Project 1

Personal Loan Modeling

Prediction of whether a customer will respond to a Personal Loan Campaign:

A case study to devise campaigns with better target marketing

Data based on Thera Bank

GROUP 2

K.A.I. Kandewatte-s13994
S.H. Karunadhika- s13995
B.D.S. Suwaris –s14024

Abstract

This report is based on the descriptive analysis carried out on the Thera Bank, USA dataset sourced from the Kaggle website. The analysis aims to identify the potential clients for personal loans depending on the demographic information (age, income, etc.) and the customer's relationship with the bank (mortgage, securities account, etc.), thereby to identify the most influential predictors to predict whether a customer will respond to a Personal Loan Campaign. Data can be considered as an asset in the process of marketing. By utilizing such data, marketing department of the Thera Bank can begin to build better marketing campaigns in addition to creating stronger product offerings. Univariate and bivariate plots between the response and predictor variables led the pathway to identify the different techniques that could be used for the advanced analysis. The results showed an association between accepting a personal loan and the variables: location, income, family size, education, income, CCAvg, Security Account, Mortgage and CD Account. The results suggest that techniques such as SMOTE, ADASYN are needed to handle the effect of class imbalance that exist in the dataset and shrinkage methods must be used to handle the multicollinearity observed in order to predict whether a customer will respond to a personal loan campaign.

Table of Content

Page No.

List of figures

List of tables

1. Introduction

There is a rising role of marketing in the banking sector and it the function which gets personality and image for bank on its customers' mind. Data can be considered as an asset in the process of marketing. By utilizing such data, marketing department of the Thera Bank can begin to build better marketing campaigns in addition to creating stronger product offerings. This case study focuses on predicting whether a customer will respond to a personal loan campaign to devise campaigns with better target marketing. Unlike a business or a commercial loan, a personal loan is an advance given to an individual for his or her personal use that can be used for large purchases, debt consolidation, emergency expenses and much more. The personal loan market is exploding. According to U.S. web-based lending company Lending Club, personal loan originations from 2017 to 2018 were up 20%. American consumer credit reporting agency TransUnion says personal loans are by far the fastest-growing U.S. consumer-lending category. This explains the competitiveness that prevail in personal loan market. The Thera Bank which is located in USA provides such personal loans and this analysis will provide insights to identify potential clients for loans. In this report, we provide you an exploratory analysis on the factors that are associated with the conversion of liability customers of the Thera Bank to personal loan customers while retaining them as depositors.

2.The Question Description

Our main attention was paid to the variable "Personal Loan". We, as the marketing department of Thera Bank is concerned about exploring ways of converting liability customers of our bank to personal loan customers while retaining them as depositors. Furthermore, it is important to know which customers are likely to accept a personal loan in order to perform a market segmentation. This will result in lowering of marketing costs and optimization of advertising efforts since we can target the potential highest-yield customers in the next campaigns.
The main objectives of this project are as follows:

1. To identify the potential clients for loans by minimizing loss of resources and opportunities.
2. To construct a suitable model to predict whether a customer will accept to a Personal Loan Campaign based on customer demographic information and the customer's relationship with the bank.

3. Description of the Dataset

The Thera Bank dataset obtained from Kaggle website is a collection of observations from 5000 customers and 14 variables. 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).

No.	Variable	Description
1	ID	Customer ID
2	Age	Customer's age in completed years
3	Experience	Number of years of professional experience
4	Income	Annual income of the customer ($000)
5	ZIP Code	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	CreditCard	Does the customer use a credit card issued by this Bank?

Table 3.1- Table -description of the variables

4. Data Cleaning

Data cleaning was performed to correct and remove inaccurate and corrupt data since they can drive the analysis to wrong conclusions. There were no missing values or duplicates in this dataset. The variable "ID" does not add any benefit to the analysis, so it was removed. Some values of the variable "Experience" were negative and converted them into their absolute values.

