**D209 Data Mining 1**

Nicole Reiswig

College of Information Technology, Western Governors University

Festus Elleh

January 14, 2024

**Table of Contents**

**Part I: Research Question**

A.

1.  Research question

2.  Data analysis goal

**Part II: Method Justification**

B.

1.  Classification method expected outcomes

2.  Classification method assumption

3.  Packages and libraries

**Part III: Data Preparation**

C.

1.  Goal

2.  Variables

3.  Data preparation

4.  Cleaned data set

**Part IV: Analysis**

D.

1.  Split data, training, and test set

2.  Analysis technique

3.  Code

**Part V: Data Summary and Implications**

E.

1.  AUC

2.  Results and implications

3.  Limitation

4.  Recommendation

**Part VI: Demonstration**

F.  Panopto video

G.  Sources

# Analysis

**A.  Part I: Research Question**

**1.** The research question that is to be answered in this data mining report is, “What are the predictor variables in predicting Churn from the telecom churn database?” The research question will be answered utilizing the Naïve Bayes classification method.

**2.** The goal of this data mining report and classification analysis is to identify major predictors of churn, determine which factors it is based on, and determine if churn can be predicated based on the trained model. This project aims to forecast churn risk factors so that appropriate actions can be taken to mitigate risk. In this project, we will clean, evaluate, and create a supervised machine-learning model.

**B. Part II: Method Justification**

**1.** Naive Bayes is a classification method that is used to analyze the research question, “What are the predictor variables in predicting Churn from the telecom churn database?” Naïve Bayes was chosen as the best method for this analysis because it is most suitable for large datasets when all categorical parameters are independent. Naïve Bayes will be used on a telecom churn dataset containing 10,000 observations and 50 variables. According to Elleh, F. (2023), when the calibration sample size increases Naïve Bayes classifiers outperform other highly advanced classification methods in terms of classification accuracy. Naïve Bayes classification is based on Bayes Theorem which combines previous knowledge with new information obtained from observed data. Bayes Theorem can be used to identify which class, churn or no churn, is most likely for the provided predictor variables.

**2.** Assumptions of Naïve Bayes classification are that it is assumed that variables are independent, is easy to build, most useful when used on large datasets, and is known to surpass other highly advanced classification systems.

**3.** The programming language utilized to perform this classification analysis is the R programming language in R Studios. The libraries and packages used are tidyverse which is essential to data science collection, ggplot2 which is used to generate plots, dplyr which is essential for data frame manipulation, naivebayes which is used for the Naïve Bayes model, psych which is used for graphing variable independence, caret which is a predictive model package, e1071 which is required for confusion matrix, ROCR which is required to create a performance curve, rpart which is used for building classification and regression trees, data.table used for reading and manipulation of data, glmnet for regression, caret for modeling, xgboost for building models, cowplot for combining multiple plots, and pROC used for auc and roc models.

**C. Part III: Data Preparation**

**1.** A preprocessing goal of this analysis is to encode the many categorical variables from yes/no to 0/1 utilizing dummy variables to fit the Naive Bayes Model for analysis.

**2.**  The dataset variables in the telecom churn database utilized to answer the research question in section A1 utilizing the Naïve Bayes classification method are as follows, see data class for classification of categorical or numerical variable type:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Data Type** | **Description** | **Example** | **Data Class** |
| Case order | Qualitative | Placeholder variable | 1 | Categorical |
| Customer ID | Qualitative | Unique customer id | K409198 | Categorical |
| Interaction uid | Qualitative | Transaction id | Aa90260b-4141… | Categorical |
| City | Qualitative | Customer residence | Point Baker | Categorical |
| State | Qualitative | Customer residence | Arkansas | Categorical |
| County | Qualitative | Customer residence | Cook | Categorical |
| Zip | Qualitative | Customer residence | 99362 | Categorical |
| Lat | Qualitative | GPS coordinate | 56.251 | Categorical |
| Lng | Qualitative | GPS coordinate | -133.376 | Categorical |
| Population | Quantitative | Population within 1mile of customer | 38 | Numeric |
| Area | Qualitative | Area type | Rural | Categorical |
| Timezone | Qualitative | Customer timezone | America/Chicago | Categorical |
| Job | Qualitative | Customer job | Engineer | Categorical |
| Children | Quantitative | # of children in household | 2 | Numeric |
| Age | Quantitative | Age of customer | 30 | Numeric |
| Education | Qualitative | Customer education | Masters | Categorical |
| Employment | Qualitative | Customer employment | Employed | Categorical |
| Income | Quantitative | Income of customer | 28561 | Numeric |
| Marital | Qualitative | Customer marital status | Single | Categorical |
| Gender | Qualitative | Gender of customer | Female | Categorical |
| Churn | Qualitative | If customer d/c service | Yes | Categorical |
| Outage sec per week | Quantitative | Avg # of outage seconds a week | 6.9 | Numeric |
| Email | Quantitative | # of emails sent to customer in week | 5 | Numeric |
| Contacts | Quantitative | # of times customer contacted tech support | 0 | Numeric |
| Yearly equip failure | Quantitative | # of customer equipment failures a year | 0 | Numeric |
| Techie | Qualitative | Customer is technically inclined | Yes | Categorical |
| Contract | Qualitative | Contract terms | One year | Categorical |
| Port modem | Qualitative | Customer has port modem | Yes | Categorical |
| Tablet | Qualitative | Customer has tablet | Yes | Categorical |
| Internet service | Qualitative | Customers internet service provider | Fiber optics | Categorical |
| Phone | Qualitative | Customer has phone service | Yes | Categorical |
| Multiple | Qualitative | Customer has multiple lines | Yes | Categorical |
| Online security | Qualitative | Customer has add on security | Yes | Categorical |
| Online backup | Qualitative | Customer has add on backup | Yes | Categorical |
| Device protection | Qualitative | Customer has device protection | yes | Categorical |
| Tech support | Qualitative | Customer has add on tech support | Yes | Categorical |
| Streaming tv | Qualitative | Customer has streaming tv | Yes | Categorical |
| Streaming movies | Qualitative | Customer has streaming movies | Yes | Categorical |
| Paperless billing | Qualitative | Customer has paperless billing | Yes | Categorical |
| Payment method | Qualitative | Customers payment method | Check | Categorical |
| Tenure | Quantitative | # of months customer has been with provider | 12 | Numeric |
| Monthly charge | Quantitative | Amount of monthly charge | 171.45 | Numeric |
| Bandwidth gb year | Quantitative | Amount of gb used per year | 904 | Numeric |
| Timely response | Quantitative | Survey response rating importance of timely response | 1 | Numeric |
| Timely fixes | Quantitative | Survey response rating importance of timely fixes | 2 | Numeric |
| Timely replacements | Quantitative | Survey response rating importance of timely replacements | 3 | Numeric |
| Reliability | Quantitative | Survey response rating importance of reliability | 4 | Numeric |
| Options | Quantitative | Survey response rating importance of having options | 5 | Numeric |
| Respectful response | Quantitative | Survey response rating importance of respectful response | 6 | Numeric |
| Courteous exchange | Quantitative | Survey response rating importance of courteous exchange | 7 | Numeric |
| Evidence of active listening | Quantitative | Survey response rating importance of active listening | 8 | Numeric |

