Data Mining For Business Applications

**Instructions**

Before starting this homework assignment, please execute the code below. You will need the **tidyverse**, **MASS**, **e1071**, and **readxl** packages. We will be working with the **Telco Customer Churn** data set in this assignment. Please submit your answers as an **R** file that follows the same format as in Homework 2. For the questions in which I ask for your analysis, you can type your answers directly in your **R** file.

**library**(MASS) **library**(tidyverse) **library**(readxl) **library**(e1071)

*# Adjust the path as needed*

*# Telco Customer Churn*

telco <- **read\_excel**(path = "./data/telco customer churn.xlsx")

*# Recode "Churn" variable for classification*

telco <- telco **%>% mutate**(Churn = **recode\_factor**(Churn, "No" = "No", "Yes" = "Yes"))

*# Function for analyzing confusion matrices*

confusion\_matrix\_results <- **function**(table\_matrix, positive\_value) {

pos\_row = **which**(**rownames**(table\_matrix) **==** positive\_value)

pos\_col = **which**(**colnames**(table\_matrix) **==** positive\_value)

TP <- table\_matrix[pos\_row, pos\_col] FN <- table\_matrix[pos\_row, **-**pos\_col] FP <- table\_matrix[**-**pos\_row, pos\_col] TN <- table\_matrix[**-**pos\_row, **-**pos\_col]

results <-

**data.frame**(correct = **c**(TN **+** TP, **round**((TN **+** TP)**/sum**(table\_matrix),3)), misclassified = **c**(FN **+** FP, **round**((FN **+** FP)**/sum**(table\_matrix),3)), true\_pos = **c**(TP, **round**(TP**/sum**(table\_matrix[pos\_row,]),3)), false\_neg = **c**(FN, **round**(FN**/sum**(table\_matrix[pos\_row,]),3)), true\_neg = **c**(TN, **round**(TN**/sum**(table\_matrix[**-**pos\_row,]),3)), false\_pos = **c**(FP, **round**(FP**/sum**(table\_matrix[**-**pos\_row,]),3)))

**rownames**(results) <- **c**("Observations", "Rate")

**return**(results)

}

Write **one** expression linked by %>% operators that uses **dyplr** functions to produce the data frame below.

**Hint**: remember that the sum of a logical vector returns the total number of TRUE’s.

# A tibble: 4 x 5

PaymentMethod customers avg\_tenure avg\_monthly\_charges churn\_yes\_rate

<chr> <int> <dbl> <dbl> <dbl>

1 Bank transfer (automatic) 1544 43.65674 67.19265 0.1670984

2 Credit card (automatic) 1522 43.26938 66.51239 0.1524310

3 Electronic check 2365 25.17463 76.25581 0.4528541

4 Mailed check 1612 21.83002 43.91706 0.1910670

**Part (b) 2.5 Points**

Write **one** expression linked by %>% operators that uses **dyplr** and **tidyr** functions to produce the data frame below. This table displays the proportion of customers who left the company service (Churn = “Yes”) by *Contract* and *PaymentMethod*.

# A tibble: 4 x 4

PaymentMethod `Month-to-month` `One year` `Two year`

\* <chr> <dbl> <dbl> <dbl>

1 Bank transfer (automatic) 0.3412564 0.09718670 0.033687943

2 Credit card (automatic) 0.3278085 0.10301508 0.022375215

3 Electronic check 0.5372973 0.18443804 0.077380952

4 Mailed check 0.3157895 0.06824926 0.007853403

Use **ggplot** to produce the plot below.

Tensure Months by Contract and Payment Type for Customers With Internet Service

DSL Fiber optic

|  |  |  |
| --- | --- | --- |
|  | |  |
|  | |  |
|  |
|  |  | |

Bank transfer (automatic)

60

40

20

0

60

40

20

Churn

Credit card (automatic)

0

No

Yes

|  |  |
| --- | --- |
|  | |
|  | |
|  | |
|  |  |
|  | |
|  | |
|  | |

Electronic check

60

TenureMonths

40

|  |  |  |
| --- | --- | --- |
|  | |  |
|  | |  |
|  |  | |

20

0

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
|  | |  |  |
|  | |  | |
|  |  |  | |

60

40

20

Mailed check

0

Month−to−month One year Two year Month−to−month One year Two year

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | | |
|  | |  | |
|  | |  |  |
|  | |  |  |