5. Descriptive Analysis

5.1 Response Variable: Personal Loan

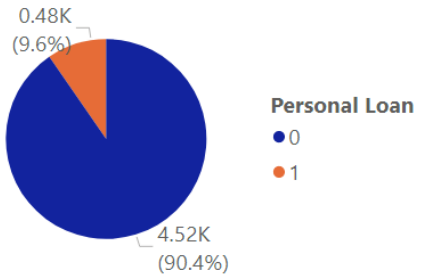

Figure 5.1.1 Pie chart of personal loan

As observed from the figure, the conversion rate of the personal loan marketing campaign is 9.6%. By carefully identifying the potential clients for personal loans, marketing costs of the department can be optimized. 90.4% of the customers have not applied for personal loans after the campaign. This highlights the importance of market segmentation. This also indicates that the dataset obtained is highly imbalanced.

5.2 Location

As observed from the figure (**5.2.1**), customers in some parts of the country have not obtained a personal loan after the campaign. One such city is Tahoe. According to *city-data site of USA*, 13.0% of South Lake Tahoe, CA residents had an income below the poverty level in 2019, which was 9.9%. In the given figure, the size of each circle denotes the average income of the city. Also, the shaded portion represents the proportion of average income of customers who accepted the personal loan Vs who did not. It can clearly be seen that average income of the cities where there are customers who have accepted the loan is very high comparatively to the other cities. Furthermore, by considering the size of the circle, it is worth noting that the average income is very low in parts of the country where none of the customers have obtained a personal loan.

Figure 5.2.1 Mean income of customers who accepted the personal loan vs. who did not

5.3 Income

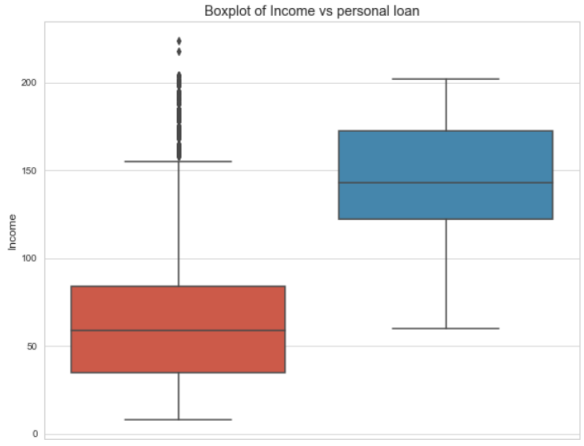

Figure 5.3.1 Boxplot of income vs. personal loan

As we can observe from the boxplot, the median annual income of customers who have applied for personal loans is higher than that of customers who haven't applied for personal loans. This is clearly visible form the figure. According to *Forbes Advisor journal*, Lenders impose income requirements on borrowers to ensure they have the means to repay a new loan. Minimum income requirements vary by lender. The SoFi finance company in USA imposes a minimum salary requirement of $45,000 per year.

According to the figure (**5.3.1**), we can assume that the minimum annual income requirement of Thera Bank is approximately $50,000 per year. As per *Ascent personal loan statistics*, Americans with income over $100,000 are more likely to consider taking out a personal loan than those with lower incomes. This is also evident from the figure (**5.3.2**). Thus, targeting the wrong market (i.e. customers with low income) will costs a lot for the marketing department.

