2-Way ANOVA

Hypothesis:

H0 = The average income of 3 education groups and family size is not different.

H1 = The average income of 3 education groups and family size are are different.

Code:

```
anova2 <-aov(Income~Family*Education)

summary(anova2)
```

Output:

```
                    Df   Sum Sq   Mean Sq F  value Pr(>F)
Family               1    262785    262785       135.7 <2e-16 ***
Education            1    334409    334409       172.7 <2e-16 ***
Family:Education     1    323656    323656       167.2 <2e-16 ***
Residuals         4996  9672553    1936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

As the significance level is less than 0.05. The null hypothesis is Rejected. It means that the mean income is different for each education group and family size.

5.2.4 ANCOVA

ANCOVA is commonly used for the analysis of quasi-experimental studies when the treatment groups are not randomly assigned and the researcher wishes to statistically "equate" groups on one or more variables which may differ across groups.

Hypothesis:

H0 = Means of CCAvg are equal after controlling the effect of Income in the Education Group

H1 = At least, one CCAvg mean is different from other Education groups after controlling the Income.

Code:

```
ancova<-aov(CCAvg~Income+Education)
summary(ancova)
```

Output:

```
             Df  Sum Sq  Mean Sq F value Pr(>F)
Income        1    6372    6372          3580 <2e-16 ***
Education     1       4       4          20.157
Residuals  4997    8893       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

As the significance level is more than 0.05. The null hypothesis is Rejected. It means that At least, one CCAvg mean is different from other Education groups after controlling the Income.

5.3 Building the Machine Learning Model:

As we saw in our business objective we have to build a machine-learning model to predict whether the customer will get it or not. Since for this business problem, the outcome is in Boolean, so we can use logical regression to train our machine learning model.

Step 1:

Installing necessary packages for building the machine learning model.

Code:

```
install.packages("caTools")
install.packages("ROCR")
install.packages("xlsx")
library("xlsx")
library(“caTools”)
library(“ROCR”)
```

Step 2:

After installing the necessary packages then we can load our data set.

Code:

```
dataset<-read.xlsx("Bank_Personal_Loan_Modelling.xlsx",sheetIndex =2)
attach(dataset)
```

Step 3:

Now our dataset is in the program and now we can split the dataset into a training set and a test set. By splitting the dataset into training and test data we can use the training data to train our model and test data to check the accuracy. We split our data set into ratios of 80% and 20%. 80% training set and 20% for the test set.

Code:

```
Splitting dataset
split <- sample.split(dataset$PersonalLoan, SplitRatio = 0.8)
train_reg <- subset(dataset, split == "TRUE")
test_reg <- subset(dataset, split == "FALSE")
```

Step 4:

Now we split the dataset into two parts, it's time to train our model. To train out model we use the training set “train_reg”. We loaded our prediction in the variable called “logistic_model”

Code:

```
logistic_model <-
glm(PersonalLoan~Age+Experience+Income+Family+CCAvg+Education+Mortgage,dat
a = train_reg, family = "binomial")
```

Step 5:

Now we can test our model without test data.

Code:

```
predict_reg <- predict(logistic_model,test_reg, type = "response")
predict_reg <- ifelse(predict_reg >0.5, 1, 0)
```

Step 6:

Now we tested out the model with our test data. It's time to check the accuracy of our model. For that, we can form a matrix and get the accuracy percentage.

Code:

```
table(test_reg$PersonalLoan, predict_reg)
missing_classerr <- mean(predict_reg != test_reg$PersonalLoan)
print(paste('Accuracy =', 1 - missing_classerr))
```

Output:

```
"Accuracy = 0.95"
```

Optional step:

To evaluate and visualise the performance of our model we use ROCR.

Code:

```
ROCPred <- prediction(predict_reg, test_reg$PersonalLoan)
ROCPer <- performance(ROCPred, measure = "tpr",x.measure = "fpr")
auc <- performance(ROCPred, measure = "auc")
auc <- auc@y.values[[1]]
```

After evaluating the model then we plot the result in the chart.