**3.** Steps to prepare the dataset for the analysis:

* + Import CSV data file into R studio
  + Profile data utilizing str()
  + Check dimensions of data utilizing dim()
  + Load packages needed for each step of the analysis
  + View data with str() and glimpse()
  + Check for duplicates using duplicated()
  + Check for missing values using colSums(is.na()) or vismis()
  + Rename columns with misspellings or awkward names
  + Explore data using univariate and bivariate visuals such as histograms, boxplots, bar plots and scatterplots
  + When appropriate, impute, retain, or exclude values
  + Check the correlation of variables
  + Encode yes/no variables with dummy variables as 0/1
  + Export the clean data to CSV. See attached CSV of clean data.

#Check the working directory

getwd()

#Data profiling

str("~/MSDA/churn\_clean")

#Dimensions of data

dim(churn\_clean)

#1000 50

# load packages

#for reading and manipulation of data

library(data.table)

# used for data manipulation and joining

library(dplyr)

# used for regression

library(glmnet)

# used for plotting

library(ggplot2)

# used for modeling

library(caret)

# used for building XGBoost model

library(xgboost)

# used for skewness

library(e1071)

# used for combining multiple plots

library(cowplot)

#collection of r packages for data science

library(tidyverse)

#Perform Naive Bayes Classification

library(naivebayes)

#

library(psych)

#Tools for training regression and classification models

library(caret)

#

library(ROCR)

#

library(rpart)

#

library(pROC)

#View data

glimpse(churn\_clean)

str(churn\_clean)

#Detect duplicates

duplicated("~/MSDA/churn\_clean")

#No duplicates

#Detect missing values

colSums(is.na(churn\_clean))

#No missing values

#Rename column names item 1- 8 [in-text citation: (Zach, 2022)]

colnames(churn\_clean)[colnames(churn\_clean) == 'Item1'] <- 'Timely\_Response'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item2'] <- 'Timely\_Fixes'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item3'] <- 'Timely\_Replacements'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item4'] <- 'Reliability'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item5'] <- 'Options'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item6'] <- 'Respectful\_Response'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item7'] <- 'Courteous\_Exchange'

colnames(churn\_clean)[colnames(churn\_clean) == 'Item8'] <- 'Active\_Listening'

#Verify columns renamed successfully

glimpse(churn\_clean)

#Cross tabulation

xtabs(~ Churn + Gender, data = churn\_clean)

xtabs(~ Churn + Age, data = churn\_clean)

xtabs(~ Churn + Children, data = churn\_clean)

xtabs(~ Churn + Area, data = churn\_clean)

xtabs(~ Churn + TimeZone, data = churn\_clean)

xtabs(~ Churn + State, data = churn\_clean)

xtabs(~ Churn + Marital, data = churn\_clean)

xtabs(~ Churn + Email, data = churn\_clean)

xtabs(~ Churn + Contacts, data = churn\_clean)

xtabs(~ Churn + Yearly\_equip\_failure, data = churn\_clean)

xtabs(~ Churn + Techie, data = churn\_clean)

xtabs(~ Churn + Contract, data = churn\_clean)

xtabs(~ Churn + Port\_modem, data = churn\_clean)

xtabs(~ Churn + Tablet, data = churn\_clean)

xtabs(~ Churn + InternetService, data = churn\_clean)

xtabs(~ Churn + Phone, data = churn\_clean)

xtabs(~ Churn + Multiple, data = churn\_clean)

xtabs(~ Churn + OnlineSecurity, data = churn\_clean)

xtabs(~ Churn + OnlineBackup, data = churn\_clean)

xtabs(~ Churn + DeviceProtection, data = churn\_clean)

xtabs(~ Churn + TechSupport, data = churn\_clean)

xtabs(~ Churn + StreamingTV, data = churn\_clean)

xtabs(~ Churn + StreamingMovies, data = churn\_clean)

xtabs(~ Churn + PaperlessBilling, data = churn\_clean)

xtabs(~ Churn + PaymentMethod, data = churn\_clean)

xtabs(~ Churn + Timely\_Response, data = churn\_clean)

xtabs(~ Churn + Timely\_Fixes, data = churn\_clean)

xtabs(~ Churn + Timely\_Replacements, data = churn\_clean)

xtabs(~ Churn + Reliability, data = churn\_clean)

xtabs(~ Churn + Options, data = churn\_clean)

xtabs(~ Churn + Respectful\_Response, data = churn\_clean)

xtabs(~ Churn + Courteous\_Exchange, data = churn\_clean)

xtabs(~ Churn + Active\_Listening, data = churn\_clean)

#Visualizations

#Boxplot of each variable - detect outliers - Univariate Graphs of each variable

boxplot(churn\_clean$CaseOrder, xlab = "Case Order")

boxplot(churn\_clean$Zip, xlab = "Zip Code")

boxplot(churn\_clean$Lat, xlab = "Lat")

boxplot(churn\_clean$Lng, xlab = "Lng")

boxplot(churn\_clean$Population, xlab = "Population")

boxplot(churn\_clean$Children, xlab = "Children")

boxplot(churn\_clean$Age, xlab = "Age")

boxplot(churn\_clean$Income, xlab = "Income")

boxplot(churn\_clean$Outage\_sec\_perweek, xlab = "Outage Sec Per Week")

boxplot(churn\_clean$Email, xlab = "Email")

boxplot(churn\_clean$Contacts, xlab = "Contacts")

boxplot(churn\_clean$Yearly\_equip\_failure, xlab = "Yearly Equip Failures")

boxplot(churn\_clean$Tenure, xlab = "Tenure")

boxplot(churn\_clean$MonthlyCharge, xlab = "Monthly Charge")

boxplot(churn\_clean$Bandwidth\_GB\_Year, xlab = "Bandwidth GB Year")

#Count and range of variables with outliers

Children <- churn\_clean[which(churn\_clean$Children > 7), ]

str(Children)

Income <- churn\_clean[which(churn\_clean$Income > 100000), ]

str(Income)

Outage <- churn\_clean[which(churn\_clean$Outage\_sec\_perweek > 20), ]

str(Outage)

Outage2 <- churn\_clean[which(churn\_clean$Outage\_sec\_perweek < 0), ]

str(Outage2)