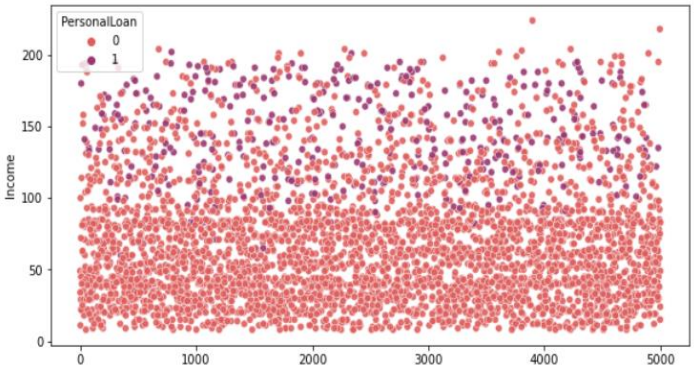

Figure 5.3.2 Scatter plot of Income with personal loan

5.4 Family Size

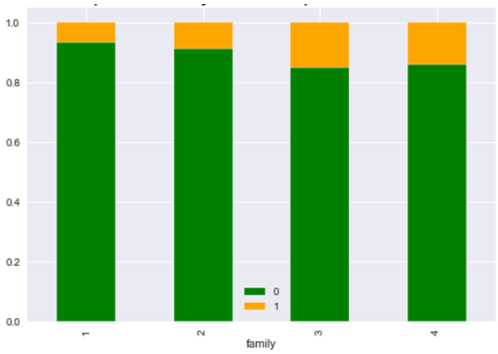

Figure 5.4.1 Stack Bar plot of family by personal loan

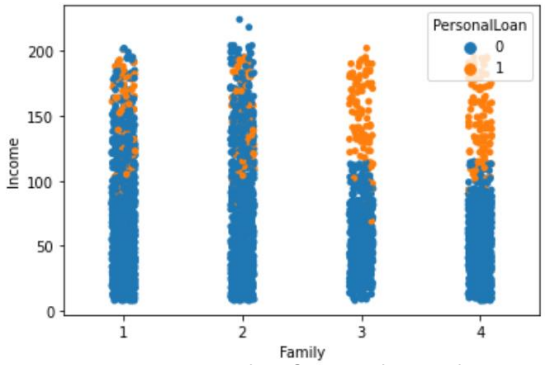

Figure 5.4.2 strip plot of income by Family size

According to figure (**5.4.1**), the proportion of personal loans taken by a family of 3 is the highest. There is an interesting pattern that shows larger families feel the need to get personal loans more. The expenses of large families are high. Another important insight that can be gained from the figure (**5.4.2**) is that only the higher income families out of the families of larger sizes have obtained a personal loan. The minimum income requirements imposed when obtaining loan personal loans may be the possible reason for this.

5.5 Education

When considering the education, the customers with Advanced/professional education (3) have accepted the personal loan more than the undergraduates (1) and graduates (2). The possible reason for this can be since the government of USA issues scholarships and grants for undergraduates, they may not find it necessary to buy a personal loan to fund their education (*Federal Student Aid Office*).

Mainly, the undergraduates do not have a proper source of income to repay the loans. Most types of grants, unlike loans, are sources of free money that generally do not have to be repaid. Also, higher the education level the more the customer is confident and open to opt for personal loans.

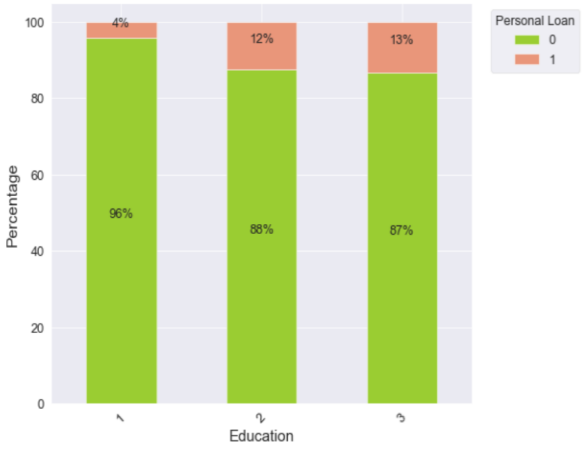

Figure 5.5.1 Stack Bar Chart of personal loan by Education

5.6 Age

Table 5.6.1- Table containing distributional summary statistics for Age variable

Mean	Min	25%	Median	75%	Max
45.446750	23.0	36.0	46.0	55.0	67.0

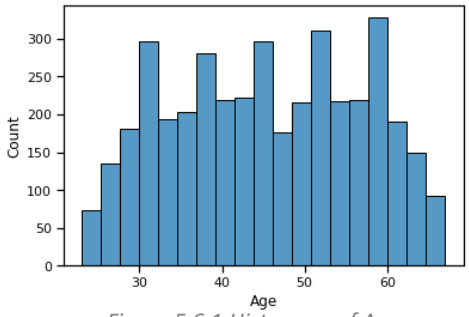

Figure 5.6.1-Histogram of Age

It can be seen that most of the customers fall into 36 to 55 years of age. Through the analysis, it was observed that age does not affect in accepting the personal loan. According to *Forbes advisor magazine*, in USA most lenders require applicants to provide at least two forms of government-issued identification to prove they are at least 18 years old. Since all the customers of the Thera Bank are above 23 years, any customer will be eligible for a loan based on their age even though there are other limitations as explained previously. Furthermore, it was observed that there is a very strong positive relationship between experience and age as observed in the figure (**5.6.2**). An "aged" person can be said to have lot of experience.