Code:

```
plot(ROCPer)
plot(ROCPer, colorize = TRUE, print.cutoffs.at = seq(0.1, by = 0.1), main = "ROC CURVE")
abline(a = 0, b = 1)
auc <- round(auc, 4)
legend(.6, .5, auc, title = "AUC", cex = 1)
```

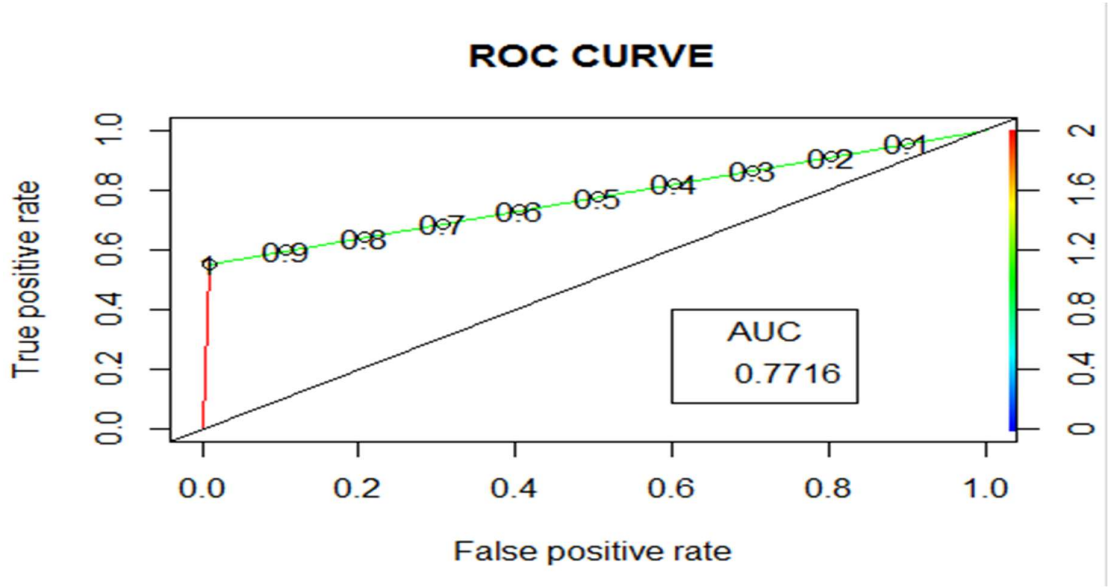

Fig:5.5 Roc chart

5.4 Tableau Dashboard

5.4.1 Tableau Dashboard 1

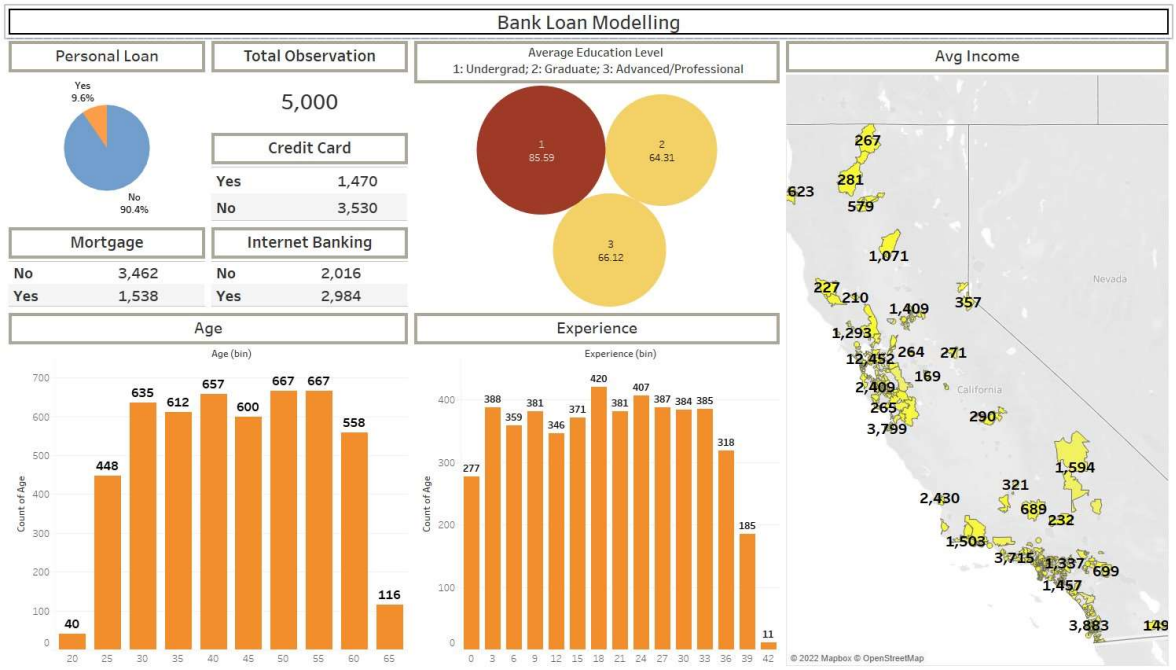

Fig:5.6 Dashboard 1

The dataset consists of 5000 observations in that most of the observations are in the age range from 25 to 63 with 2 to 26 years of experience.

