Noah Sprunk Assignment 3

Code ▾

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
library(ggplot2)
library(lattice)
library(caret)
library(class)
library(ISLR)
library(e1071)
library(gmodels)

Bank <- read.csv("C:\\Users\\Noah\\Downloads\\UniversalBank.csv")

summary(Bank)
```

```
       ID             Age          Experience       Income          ZIP.Code         Family     
CCAvg          Education    
 Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307   Min.   :1.000  
Min.   : 0.000   Min.   :1.000  
 1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911   1st Qu.:1.000  
1st Qu.: 0.700   1st Qu.:1.000  
 Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437   Median :2.000  
Median : 1.500   Median :2.000  
 Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153   Mean   :2.396  
Mean   : 1.938   Mean   :1.881  
 3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608   3rd Qu.:3.000  
3rd Qu.: 2.500   3rd Qu.:3.000  
 Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651   Max.   :4.000  
Max.   :10.000   Max.   :3.000  
    Mortgage     Personal.Loan   Securities.Account  CD.Account            Online          Credit
Card     
 Min.   :  0.0   Min.   :0.000   Min.   :0.0000     Length:5000        Min.   :0.0000   Min.   :
0.000  
 1st Qu.:  0.0   1st Qu.:0.000   1st Qu.:0.0000     Class :character   1st Qu.:0.0000   1st Qu.:
0.000  
 Median :  0.0   Median :0.000   Median :0.0000     Mode  :character   Median :1.0000   Median :
0.000  
 Mean   : 56.5   Mean   :0.096   Mean   :0.1044                        Mean   :0.5968   Mean   :
0.294  
 3rd Qu.:101.0   3rd Qu.:0.000   3rd Qu.:0.0000                        3rd Qu.:1.0000   3rd Qu.:
1.000  
 Max.   :635.0   Max.   :1.000   Max.   :1.0000                        Max.   :1.0000   Max.   :
1.000  
```

Hide

```
Bank$Loan_Category='Accepted'
Bank$Loan_Category[Bank$Personal.Loan < 1]='Declined'

Bank$CreditCard=as.factor(Bank$CreditCard)
Bank$Online=as.factor(Bank$Online)

Bank$Loan_Category=as.factor(Bank$Loan_Category)
Bank$Personal.Loan<-NULL
summary(Bank)
```

```
       ID            Age          Experience        Income          ZIP.Code         Family
CCAvg          Education
 Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307   Min.   :1.000  
Min.   : 0.000   Min.   :1.000  
 1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911   1st Qu.:1.000  
1st Qu.: 0.700   1st Qu.:1.000  
 Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437   Median :2.000  
Median : 1.500   Median :2.000  
 Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153   Mean   :2.396  
Mean   : 1.938   Mean   :1.881  
 3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608   3rd Qu.:3.000  
3rd Qu.: 2.500   3rd Qu.:3.000  
 Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651   Max.   :4.000  
Max.   :10.000   Max.   :3.000  
    Mortgage     Securities.Account  CD.Account        Online   CreditCard  Loan_Category
 Min.   :  0.0   Min.   :0.0000     Length:5000        0:2016   0:3530     Accepted: 480  
 1st Qu.:  0.0   1st Qu.:0.0000     Class :character   1:2984   1:1470     Declined:4520  
 Median :  0.0   Median :0.0000     Mode  :character                                      
 Mean   : 56.5   Mean   :0.1044                                                           
 3rd Qu.:101.0   3rd Qu.:0.0000                                                           
 Max.   :635.0   Max.   :1.0000                                                           
```

Hide

```
Train_Index = createDataPartition(Bank$Loan_Category,p=0.6, list=FALSE)
Train.df=Bank[Train_Index,]
Validation.df=Bank[-Train_Index,]