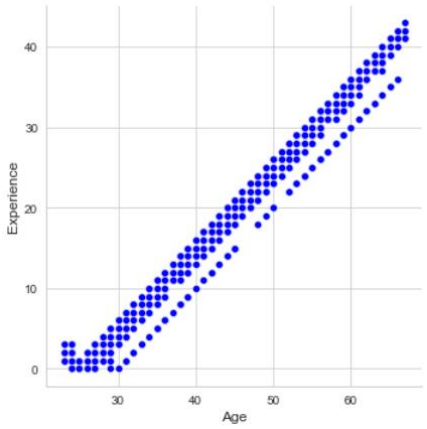

Figure 5.6.2 scatter plot of Experience vs. Age

5.7 Credit Card Spending

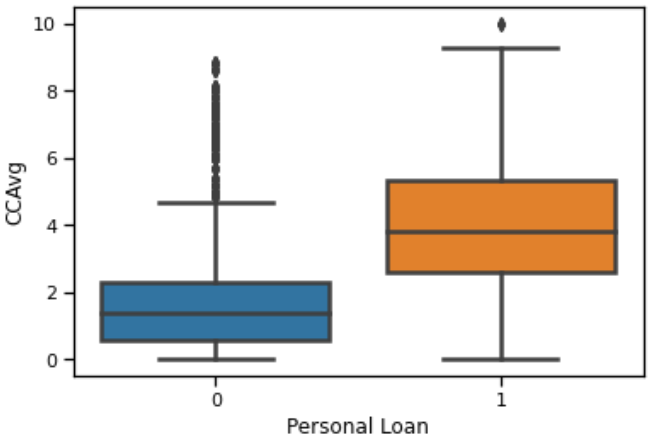

Figure 5.7.1 Boxplot of CCAvg by personal loan

Figure 5.7.2 scatter plot of CCAvg vs. income with personal loan

According to the given boxplot, customers who have a high average spending on Credit Card per month have decided to get a Personal Loan offered in the last campaign. According to *Megan DeMatteo*, who is an editor at *Business Insider*, the credit cards have the highest interest rates out of every kind of credit product. The *website cnbc.com* which provides high-quality personal finance advice states the solution for the above high interest rates of credit cards is to use a personal loan to consolidate the credit card debt into one monthly payment. This provides evidence as to why customers with a high average spending on credit cards tend to buy personal loan than the other customers. Apart, according to a research that was carried out by LendingTree, the vast majority of borrowers in USA are using personal loans to consolidate debt and refinance credit cards, combining for a total of 61% for all personal loans. The trends discussed in the LendingTree analysis are immediately clear in our visual.

Furthermore, from the scatter plot, it is clear that customers with a higher income have a higher average spending on credit card. Since the credit cards have the highest interest rates out of every kind of credit product as mentioned above, only the customers with a higher income can afford to pay the interest rates.

5.8 Certificate of Deposit Account

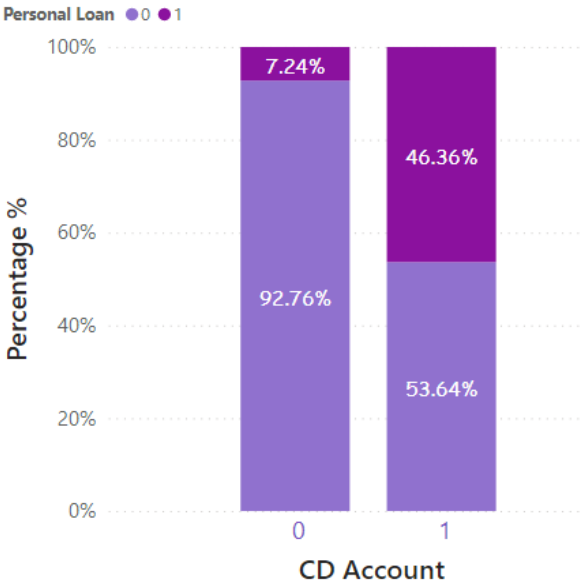

Figure 5.8.1 Bar plot of CD Account by personal loan

A certificate of deposit (CD) is a type of savings account usually issued by commercial banks, which restricts your access to the money you invest but offers much higher interest rates than those associated with regular savings accounts. According to the bar chart given, a significant percentage i.e. 46.36% of customers who have CD accounts have accepted a personal loan after the last campaign. Thus, the impact of CD accounts is very high on the purchase of a personal loan.