Email <- churn\_clean[which(churn\_clean$Email > 20), ]

str(Email)

Email2 <- churn\_clean[which(churn\_clean$Email < 4), ]

str(Email2)

Contacts <- churn\_clean[which(churn\_clean$Contacts > 5), ]

str(Contacts)

Equipment <- churn\_clean[which(churn\_clean$Yearly\_equip\_failure > 2), ]

str(Equipment)

Monthly <- churn\_clean[which(churn\_clean$MonthlyCharge > 300), ]

str(Monthly)

Timely\_Response <- churn\_clean[which(churn\_clean$Timely\_Response > 5), ]

str(Timely\_Response)

Timely\_Response1 <- churn\_clean[which(churn\_clean$Timely\_Response < 2), ]

str(Timely\_Response1)

Timely\_Fixes <- churn\_clean[which(churn\_clean$Timely\_Fixes > 5), ]

str(Timely\_Fixes)

Timely\_Fixes2 <- churn\_clean[which(churn\_clean$Timely\_Fixes < 2), ]

str(Timely\_Fixes2)

Timely\_Replacements <- churn\_clean[which(churn\_clean$Timely\_Replacements > 5), ]

str(Timely\_Replacements)

Timely\_Replacements3 <- churn\_clean[which(churn\_clean$Timely\_Replacements < 2), ]

str(Timely\_Replacements3)

Reliability <- churn\_clean[which(churn\_clean$Reliability > 5), ]

str(Reliability)

Reliability4 <- churn\_clean[which(churn\_clean$Reliability < 2), ]

str(Reliability4)

Options <- churn\_clean[which(churn\_clean$Options > 5), ]

str(Options)

Options5 <- churn\_clean[which(churn\_clean$Options < 2), ]

str(Options5)

Respectful\_Response <- churn\_clean[which(churn\_clean$Respectful\_Response > 5), ]

str(Respectful\_Response)

Respectful\_Response6 <- churn\_clean[which(churn\_clean$Respectful\_Response < 2), ]

str(Respectful\_Response6)

Courteous\_Exchange <- churn\_clean[which(churn\_clean$Courteous\_Exchange > 5), ]

str(Courteous\_Exchange)

Courteous\_Exchange7 <- churn\_clean[which(churn\_clean$Courteous\_Exchange < 2), ]

str(Courteous\_Exchange7)

Active\_Listening <- churn\_clean[which(churn\_clean$Active\_Listening > 5), ]

str(Active\_Listening)

Active\_Listening8 <- churn\_clean[which(churn\_clean$Active\_Listening < 2), ]

str(Active\_Listening8)

#Summary statistics

summary(churn\_clean)

#Histograms of variables with outliers- Univariate graphs

hist(churn\_clean$Population, col = 'turquoise', main = "Population")

hist(churn\_clean$Children, col = 'turquoise', main = "Children")

hist(churn\_clean$Income, col = 'turquoise', main = "Income")

hist(churn\_clean$Outage\_sec\_perweek, col = 'turquoise', main = "Outage Sec Perweek")

hist(churn\_clean$Email, col = 'turquoise', main = "Email")

hist(churn\_clean$Contacts, col = 'turquoise', main = "Contacts")

hist(churn\_clean$Yearly\_equip\_failure, col = 'turquoise', main = "Yearly Equip Failure")

hist(churn\_clean$Timely\_Response, col = 'turquoise', main = "Timely Response")

hist(churn\_clean$Timely\_Fixes, col = 'turquoise', main = "Timely Fixes")

hist(churn\_clean$Timely\_Replacements, col = 'turquoise', main = "Timely Replacements")

hist(churn\_clean$Reliability, col = 'turquoise', main = "Reliability")

hist(churn\_clean$Options, col = 'turquoise', main = "Options")

hist(churn\_clean$Respectful\_Response, col = 'turquoise', main = "Respectful Response")

hist(churn\_clean$Courteous\_Exchange, col = 'turquoise', main = "Courteous Exchange")

hist(churn\_clean$Active\_Listening, col = 'turquoise', main = "Active Listening")

#Explore data variables- univariate graphs [in-text citation: (R programming 101, n.d.)]

barplot(sort(table(churn\_clean$Area)), col = 'blue', main = "Area")

barplot(sort(table(churn\_clean$TimeZone)), col = 'blue', main = "Timezone")

barplot(sort(table(churn\_clean$Children)), col = 'blue', main = "Number of Children")

barplot(sort(table(churn\_clean$Age)), col = 'blue', main = "Age")

barplot(sort(table(churn\_clean$Income)), col = 'blue', main = "Income")

barplot(sort(table(churn\_clean$Marital)), col = 'blue', main = "Marital")

barplot(sort(table(churn\_clean$Gender)), col = 'blue', main = "Gender")

barplot(sort(table(churn\_clean$Churn)), col = 'blue', main = "Churn")

barplot(sort(table(churn\_clean$Outage\_sec\_perweek)), col = 'blue', main = "Outage Sec Per Week")

barplot(sort(table(churn\_clean$Email)), col = 'blue', main = "Email")

barplot(sort(table(churn\_clean$Contacts)), col = 'blue', main = "Contacts")

barplot(sort(table(churn\_clean$Yearly\_equip\_failure)), col = 'blue', main = "Yearly Equipment Failure")

barplot(sort(table(churn\_clean$Techie)), col = 'blue', main = "Techie")

barplot(sort(table(churn\_clean$Contract)), col = 'blue', main = "Contracts")

barplot(sort(table(churn\_clean$Port\_modem)), col = 'blue', main = "Port Modem")

barplot(sort(table(churn\_clean$Tablet)), col = 'blue', main = "Tablet")

barplot(sort(table(churn\_clean$InternetService)), col = 'blue', main = "Internet Service")

barplot(sort(table(churn\_clean$Phone)), col = 'blue', main = "Phone")

barplot(sort(table(churn\_clean$Multiple)), col = 'blue', main = "Multiple")

barplot(sort(table(churn\_clean$OnlineSecurity)), col = 'blue', main = "Online Security")

barplot(sort(table(churn\_clean$OnlineBackup)), col = 'blue', main = "Online Backup")

barplot(sort(table(churn\_clean$DeviceProtection)), col = 'blue', main = "Device Protection")

barplot(sort(table(churn\_clean$TechSupport)), col = 'blue', main = "Tech Support")

barplot(sort(table(churn\_clean$StreamingTV)), col = 'blue', main = "Streaming TV")

barplot(sort(table(churn\_clean$StreamingMovies)), col = 'blue', main = "Streaming Movies")

barplot(sort(table(churn\_clean$PaperlessBilling)), col = 'blue', main = "Paperless Billing")

barplot(sort(table(churn\_clean$PaymentMethod)), col = 'blue', main = "Payment Method")