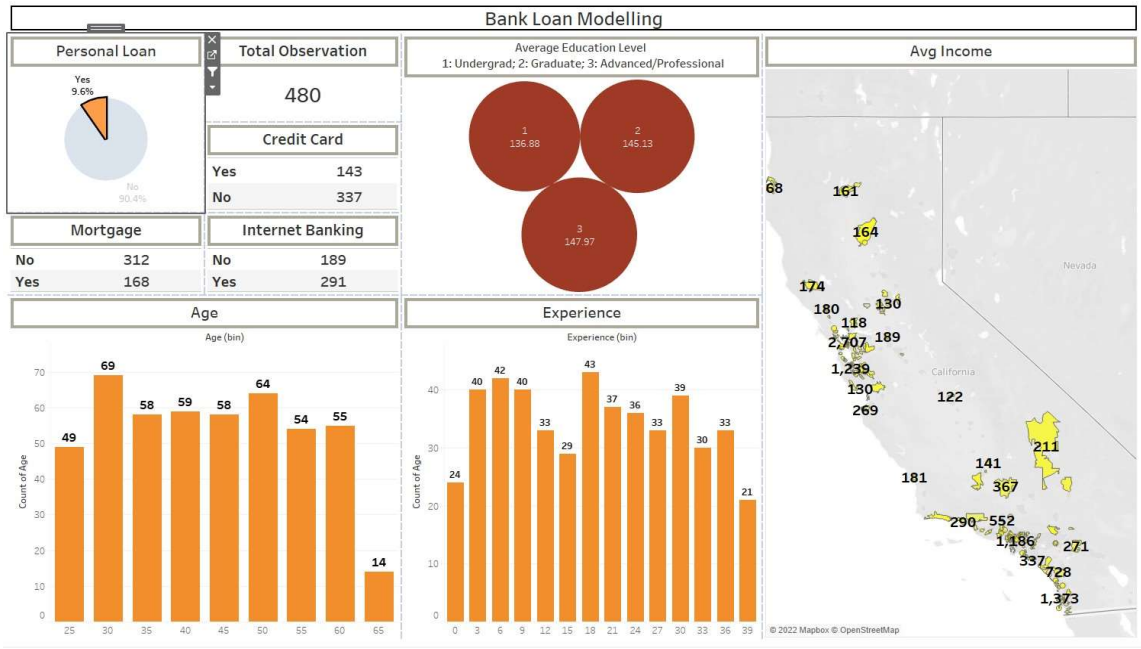

Fig:5.7 Dashboard 1.2

As we can see only 480 customers have taken the personal loan out of 5000 observations.