Since CD accounts restrict access to money, customers tend to buy personal loans to satisfy their monetary requirements. Apart, the interest rates from CD accounts are much higher which in result aids the customers to repay the loan amount.

According to the article written by *Jennifer Brozic*, there are two types of personal loans namely secured and unsecured. Secured personal loans are backed by collateral, such as a savings account or CD Account. Thus, this may be another possible reason as to why higher percentage of customers with CD accounts have accepted the personal loan.

5.9 Securities Account

Figure 5.9.1 Stack Bar plot of securities account by

Figure 5.9.2 Strip plot of CCAvg vs. securities account by personal loan

It can be observed from the given figure that most of the customers who have accepted the personal loan do not have a securities account. Securities account is an account that holds financial assets such as securities on behalf of an investor with a bank, broker or custodian. There are various risks associated with securities accounts. Some of those risks are Market (systematic) risk, Business (nonsystematic) risk, Credit risk, Liquidity (marketability) risk and Prepayment risk. These risks may demotivate the securities account holders to buy a personal loan since it adds another risk. This explains the reason as to why most of the customers who have accepted the personal loan do not have a security account. Also, it can be noted that the average credit card spending per month is low among customers who have a securities account than the other group. The securities account holders may be less motivated to spend on credit cards since they already have many risks involved with the securities account as explained earlier.

5.10 Online

It can clearly be seen from figure (**5.10.1**), that more than 50% of the customers transact online. According to the author *Jamie Gonzalez-Garcia*, in 2016, 62% of Americans cited digital banking as their primary method of banking (Bank of America Trends in Consumer Mobility Report 2016). Our analysis provides further evidence to this as 60% of the customers transact online.

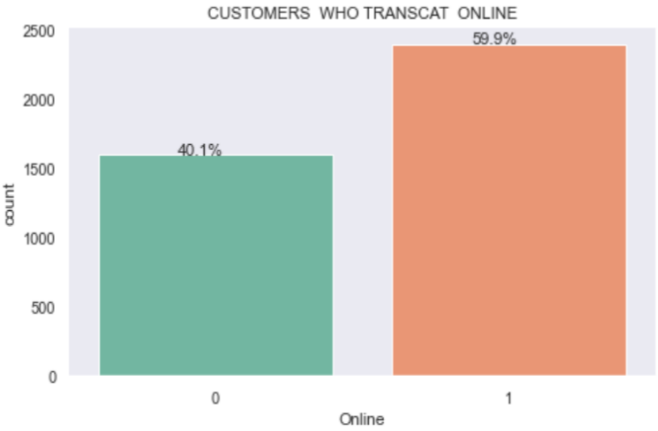

Figure 5.10.1 Bar plot of customers who transact online

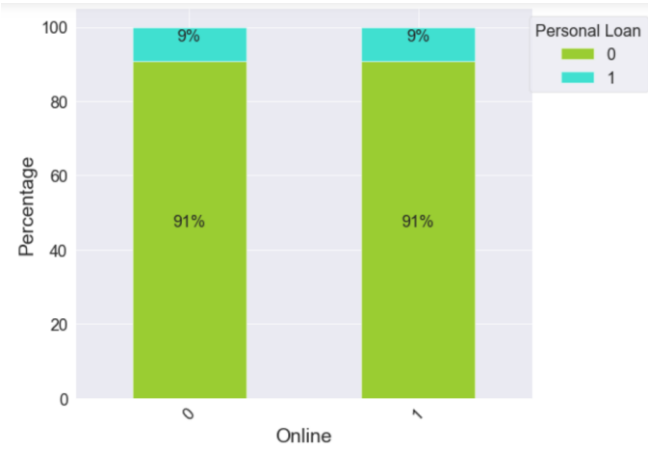

Figure 5.10.2 Stacked Bar plot of online by personal loan

Another fascinating insight that was gained through this analysis is that an equal percentage of online users and non-online users have obtained the personal loan even though the initial campaign was not done in online platforms. By making use of online platforms to market the personal loans in the next campaign will attract more customers which in result would give a healthy conversion rate.