barplot(sort(table(churn\_clean$Tenure)), col = 'blue', main = "Tenure")

barplot(sort(table(churn\_clean$MonthlyCharge)), col = 'blue', main = "Monthly Charge")

barplot(sort(table(churn\_clean$Bandwidth\_GB\_Year)), col = 'blue', main = "Bandwidth GB Year")

barplot(sort(table(churn\_clean$Timely\_Response)), main = "Timely Response")

barplot(sort(table(churn\_clean$Timely\_Fixes)), col = 'blue', main = "Timely Fixes")

barplot(sort(table(churn\_clean$Timely\_Replacements)), col = 'blue', main = "Timely Replacements")

barplot(sort(table(churn\_clean$Reliability)), col = 'blue', main = "Reliability")

barplot(sort(table(churn\_clean$Options)), col = 'blue', main = "Options")

barplot(sort(table(churn\_clean$Respectful\_Response)), col = 'blue', main = "Respecftful Response")

barplot(sort(table(churn\_clean$Courteous\_Exchange)), col = 'blue', main = "Courteous Exchange")

barplot(sort(table(churn\_clean$Active\_Listening)), col = 'blue', main = "Active Listening")

#BiVariate Graph of each variable

library(ggplot2)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Area)) +

geom\_point(color = "coral", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

#ggplot(churn\_clean, aes(x = Churn, y = Area, color = Churn)) +

#geom\_point()

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Contract)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Gender)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Marital)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = InternetService)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = PaymentMethod)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Tenure)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Techie)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Port\_modem)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Tablet)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Phone)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = Multiple)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = OnlineSecurity)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = OnlineBackup)) +

geom\_point(color = "cornflowerblue", alpha = .7,

size = 2) + geom\_smooth(method = "lm", se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = DeviceProtection)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = TechSupport)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = StreamingTV)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = StreamingMovies)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

ggplot(data = churn\_clean,

mapping = aes(x = Churn, y = PaperlessBilling)) +

geom\_point(color = "cornflowerblue", alpha = .7, size = 2) +

geom\_smooth(method = "lm", se = FALSE, linewidth = 1.5)

#Encode categorical variables [in-text citation: (Statistics Globe, n.d.)]

#Encode Churn

churn\_clean$Churn <- as.character(churn\_clean$Churn)

churn\_clean$Churn[churn\_clean$Churn == "Yes"] <- 1

churn\_clean$Churn[churn\_clean$Churn == "No"] <- 0

churn\_clean$Churn <- as.factor(churn\_clean$Churn)

#Encode Techie

churn\_clean$Techie <- as.character(churn\_clean$Techie)

churn\_clean$Techie[churn\_clean$Techie == "Yes"] <- 1

churn\_clean$Techie[churn\_clean$Techie == "No"] <- 0

churn\_clean$Techie <- as.factor(churn\_clean$Techie)

#Encode Port Modem

churn\_clean$Port\_modem <- as.character(churn\_clean$Port\_modem)

churn\_clean$Port\_modem[churn\_clean$Port\_modem == "Yes"] <- 1

churn\_clean$Port\_modem[churn\_clean$Port\_modem == "No"] <- 0

churn\_clean$Port\_modem <- as.factor(churn\_clean$Port\_modem)

#Encode Port Tablet

churn\_clean$Tablet <- as.character(churn\_clean$Tablet)

churn\_clean$Tablet[churn\_clean$Tablet == "Yes"] <- 1

churn\_clean$Tablet[churn\_clean$Tablet == "No"] <- 0

churn\_clean$Tablet <- as.factor(churn\_clean$Tablet)

#Encode Phone

churn\_clean$Phone <- as.character(churn\_clean$Phone)

churn\_clean$Phone[churn\_clean$Phone == "Yes"] <- 1

churn\_clean$Phone[churn\_clean$Phone == "No"] <- 0

churn\_clean$Phone <- as.factor(churn\_clean$Phone)

#Encode Multiple

churn\_clean$Multiple <- as.character(churn\_clean$Multiple)

churn\_clean$Multiple[churn\_clean$Multiple == "Yes"] <- 1

churn\_clean$Multiple[churn\_clean$Multiple == "No"] <- 0

churn\_clean$Multiple <- as.factor(churn\_clean$Multiple)

#Encode Online Security

churn\_clean$OnlineSecurity <- as.character(churn\_clean$OnlineSecurity)

churn\_clean$OnlineSecurity[churn\_clean$OnlineSecurity == "Yes"] <- 1

churn\_clean$OnlineSecurity[churn\_clean$OnlineSecurity == "No"] <- 0

churn\_clean$OnlineSecurity <- as.factor(churn\_clean$OnlineSecurity)

#Encode Online Backup

churn\_clean$OnlineBackup <- as.character(churn\_clean$OnlineBackup)

churn\_clean$OnlineBackup[churn\_clean$OnlineBackup == "Yes"] <- 1

churn\_clean$OnlineBackup[churn\_clean$OnlineBackup == "No"] <- 0

churn\_clean$OnlineBackup <- as.factor(churn\_clean$OnlineBackup)

#Encode Device Protection

churn\_clean$DeviceProtection <- as.character(churn\_clean$DeviceProtection)

churn\_clean$DeviceProtection[churn\_clean$DeviceProtection == "Yes"] <- 1

churn\_clean$DeviceProtection[churn\_clean$DeviceProtection == "No"] <- 0

churn\_clean$DeviceProtection <- as.factor(churn\_clean$DeviceProtection)

#Encode Tech Support

churn\_clean$TechSupport <- as.character(churn\_clean$TechSupport)

churn\_clean$TechSupport[churn\_clean$TechSupport == "Yes"] <- 1

churn\_clean$TechSupport[churn\_clean$TechSupport == "No"] <- 0

churn\_clean$TechSupport <- as.factor(churn\_clean$TechSupport)

#Encode Streaming TV

churn\_clean$StreamingTV <- as.character(churn\_clean$StreamingTV)

churn\_clean$StreamingTV[churn\_clean$StreamingTV == "Yes"] <- 1

churn\_clean$StreamingTV[churn\_clean$StreamingTV == "No"] <- 0

churn\_clean$StreamingTV <- as.factor(churn\_clean$StreamingTV)

#Encode Streaming Movies

churn\_clean$StreamingMovies <- as.character(churn\_clean$StreamingMovies)

churn\_clean$StreamingMovies[churn\_clean$StreamingMovies == "Yes"] <- 1

churn\_clean$StreamingMovies[churn\_clean$StreamingMovies == "No"] <- 0

churn\_clean$StreamingMovies <- as.factor(churn\_clean$StreamingMovies)