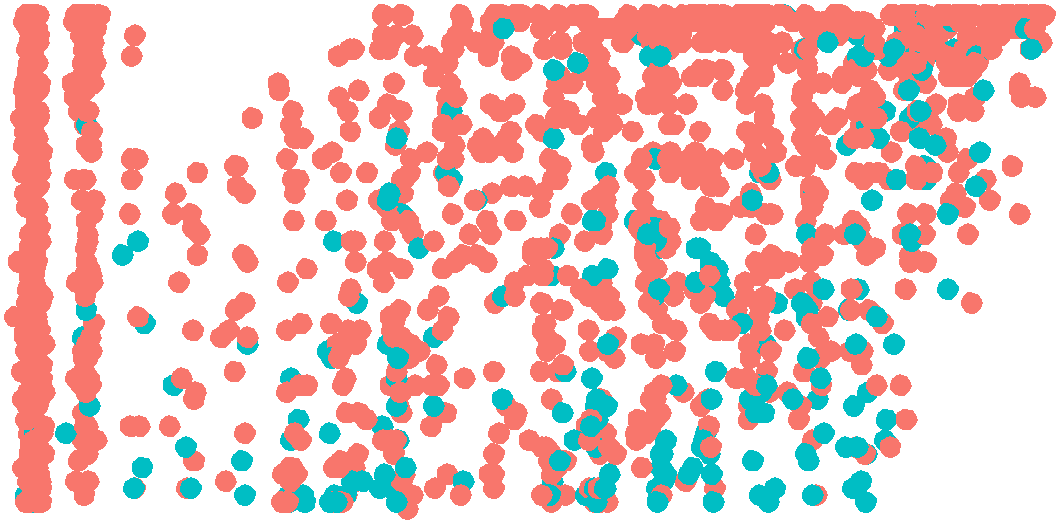
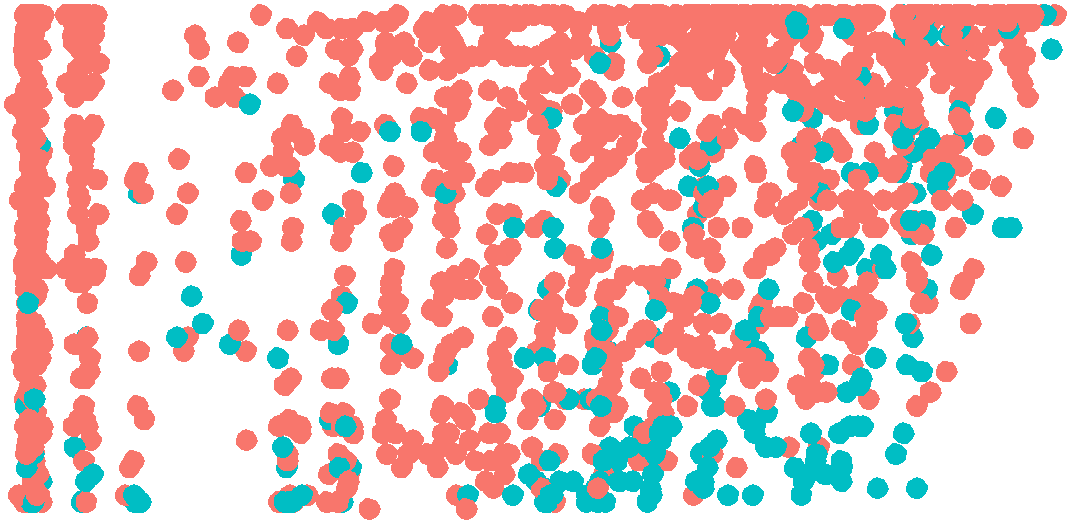
Contract

**Part (b) 2.5 Points**

Use **ggplot** to produce the plot below.

Tensure Months vs Monthly Charges by Payment Type for All Customers

Bank transfer (automatic) Credit card (automatic)