5.11 Mortgage

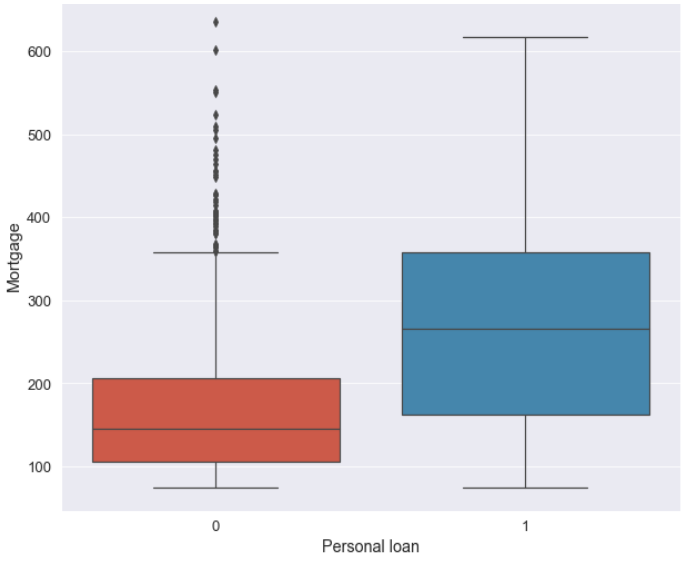

Figure 5.11.1 boxplot of mortgage by personal loan

A mortgage is a necessity if you can't pay the full cost of a home out of pocket. Mortgage is a type of secured loan you can use to buy or refinance a home. In other words, mortgages are a way to buy a home without having all the cash up front. If the borrower fails to make the payments, the lender can take possession of the home, in a process known as foreclosure. It's observable that the median mortgage amount is relatively high among the customers who have accepted the personal loans. This behavior may be due to the inability to pay the installments for the mortgages, especially among those who have obtained high mortgages.

Another interesting pattern can be observed from the scatter plot (**5.11.2**). The customers with higher incomes have obtained mortgages with higher values. This implies that the house price of the house they are intending to refinance is also high.

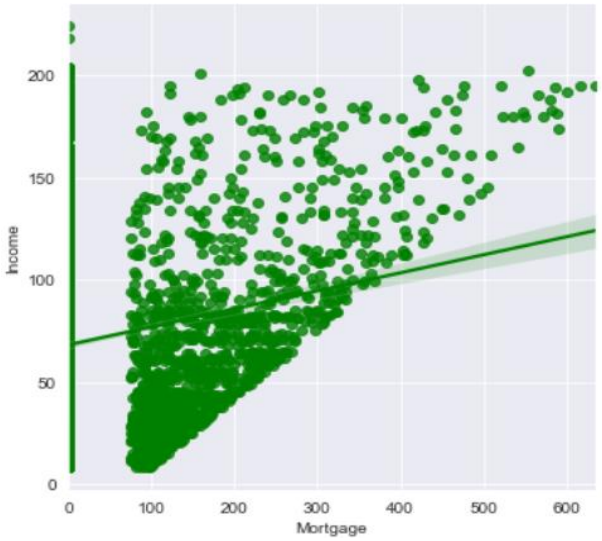

Figure 5.11.2 scatter plot of mortgage vs. income

6. Suggestions for Advanced Analysis

Univariate and bivariate plots between the response and predictor variables led the pathway to identify the different techniques that could be used for the advanced analysis. Spearman rank correlation test was performed to get an idea about the association between continuous predictor variables and Goodman and Kruskal's tau measures were used to check the association between categorical variables.