#Encode Paperless Billing

churn\_clean$PaperlessBilling <- as.character(churn\_clean$PaperlessBilling)

churn\_clean$PaperlessBilling[churn\_clean$PaperlessBilling == "Yes"] <- 1

churn\_clean$PaperlessBilling[churn\_clean$PaperlessBilling == "No"] <- 0

churn\_clean$PaperlessBilling <- as.factor(churn\_clean$PaperlessBilling)

#One-Hot Encoding Area

library(fastDummies)

A\_treat <- dummy\_cols(churn\_clean, select\_columns = "Area")

#One-Hot Encoding Marital

M\_treat <- dummy\_cols(churn\_clean, select\_columns = "Marital")

#One-Hot Encoding Gender

G\_treat <- dummy\_cols(churn\_clean, select\_columns = "Gender")

#One-Hot Encoding Contract

C\_treat <- dummy\_cols(churn\_clean, select\_columns = "Contract")

#One-Hot Encoding InternetService

I\_treat <- dummy\_cols(churn\_clean, select\_columns = "InternetService")

#One-Hot Encoding PaymentMethod

P\_treat <- dummy\_cols(churn\_clean, select\_columns = "PaymentMethod")

library(tidyverse)

#Put all df in a list [in-text citation: (Zach, 2021)]

df\_list <- list(churn\_clean, A\_treat, M\_treat, G\_treat, C\_treat, I\_treat, P\_treat)

#Merge all data frames together using mutate[in-text citation: (Zach, 2021)]

library (dplyr)

Mutate\_Churn <- mutate(churn\_clean, A\_treat, C\_treat, G\_treat, I\_treat, M\_treat, P\_treat)

#Drop columns by name that were duplicates to the variables one-hot encoded

Mutate\_Churn1 <- subset(Mutate\_Churn, select = -c(Area, Contract, Gender, InternetService, Marital, PaymentMethod))

#Rename columns with unexpected symbol

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'Marital\_Never Married'] <- 'Marital\_NeverMarried'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'Contract\_One year'] <- 'Contract\_OneYear'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'Contract\_Two Year'] <- 'Contract\_TwoYear'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'Contract\_Month-to-month'] <- 'Contract\_Month\_To\_Month'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'InternetService\_Fiber Optic'] <- 'InternetService\_FiberOptic'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'PaymentMethod\_Bank Transfer(automatic)'] <- 'PaymentMethod\_BankTransfer'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'PaymentMethod\_Credit Card (automatic)'] <- 'PaymentMethod\_CreditCard'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'PaymentMethod\_Electronic Check'] <- 'PaymentMethod\_ElectronicCheck'

colnames(Mutate\_Churn1)[colnames(Mutate\_Churn1) == 'PaymentMethod\_Mailed Check'] <- 'PaymentMethod\_MailedCheck'

#Reduce Model Payment Method, Marital as cardinality is too great and we do not want to proliferate

Mutate\_Churn2 <- subset(Mutate\_Churn1, select = -c(Job, TimeZone, Population,

Lat, Lng, Zip, County, State,

City, UID, Interaction, Customer\_id,

PaymentMethod\_BankTransfer,

PaymentMethod\_CreditCard,

PaymentMethod\_ElectronicCheck,

PaymentMethod\_MailedCheck,

Marital\_Divorced, Marital\_Married,

Marital\_NeverMarried, Marital\_Widowed,

Marital\_Separated, CaseOrder))

str(Mutate\_Churn2)

#Export clean data

write.csv(Mutate\_Churn2, "C:/Users/ntrei/OneDrive/Documents/MSDA/MutateChurn2.209.csv")

#Visualize correlation

library(corrr)

Mutate\_Churn2 %>%

dplyr::select(where(is.numeric)) %>%

correlate() %>%

shave() %>%

rplot(print\_cor = TRUE) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

str(Mutate\_Churn2)

describe(Mutate\_Churn2)

####DF

DF <- Mutate\_Churn2[c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None',

'Churn')]

#Partition data, set seed

set.seed(1234)

#Proportion for reproducibility- Train

Train\_prop <- 0.7

#Partition data into training and test sets

Train\_indices <- createDataPartition(DF$Churn, p = Train\_prop, list = FALSE)

Train\_data <- DF[Train\_indices, ]

Test\_data <- DF[-Train\_indices, ]

#CSV files

write.csv(Train\_data, file = "Training\_data.csv", row.names = FALSE)

write.csv(Test\_data, file = "Test\_data.csv", row.names = FALSE)

#Load test

Train\_data <- read.csv("Training\_data.csv")

Test\_data <- read.csv("Test\_data.csv")

#Extract variables

X\_Train <- Train\_data[, c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None')]

X\_Test <- Test\_data[, c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None')]

#Create train and test sets

Y\_train <- Train\_data$Churn

Y\_test <- Test\_data$Churn

#Export Training and Test Data

write.csv(X\_Train, "C:/Users/ntrei/OneDrive/Documents/MSDA/XTrain209.csv")

write.csv(X\_Test, "C:/Users/ntrei/OneDrive/Documents/MSDA/XTest209.csv")

write.csv(Y\_train, "C:/Users/ntrei/OneDrive/Documents/MSDA/YTrain209.csv")

write.csv(Y\_test, "C:/Users/ntrei/OneDrive/Documents/MSDA/YTest209.csv")

#Check dimensions of the split

prop.table(table(Train\_data$Churn)) \* 100

prop.table(table(Test\_data$Churn)) \* 100

#Feature scaling

Train\_data[-43] = scale(Train\_data[-43])

Test\_data[-43] = scale(Test\_data[-43])

####Fit Naive Bayes on Training data

#library(naivebayes)

#NBmodel <- naive\_bayes(Churn ~ ., data = Train\_data)

library(e1071)

Classifier = naiveBayes(x = Train\_data[-43], y = Train\_data$Churn)

Classifier

##Make a prediction

#Pred <- predict(NBmodel, )

Y\_predict = predict(Classifier, newdata = Test\_data[-43])

Y\_predict

#Evaluate accuracy of the model

table(Y\_predict, Test\_data$Churn)

#Confusion Matrix

CM = table(Test\_data[, 43], Y\_predict)

#Visualization

plot(Y\_predict)

plot(CM)

#Define object to plot

class(Y\_predict)

rocobj <- roc(Test\_data$Churn, as.ordered(Y\_predict))

rocobj

#Create ROC plot

plot(rocobj, main="ROC curve")