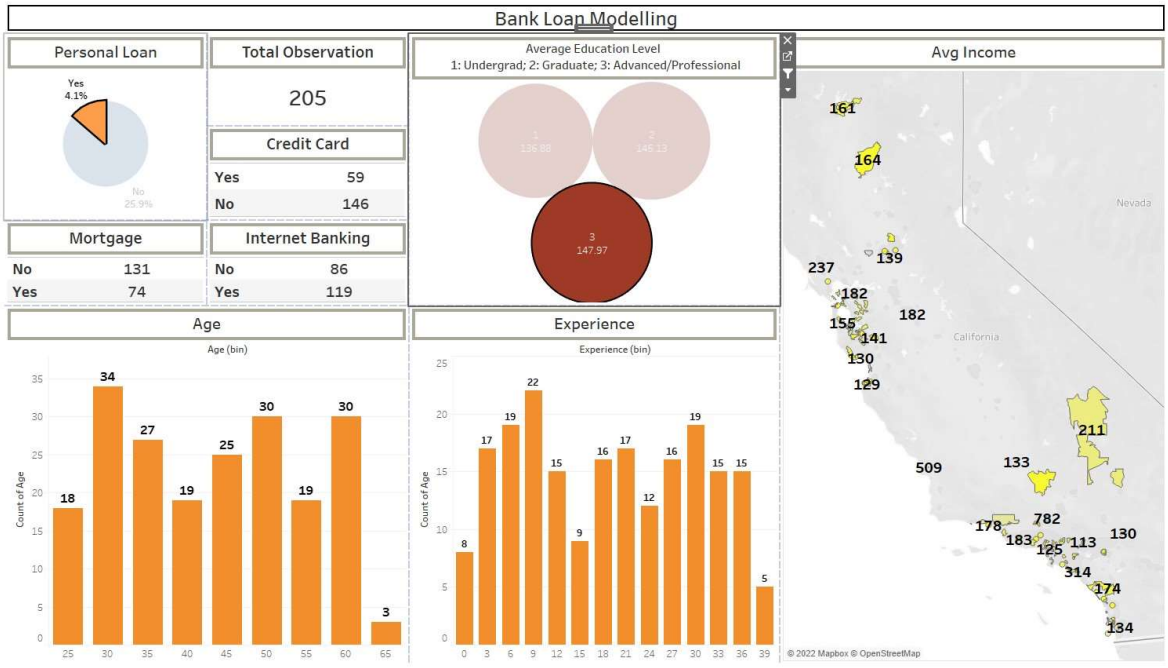

Fig:5.8 Dashboard 1.3

The highest number of people who have taken loans are in advanced level education and more than 70% of people who have taken personal loans also have a credit card. So, we can target the people who already have personal loans to buy credit cards. People who are less than 20 years old didn't take any personal loans.

5.4.2 Tableau Dashboard 2

Fig:5.9 Dashboard 2.1

In this dashboard, we can see the relationship between income and CCavg. By applying cluster analysis, we can see that we have 2 clusters in this chart. One is with low income and low CCAvg and the other one is with high income and high Avg. The mortgage and income are positively skewed distribution. Which means a greater number of observations lies below the mean.

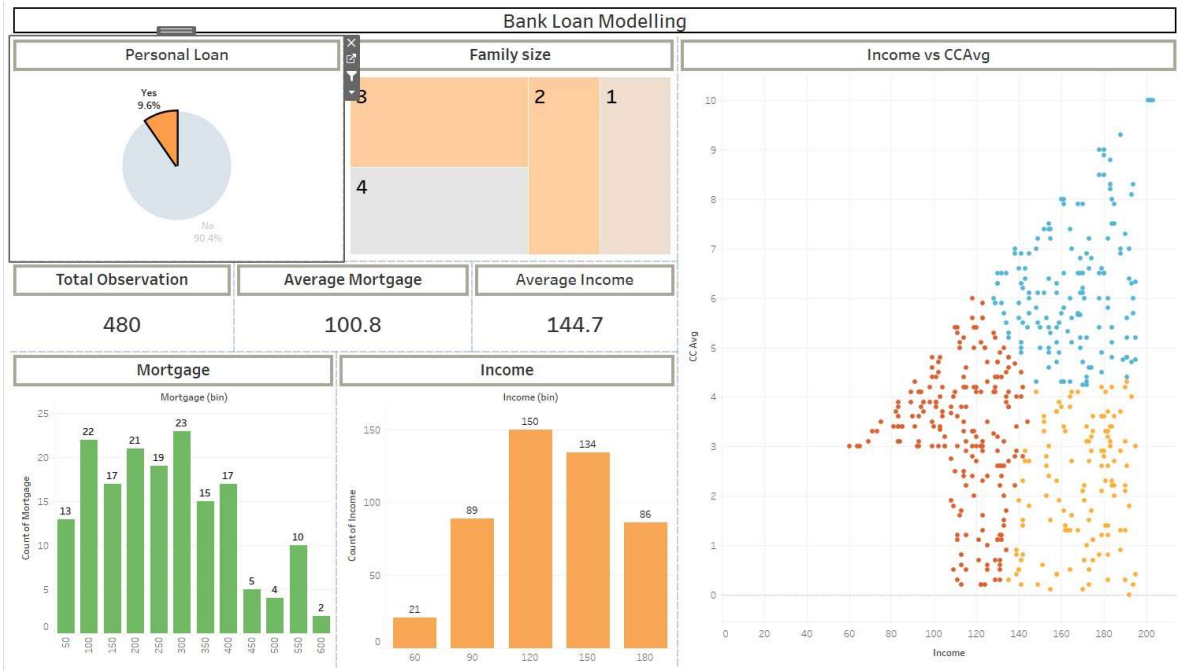

Fig:5.10 Dashboard 2.2

When we select only observations taking a personal loan, we can that people with low CCavg and income are not taking loans. More than that this plot clusters into 3 parts one with low ccavg and high income, high ccavg and high income and medium ccavg and medium income. The income distribution is negatively skewed. It means people with high incomes took the personal loan and most took low to medium mortgages.