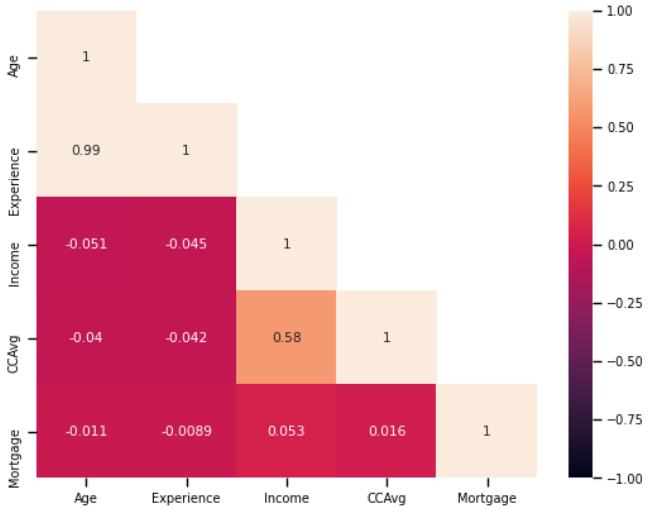

Figure 6.1 Spearman rank correlation plot

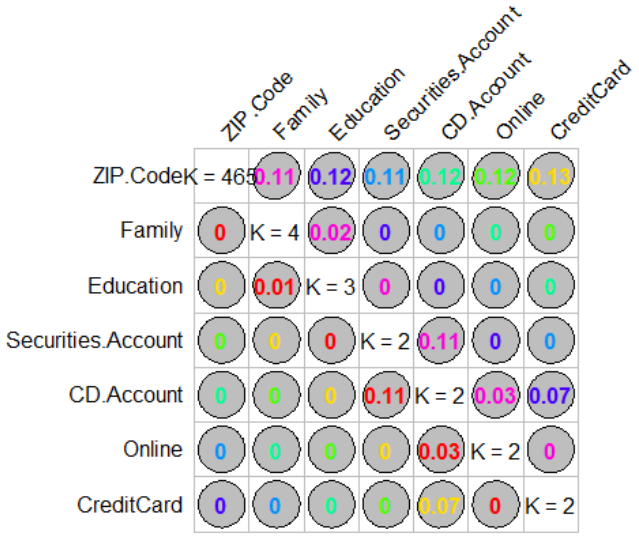

Figure 6.2 Goodman and Kruskal's plot

It can be observed that there exist several multicollinearities between some predictors. Especially between the years of experience and Age which indicates almost perfect multicollinearity (0.99). Thus, Logistic Regression with shrinkage methods such as ridge, lasso, elastic net and classification trees such as XGBoost and random forest will be used for the advanced analysis. Also, we suggest to use under sampling and over sampling techniques since the dataset is imbalanced as discussed previously.

7. Appendix

The map in Fig 5.2.1, pie chart in Fig 5.1.1 and bar chart in Fig 5.8.1 were drawn by using *Power BI, and* Goodman Kruskal's plot in Fig 6.2 was drawn by using *R*. All other figures were drawn using *Python* and the codes are given below.

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import seaborn as sns

df=pd.read_csv("./archive/Bank_Personal_Loan_Modelling.CSV")
df.head()
df.info()
df.describe() # descriptive statistics
df.duplicated().any() #no duplicates
df.isnull().sum().sum() # no missing values
```

```
# Data Cleaning

# Drop ID since it doesn't affect to analysis
df.drop("ID",inplace=True,axis=1)

# Take absolute value of experiance
df['Experience']=df['Experience'].abs()

# Remove space in Col names
df.columns = df.columns.str.replace(' ', '')

# Change data types
df.PersonalLoan = df.PersonalLoan.astype("category")
df.SecuritiesAccount = df.SecuritiesAccount.astype("category")
df.CDAccount = df.CDAccount.astype("category")
df.Online = df.Online.astype("category")
df.CreditCard= df.CreditCard.astype("category")
df.Education= df.Education.astype("category")
df.Family= df.Family.astype("category")
df['ZIPCode'] = df['ZIPCode'].astype(str).str.zfill(5)

df.info() #check data types again

#Splitting
axis_x = df.drop(['PersonalLoan'],axis=1)
axis_y = df['PersonalLoan']

#spliting the data into 80/20
x_train,x_test,y_train,y_test = train_test_split(axis_x,axis_y,test_size=0.2,random_state=100)
```

```
x_train.head()

# for ease of analysis, combine X_train and y_train and name it as df_train
train=pd.concat([x_train,y_train],axis=1)
train.head()
```

```
# Descriptive Analysis

# Figure 5.3.1 - Boxplot of income vs. personal loan
sns.reset_defaults()
sns.boxplot(x="PersonalLoan",y="Income",data=train)
plt.show()