#Define object to plot and calculate AUC

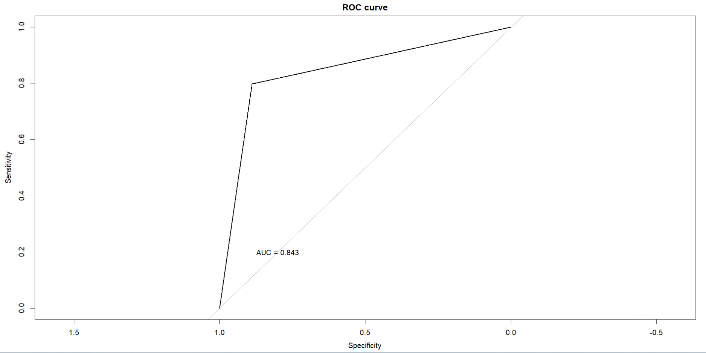
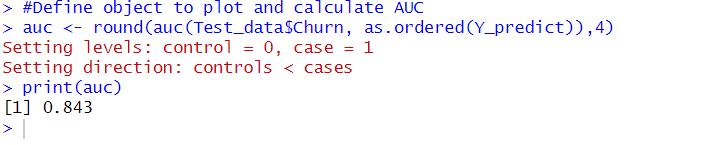
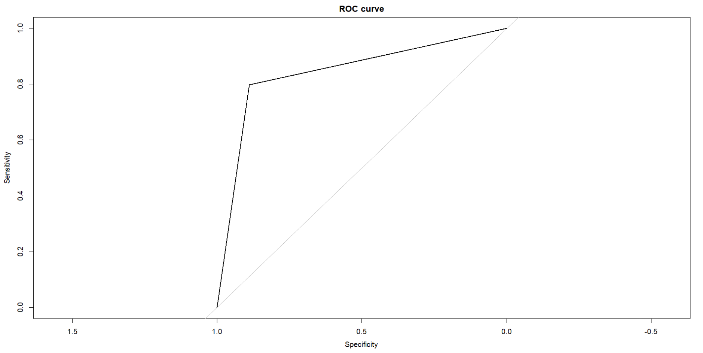
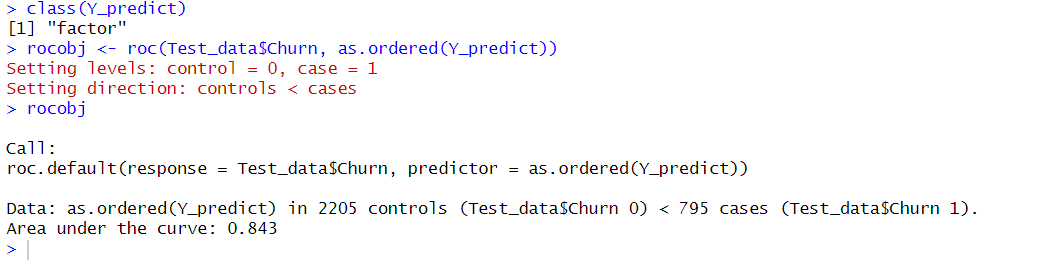
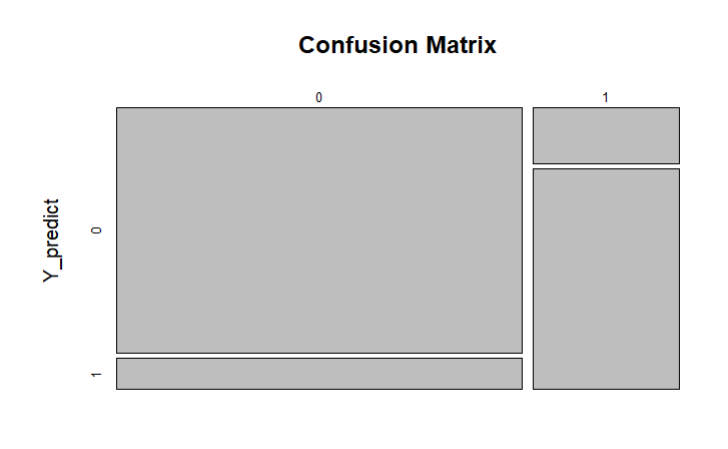
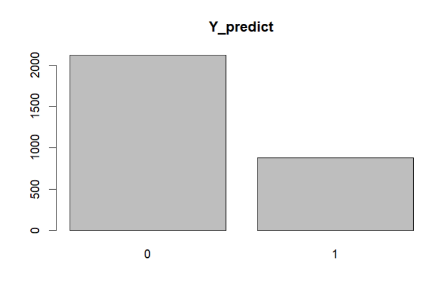
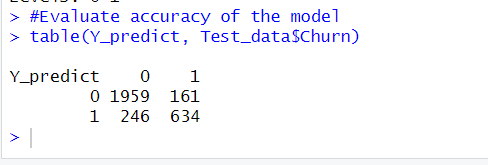
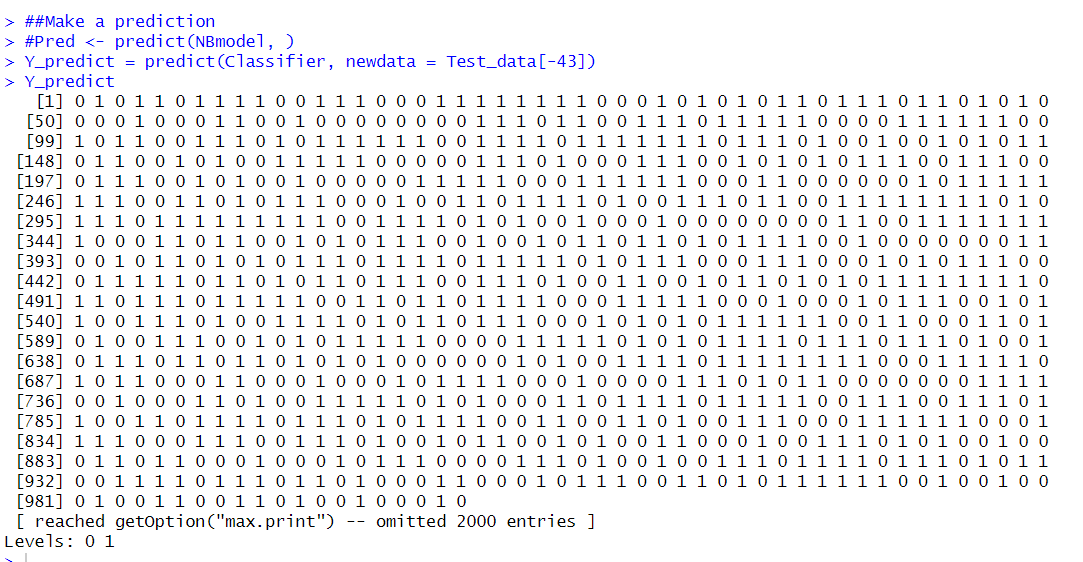
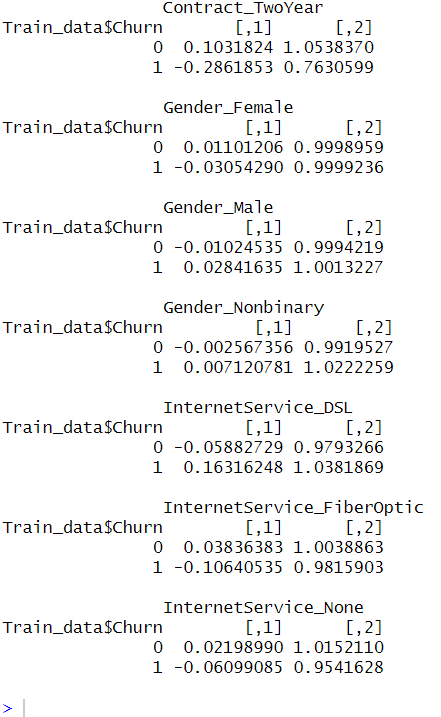
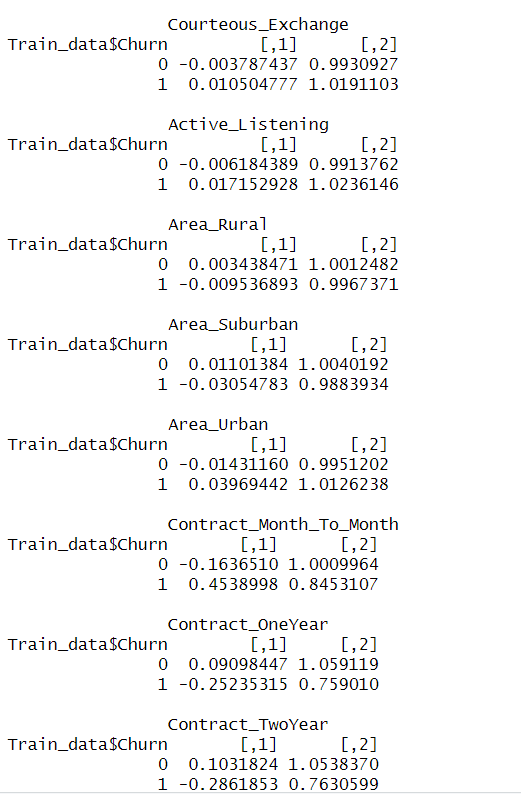
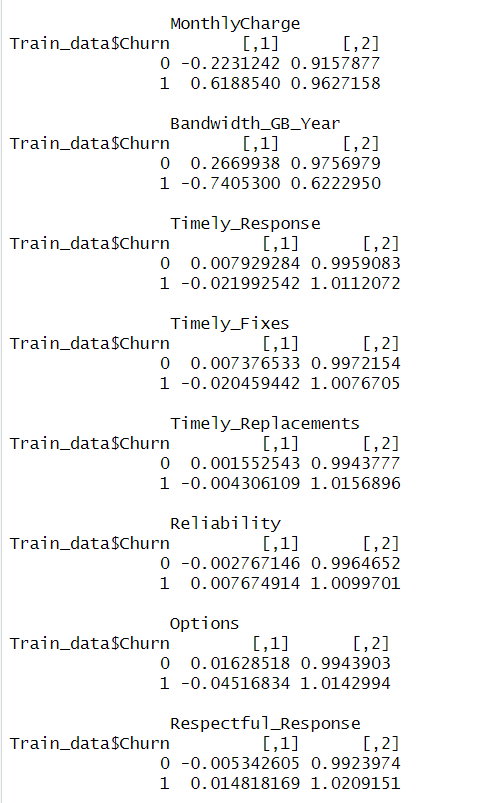
