Assignment 2: Big Data Analytics
Naïve Bayes, Confusion Matrix, ROC
Valay Shah

Managerial Conclusion

From conducting the Naive Bayes, confusion matrix, and ROC analyses we can conclude that the variables of the customer being active in online bank services and having a credit card has no bearing on them taking out a personal loan. In fact, flipping a coin and letting heads be the probability of the customer taking out a personal loan and tails being not will yield the same results. As a result of this, in future findings, other variables need to be considered for their predictive power to finding if a customer is likely to take a customer loan.

Purpose

The purpose of this data exercise was to use bank data on 5000 customers to predict the impact of whether a customer is an active user of online services and the impact of the customer having a bank-issued credit card on a customer's response to a personal loan campaign. Specifically, we want to see if the bank customer is more likely to take out a loan if they have an online account with the bank and/or if they have a credit card issued by the bank.

Methodology

To observe the relationship between these variables and the impact that the variables of online and credit card have on personal loan, a Naive Bayes methodology was utilized to see how customers that do and do not use the online and credit card services respond to the personal loan campaign. To begin, training and validation sets were selected. The training set was selected as 60% of the entirety of the data with the remaining 40% being the validation set. After this, naïve bayes was run with the training dataset which produced the following results:

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
         0          1
0.90733333 0.09266667

Conditional probabilities:
   Online
Y           0         1
  0 0.4077884 0.5922116
  1 0.4028777 0.5971223

   CreditCard
Y           0         1
  0 0.7072006 0.2927994
  1 0.6870504 0.3129496
```

From the table above we should pay close attention to the conditional probability tables for both, online, and credit card. The first table with the Y value being 0 is someone who will not take out a personal loan and 1 being someone who will take out a loan. From the table, we can clearly see that a customer who has online banking

setup will take out a loan 60% of the time and will not take out the loan with around the same probability. The same holds true for a customer who does not have an online services account set up that they actively use as the probability for not taking out a loan and taking out a loan for someone without online services remains the same at 40%. This heavily seems to imply that a customer having online services has no bearing on whether or not that customer will take out a loan.

In the next table, we can observe a similar pattern. Whether or not a customer has a credit card with the bank has a negligible impact on whether said customer will take out a loan. Not having a credit leads to 69%-71% of customers taking or not taking out a loan. Whereas for those that have a credit card, they are not taking the loan or taking the loan at 29%-31% which means that a customer is equally likely to take out a loan and not take out a loan if they have a credit card at around 30% and if they do not have a credit card at around 70%.

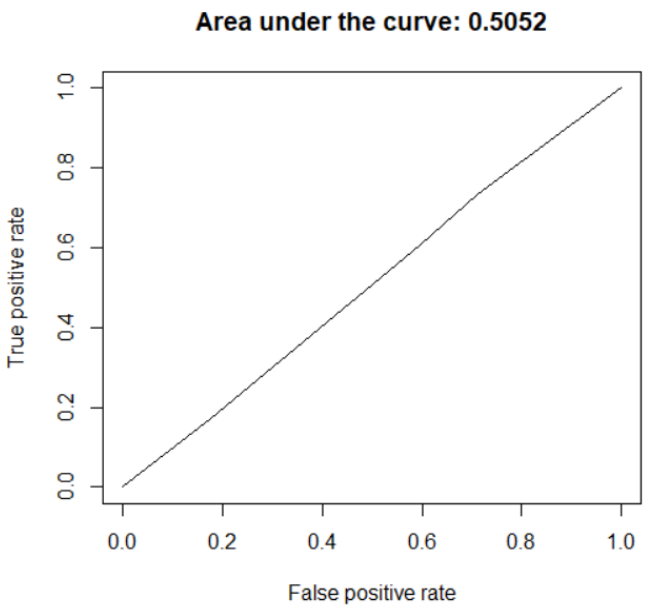
Since Naïve Bayes failed to give us conclusive results on the impact of online services and possession of a credit card on the willingness to take out a personal loan, the confusion matrix was next applied.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1798  202
         1    0    0

               Accuracy : 0.899
                 95% CI : (0.885, 0.9119)
    No Information Rate : 0.899
    P-Value [Acc > NIR] : 0.5187
```

The confusion matrix at 50% yielded strange results with all predictions being 0s which was not correct either as the predictions should be split between true and false values for the confusion matrix to display any prediction power. The accuracy was high at 90% because the prediction model captured all the negative values but also captured 100% of the false negatives which makes it a useless predictor. This led to the ROC matrix being utilized to study the relationship between true and false positive rates.



From this ROC curve, the picture becomes clear. The ROC graph yields an area under the curve of 0.5052 which means that there is a 50% chance of the two variables of a customer having online services and a credit card leading to them taking out a personal loan. This is essentially a coin flip and we can therefore determine that flipping a coin to predict if the customer will take out a personal loan is the same as using these two variables to predict if a customer will be taking out a personal loan.

Code:

```
#Naive Bayes Code
install.packages('e1071', dependencies=TRUE)
library(e1071)

#point of Naive Bayes is determine if x is more likely than y
bank.df <- read.csv("C:/Users/valay/OneDrive/Desktop/Ivey/Winter Term 2021/Big Data Analytics/Class 6/Individual Assignment/Bank-1.csv")
head(bank.df)


# Create training and validation sets
selected.var <- c(11, 14, 15)

#from the 1000 data entries, sample will be 60% of that (600)
train.index <- sample(c(1:dim(bank.df)[1]), dim(bank.df)[1]*0.6)
#c means to concatenate

train.df <- bank.df[train.index, selected.var]
head(train.df)
head(train.index)
valid.df <- bank.df[-train.index, selected.var]

# run naive bayes
bank.nb <- naiveBayes(as.factor(Personal_Loan) ~ ., data = train.df)
bank.nb

# predict probabilities
pred.prob <- predict(bank.nb, newdata = valid.df, type = "raw")
head(pred.prob)

# predict class membership
pred.class <- predict(bank.nb, newdata = valid.df)
head(pred.class)
head(valid.df)

#Install a new package-caret-and make functions available
install.packages("caret")
library(caret)

install.packages("ROCR")
library(ROCR)
#Confusion Matrix

classification <- ifelse(pred.prob[,2]>0.1, 1, 0)
actual = valid.df$Personal_Loan
confusionMatrix(as.factor(classification), as.factor(actual))


predObj <- prediction(pred.prob[,2], train.df$Personal_Loan)

rocObj = performance(predObj, measure="tpr", x.measure="fpr")
aucObj = performance(predObj, measure="auc")

plot(rocObj, main = paste("Area under the curve:", round(aucObj@y.values[[1]] ,4)))
```