5.4.3 Tableau Dashboard 3

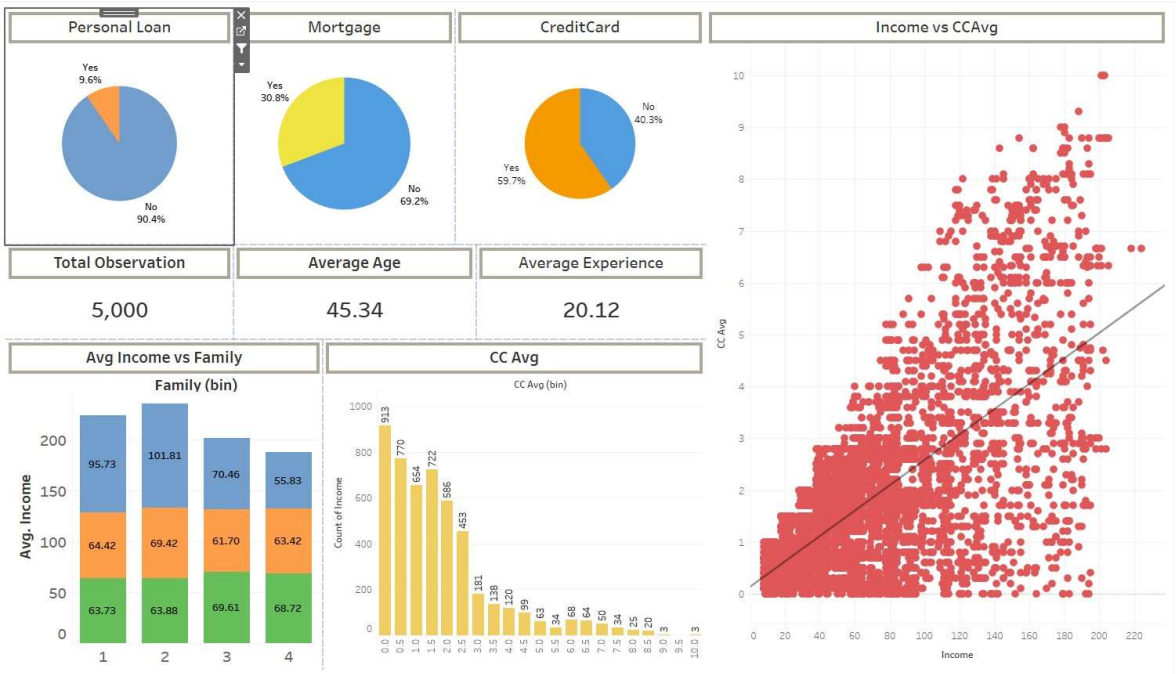

Fig:5.11 Dashboard 3.1

In this dashboard we have the personal loan, Mortgage and CreditCard to control the data flow dynamically. This dashboard contains the relationship between Income and CCavg. By applying the trend line in the scatter chart, we can see that when income increases CCavg also increases. This dashboard also shows the CCavg is positively skewed.

Fig:5.12 Dashboard 3.2

When we select the observation only people who took the personal loan, we can see that the CCavg is normally distributed and people with 3rd level of education take more loans comparatively than other education levels. People who took a personal loan have high incomes and CCAvg.

Chapter 6

Project finding

6.1 Descriptive analytics:

- There is a positive relationship between income and CCAvg.
- Income distribution is negatively skewed that mean a greater number of observations lies below the average of bank dataset.

6.2 Inferential statistics

- By doing correlation analysis we can get vale of 0.62. For the variable income and CCAvg. There is a positive moderate relationship between these 2 variables.
- There is high relationship between the customer getting accepted for personal loan getting accepted and size of the family.
- It means that the mean income is different for each education group.
- At least, one CCAvg mean is different from other Education group after controlling the Income.

6.3 Tableau

- The highest number of people who have taken loans are in advanced level education and more than 70% of people who have taken personal loans also have a credit card
- The mortgage and income are positively skewed distribution. Which means a greater number of observations lies below the mean.
- People took personal loan have high income and CCAvg.
- People with high income taken personal loan and most took low to medium mortgage.