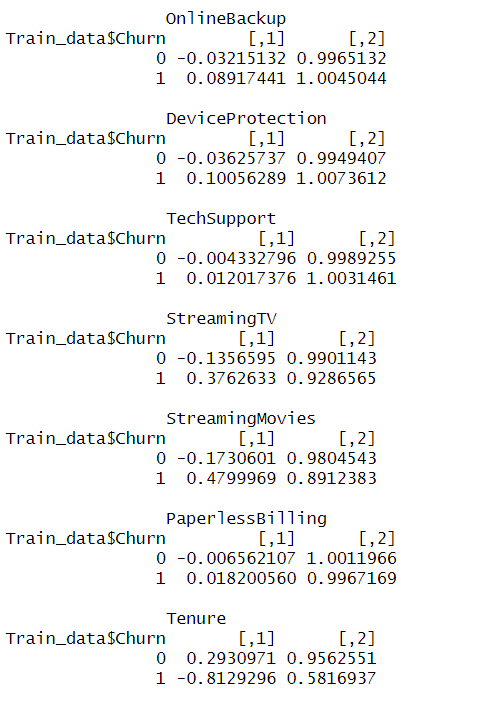
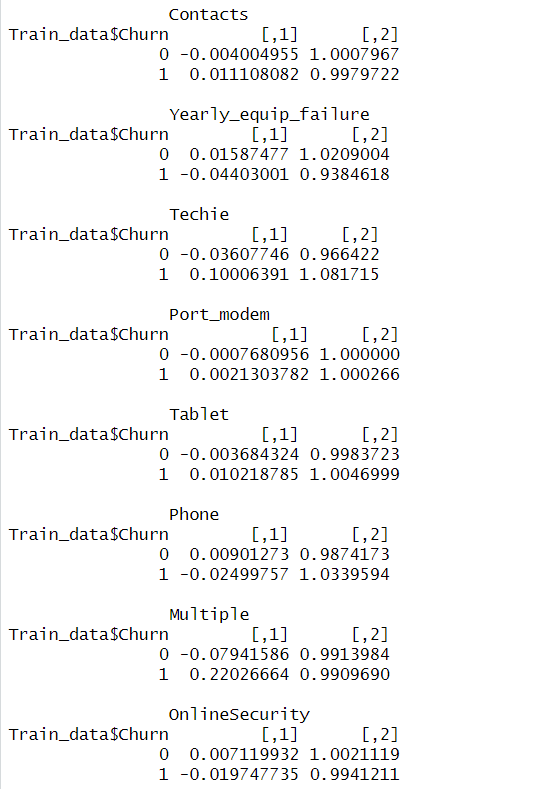
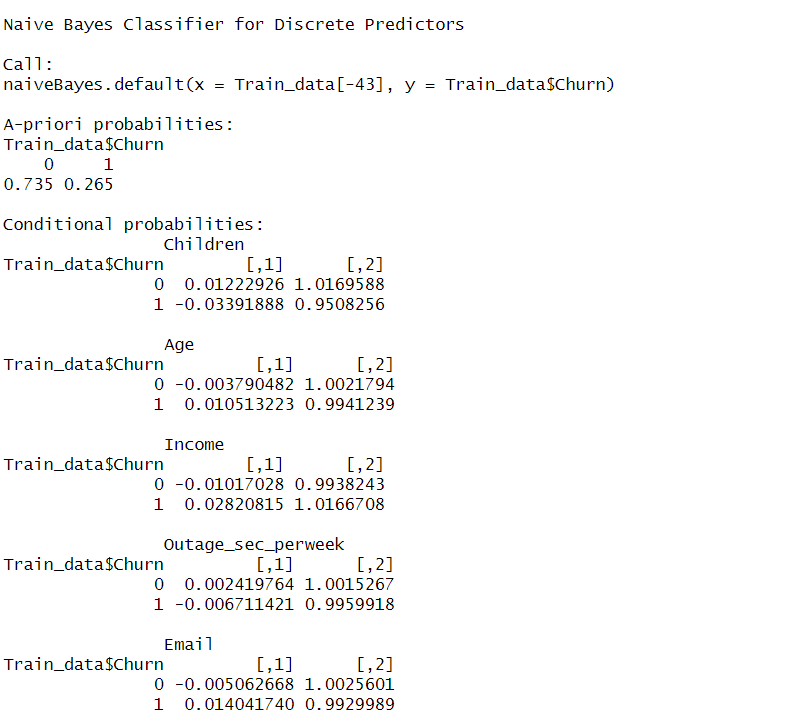
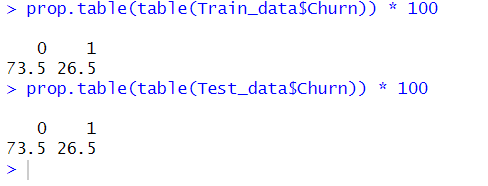
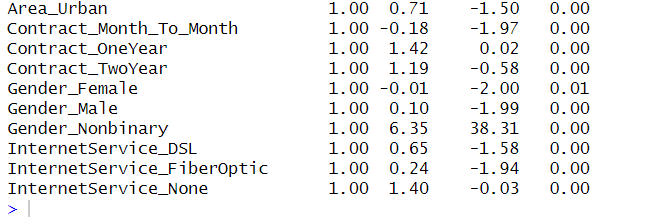
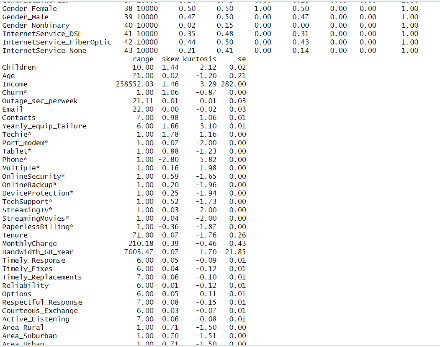
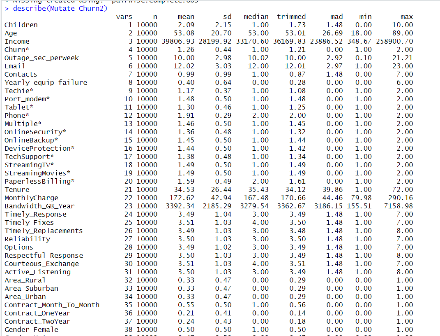
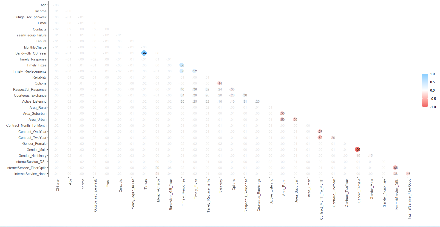
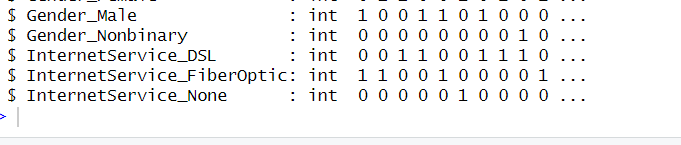
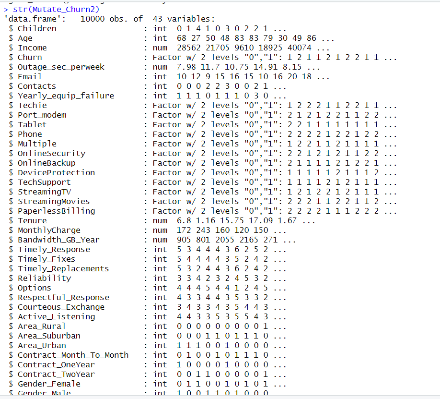
auc <- round(auc(Test\_data$Churn, as.ordered(Y\_predict)),4)