# Figure 5.3.2 - Scatter Plot of Income with personal loan
plt.figure(figsize=(10,5))
sns.scatterplot(x = train.index, y = "Income", data=train, hue = "PersonalLoan", palette="flare", alpha = 0.9)
plt.show()

# Figure 5.4.1 - Bar plot of family by personal loan
s=train.Family[train['PersonalLoan']==1].value_counts().sort_index()
col=['Red','Green','Indigo','Orange']
ax=s.plot.bar(width=.9,color=col)
plt.xlabel("Family")
plt.ylabel("Count of taken Personal Loan")
for i, v in s.reset_index().iterrows():
    ax.text(i, v.Family + 1.5, v.Family, color='Indigo')
plt.show()
```

```
# Figure 5.4.2 - Strip plot of income by Family size
sns.stripplot(x="Family",y="Income",data=train,hue=train.PersonalLoan,jitter=True)
plt.show()

# Figure 5.5.1 - Bar plot of personal loan by Education
sns.set_theme(style="dark")
sns.set_palette('Set2')
sns.countplot(x=train['PersonalLoan'],hue="Education",data=train,saturation=1.5)
plt.show()

# Figure 5.6.1 - Histogram of Age
sns.reset_defaults()
sns.histplot(x="Age",data=train)
plt.show()

# Figure 5.6.2 - Scatter plot of Experience vs. Age
sns.scatterplot(x = "Age", y = "Experience", data=train)
plt.show()

# Figure 5.7.1 - Boxplot of CCAvg by personal loan
sns.boxplot(x="PersonalLoan",y="CCAvg",data=train)
plt.show()

# Figure 5.7.2 - scatter plot of CCAvg vs. income with personal loan
plt.figure(figsize=(10,8))
sns.set_style("darkgrid")
sns.set_palette('Set2')
sns.scatterplot(x="Income",y="CCAvg",data=train,hue="PersonalLoan")
plt.show()
```

```
# Figure 5.9.1 - bar plot of securities account by personal loan
sns.reset_defaults()
sns.countplot(x=train["SecuritiesAccount"],hue=train.PersonalLoan,dodge=True)
plt.show()

# Figure 5.9.2 - strip plot of CCAvg vs. securities account by personal loan
sns.set_style("darkgrid")
sns.stripplot(x="SecuritiesAccount",y="CCAvg",data=train,hue=y_train,jitter=True)
plt.show()

# Figure 5.10.1 - Bar plot of customers who transact online
sns.set_theme(style="dark")
sns.set_palette('Set2')
ax=sns.countplot(x=train["Online"], data=train )
for p in ax.patches:
    percentage = '{:.1f}%'.format(100 * p.get_height()/len(train["Online"]))
    x = p.get_x() + p.get_width() / 2 - 0.05
    y = p.get_y() + p.get_height()
    plt.annotate(percentage, (x,y),ha='center')
plt.title('Customers  who transcat  Online'.upper())
plt.show()

# Figure 5.10.2 - Bar plot of online by personal loan
sns.reset_defaults()
sns.set_style("darkgrid")
sns.countplot(x= train.Online,hue="PersonalLoan",data=train)
plt.show()
```

```
# Figure 5.11.1 - boxplot of mortgage by personal loan
sns.set_style("darkgrid")
sns.boxplot(x="PersonalLoan",y="Mortgage",data=train[train['Mortgage']>0])
plt.show()

# Figure 5.11.2 - Scatter plot of mortgage vs. income
sns.jointplot(x="Mortgage",y="Income",data=train,kind="reg",color="green")
plt.show()

# Figure 6.1.1 - Spearman rank correlation plot
corr = train.corr(method='spearman') # spearman correlation
sns.reset_defaults()
sns.set_context("paper", font_scale=0.8, rc={"lines.linewidth": 2.5})
mask = np.zeros_like(corr)# create a mask so we only see the correlation values once
mask[np.triu_indices_from(mask, 1)] = True
sns.heatmap(corr,mask=mask, annot=True, vmin=-1,vmax=1)
plt.show()
```

R code for Goodman and Kruskal's plot

```
library(GoodmanKruskal)
var_set <- c ("ZIPCode","Family","Education","SecuritiesAccount", "CDAccount","Online","CreditCard")
df1 <- subset(train, select = var_set)
GKmatrix <- GKtauDataframe(df1)
plot(GKmatrix)
```