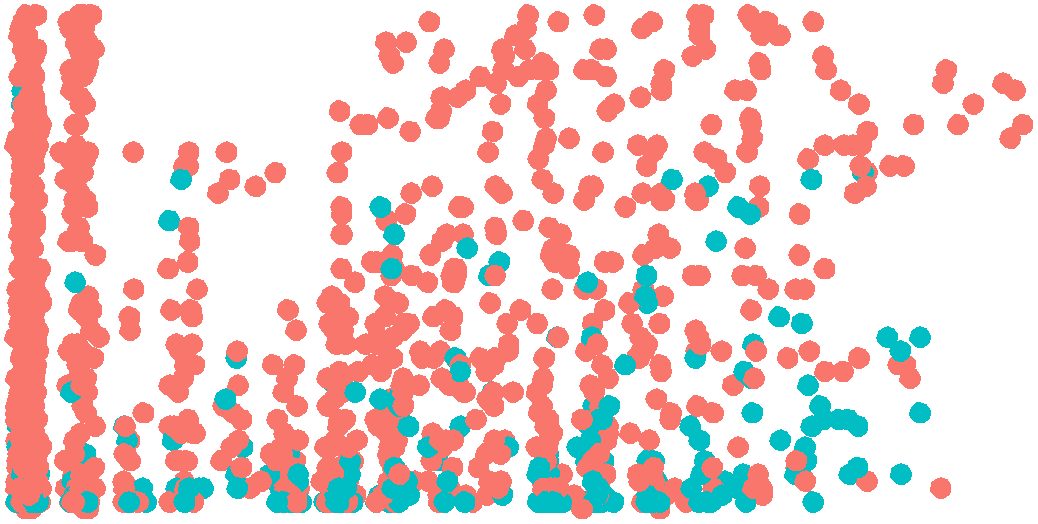
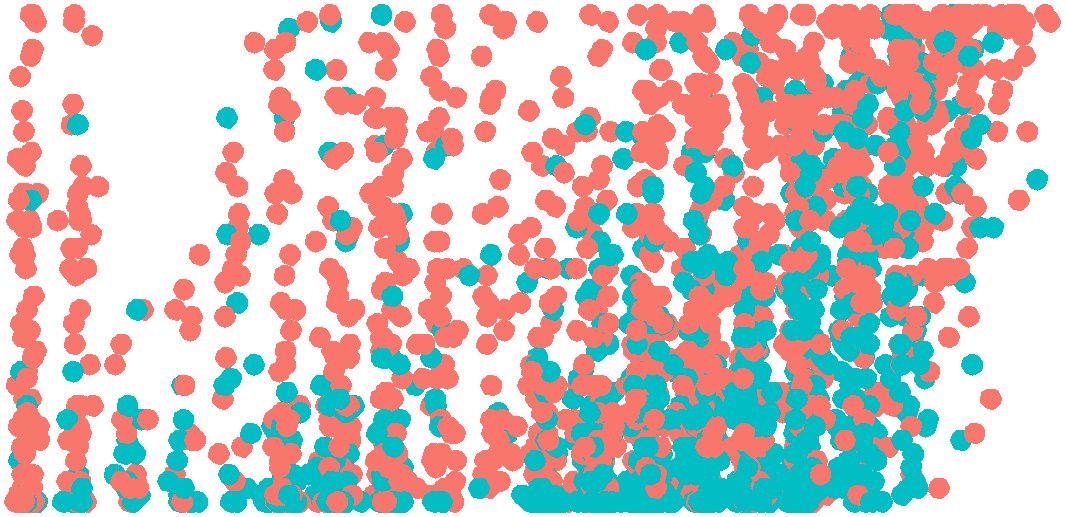
60

40

20

0

Electronic check Mailed check



Churn

No

Yes

60

TenureMonths

40

20

0

25 50 75 100 25 50 75 100

MonthlyCharges

**Problem 3 (5 Points)**

**Part (a) 2.5 Points**

Fit a logistic regression model to predict the probability that a customer will leave the company (**Churn** = “Yes”) using **Contract**, **PaymentMethod**, **MonthlyCharges**, and **TenureMonths**. Use **R** to obtain the answer to the following question:

How many times greater are the odds that a customer with an electronic check payment method, month-to- month contract, $80 monthly charge, and 30 month tenure will leave the company (Churn = “Yes”) versus a customer with a mailed check payment method, month-to-month contract, $80 monthly charge, and 30 month tenure?

**Part (b) 2.5 Points**

Use **R** to produce a data frame with a subset of the **telco** data frame variables used in the logistic regression model. Name this data frame **logistic\_subset**. Add a variable to **logistic\_subset** that has predicted **Churn** values based on the logistic regression model. Use a cut-off probability of 0.4 to classify customers as “Yes”.

**Problem 4 (5 Points)**

Fit a Naive Bayes model to predict the probability that a customer will leave the company (**Churn** = “Yes”)

using **Contract**, **PaymentMethod**, **MonthlyCharges**, and **TenureMonths**.

Use **R** to produce a data frame with a subset of the **telco** data frame variables used in the Naive Bayes model. Name this data frame **naive\_subset**. Add a variable to **naive\_subset** that has predicted **Churn** values based on the Naive Bayes model. Use a cut-off probability of 0.4 to classify customers as “Yes”.

**Problem 5 (5 Points)**

Use the **confusion\_matrix\_results()** function to compare the two models. Discuss which model you think has better performance and why. Make sure to mention important error rates in your discussion. Which model would you choose to use in presenting results to decision makers at Telco? The confusion matrix results below are what you should get for the two different models.

correct misclassified true\_pos false\_neg true\_neg false\_pos

Observations 5475.000 1568.000 1171.000 698.000 4304.000 870.000

Rate 0.777 0.223 0.627 0.373 0.832 0.168

correct misclassified true\_pos false\_neg true\_neg false\_pos

Observations 5268.000 1775.000 1375.000 494.000 3893.000 1281.000

Rate 0.748 0.252 0.736 0.264 0.752 0.248