summary(Train.df)
```

```
       ID            Age          Experience        Income          ZIP.Code         Family     
CCAvg          Education    
 Min.   :   2   Min.   :23.00   Min.   :-3.00   Min.   :  8.00   Min.   : 9307   Min.   :1.000  
Min.   : 0.000   Min.   :1.000  
 1st Qu.:1230   1st Qu.:35.00   1st Qu.:10.00   1st Qu.: 39.00   1st Qu.:91789   1st Qu.:1.000  
1st Qu.: 0.700   1st Qu.:1.000  
 Median :2482   Median :46.00   Median :21.00   Median : 65.00   Median :93460   Median :2.000  
Median : 1.600   Median :2.000  
 Mean   :2485   Mean   :45.51   Mean   :20.27   Mean   : 74.43   Mean   :93132   Mean   :2.391  
Mean   : 1.958   Mean   :1.879  
 3rd Qu.:3736   3rd Qu.:55.00   3rd Qu.:30.00   3rd Qu.: 98.00   3rd Qu.:94609   3rd Qu.:3.000  
3rd Qu.: 2.600   3rd Qu.:3.000  
 Max.   :5000   Max.   :67.00   Max.   :43.00   Max.   :224.00   Max.   :96651   Max.   :4.000  
Max.   :10.000   Max.   :3.000  
    Mortgage      Securities.Account  CD.Account          Online     CreditCard  Loan_Category
 Min.   :  0.00   Min.   :0.0000     Length:3000        0:1198   0:2127     Accepted: 288  
 1st Qu.:  0.00   1st Qu.:0.0000     Class :character   1:1802   1: 873     Declined:2712  
 Median :  0.00   Median :0.0000     Mode  :character                                      
 Mean   : 55.98   Mean   :0.1043                                                           
 3rd Qu.: 98.00   3rd Qu.:0.0000                                                           
 Max.   :635.00   Max.   :1.0000                                                           
```

Hide

```
# Task A

mytable <- xtabs(~ Online+CreditCard+Loan_Category, data=Train.df)
ftable(mytable)
```

```
                  Loan_Category Accepted Declined
Online CreditCard                                
0      0                              81      770
       1                              34      313
1      0                             129     1147
       1                              44      482
```

Hide

```
# Task B
# If we look at the pivot table from task A,
# we can do the math: 41+515=556
# then, our accepted, online, with CC
# 41/556=0.07374 or 7.374%

# Task C

table(Loan_Category=Train.df$Loan_Category, Online=Train.df$Online)
```

```
             Online
Loan_Category    0    1
     Accepted  115  173
     Declined 1083 1629
```

Hide

```
table(Loan_Category=Train.df$Loan_Category, CreditCard=Train.df$CreditCard)
```

```
             CreditCard
Loan_Category    0    1
     Accepted  210   78
     Declined 1917  795
```

Hide

```
#Task D
# P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors)
i = 80/(208+80)
i
```

```
[1] 0.2777778
```

Hide

```
# P(Online = 1 | Loan = 1)
ii = 168/(120+168)
ii
```

```
[1] 0.5833333
```

Hide

```
# P(Loan = 1) (the proportion of loan acceptors)
iii = (120+168)/((1068+1644)+(120+168))
iii
```

```
[1] 0.096
```

Hide

```
# P(CC = 1 | Loan = 0)
iv = 822/(822+1890)
iv
```

```
[1] 0.3030973
```

Hide

```
# P(Online = 1 | Loan = 0)
v = 1644/(1644+1068)
v
```

```
[1] 0.6061947
```

Hide

```
# P(Loan = 0)
vi = (1890+822)/((1890+822)+(208+80))
vi
```

```
[1] 0.904
```

Hide

```
# Task E
E = (i * iii ) / ((80+822) / ((80+822) + (208+1890)))
E
```

```
[1] 0.0886918
```

Hide

```
# Task F
# Which is more accurate, Task B or E?
# I believe that Task B was more accurate because we had to use less
# probabilities in that calculation. It is less accurate because the
# probabilities used may not be entirely independent. We are assuming that
# they are independent but realistically there are correlations
# between these aspects of finances.

# Task G
nb.model<-naiveBayes (Loan_Category~CreditCard+Online, data=Train.df)
To_Predict=data.frame(CreditCard='1', Online='1')
predict(nb.model,To_Predict,type='raw')
```

```
       Accepted  Declined
[1,] 0.08935124 0.9106488
```

Hide

```
# The number that I got in E was 8.87% compared to 8.63% in Task G.
# This shows me that if we used naive Bayes for both but got two
# different answers. We took the raw numbers of accepted vs. declined in
# Task E vs. in Task G we took the training data. They are very close together
# though which tells me that they are both accurate. I think that
# the Task E number is more accurate though because it used the
# raw counts instead of a partition trained data set. But it does tell
# me that the data set is very accurate when it needs to predict an outcome.
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.