Chapter 7

Project recommendation

7.1 Recommendation:

To get more people to buy loans we can focus on the points follows

- People with only undergraduate education take a Mortgage loan.
- We can give high spending limits on Credit cards for customers with high-income levels.
- Focus on the customer near the seashore area with only undergraduate education in the age group of 35 to 40 years old.
- Focus on the customer with 15 to 20 years of experience to sell more personal loans.
- More people take a loan when their income and CCAvg are high.
- To sell more Mortgages focus on customers within income levels of 30 to 60.
- To sell more Mortgages focus on the customers within ccavg of 0.5 to 3.5.
- To sell more credit cards focus on people with an income range of 30 to 60.
- To sell more credit cards focus on people with more than 100k mortgage.
- To increase the customer base for the bank focus on people with low income.
- The machine learning model that we created has high accuracy in predicting the future customer that can take a personal loan. So this is deployed to other data in the bank to classify the data.
- This model can be embedded into a webpage so new customers can check their eligibility before they apply for the loan. It will save the bank from possessing a lot of applications.

Chapter 8

Conclusion

8.1 Conclusion:

The model has been created with classification algorithm. Our model given the accuracy rate of more than 95%. We can deploy this model to find out the potential customers who have a higher probability of purchasing the loan. We also did AUC test. It also gave a good result that give an extra assurance to or model. Now we can deploy the model to the field for the future work.

Annexure

R Code

```
install.packages("xlsx")

library("xlsx")

dataset<-read.xlsx("Bank_Personal_Loan_Modelling.xlsx",sheetIndex =2)

attach(dataset)

head(dataset)

boxplot(Experience)

summary(Experience)

myvar<-c('Age','Income','CCAvg','Experience','Family','Education','Mortgage')

chart<-c('Age','Experience')

sapply(dataset[myvar],mean)

sapply(dataset[myvar],median)

sapply(dataset[myvar],min)

sapply(dataset[myvar],max)

table<-sapply(dataset[myvar],mean)

barplot(table,main="Average",col="red")

table_1<-sapply(dataset[chart],mean)

barplot(table_1,main="Average",col="red")

summary(dataset)
```

#Graphs

```
plot(Income,CCAvg,col="red",pch=2, cex=0.5)

hist(dataset$Income)

boxplot(Income~PersonalLoan)
```

#Chi Squared tests

```
ch<-chisq.test(Family,PersonalLoan)

ch
```

 # Correlation

```
cor.test(CCAvg,Income)

cor(dataset)
```

#Anova

```
# 1 Way ANOVA

anova<-aov(Income~Education)

summary(anova)
```

#2 Way ANOVA

```
anova2 <-aov(Income~Family*Education)
```

```
summary(anova2)

#ANCOVA

ancova<-aov(CCAvg~Income+Education)

summary(ancova)

#Linear Regression

linearregr<-lm(CCAvg~Income)

summary(linearregr)

plot(linearregr)

#Logical Regression

logicalreg<-
glm(PersonalLoan~Age+Experience+Income+Family+CCAvg+Education+Mortgage)

summary(logicalreg)

install.packages("caTools")

install.packages("ROCR")

library(caTools)

library(ROCR)
```

 # Splitting dataset

```
split <- sample.split(dataset$PersonalLoan, SplitRatio = 0.8)

train_reg <- subset(dataset, split == "TRUE")

test_reg <- subset(dataset, split == "FALSE")
```

 # Training model

```
logistic_model <- glm(PersonalLoan~Age+Experience+Income+Family+CCAvg+Education+Mortgage,data = train_reg, family = "binomial")
```

 # Summary

```
summary(logistic_model)
```

 # Predict test data based on model

```
predict_reg <- predict(logistic_model,test_reg, type = "response")
```

 # Changing probabilities

```
predict_reg <- ifelse(predict_reg >0.5, 1, 0)
```

 # Evaluating model accuracy using confusion matrix

```
table(test_reg$PersonalLoan, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$PersonalLoan)
```

```
print(paste('Accuracy =', 1 - missing_classerr))

ROCPred <- prediction(predict_reg, test_reg$PersonalLoan)

ROCPer <- performance(ROCPred, measure = "tpr",x.measure = "fpr")

auc <- performance(ROCPred, measure = "auc")

auc <- auc@y.values[[1]]
```

 # Plotting curve

```
plot(ROCPer)

plot(ROCPer, colorize = TRUE, print.cutoffs.at = seq(0.1, by = 0.1), main = "ROC CURVE")

abline(a = 0, b = 1)

auc <- round(auc, 4)

legend(.6, .5, auc, title = "AUC", cex = 1)
```