print(auc)

#Creat ROC plot with AUC

plot(rocobj, main="ROC curve")

text(0.8, 0.2, paste("AUC =", auc))



**4.** See the cleaned data set attached.

**D. Part IV: Analysis**

**1.** Splitting the data

####DF

DF <- Mutate\_Churn2[c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None',

'Churn')]

#Partition data, set seed

set.seed(1234)

#Proportion for reproducibility- Train

Train\_prop <- 0.7

#Partition data into training and test sets

Train\_indices <- createDataPartition(DF$Churn, p = Train\_prop, list = FALSE)

Train\_data <- DF[Train\_indices, ]

Test\_data <- DF[-Train\_indices, ]

#CSV files

write.csv(Train\_data, file = "Training\_data.csv", row.names = FALSE)

write.csv(Test\_data, file = "Test\_data.csv", row.names = FALSE)

#Load test

Train\_data <- read.csv("Training\_data.csv")

Test\_data <- read.csv("Test\_data.csv")

#Extract variables

X\_Train <- Train\_data[, c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None')]

X\_Test <- Test\_data[, c('Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Email',

'Contacts',

'Yearly\_equip\_failure',

'Techie',

'Port\_modem',

'Tablet',

'Phone',

'Multiple',

'OnlineSecurity',

'OnlineBackup',

'DeviceProtection',

'TechSupport',

'StreamingTV',

'StreamingMovies',

'PaperlessBilling',

'Tenure',

'MonthlyCharge',

'Bandwidth\_GB\_Year',

'Timely\_Response',

'Timely\_Fixes',

'Timely\_Replacements',

'Reliability',

'Options',

'Respectful\_Response',

'Courteous\_Exchange',

'Active\_Listening',

'Area\_Rural',

'Area\_Suburban',

'Area\_Urban',

'Contract\_Month\_To\_Month',

'Contract\_OneYear',

'Contract\_TwoYear',

'Gender\_Female',

'Gender\_Male',

'Gender\_Nonbinary',

'InternetService\_DSL',

'InternetService\_FiberOptic',

'InternetService\_None')]

#Create train and test sets

Y\_train <- Train\_data$Churn

Y\_test <- Test\_data$Churn

#Export Training and Test Data

write.csv(X\_Train, "C:/Users/ntrei/OneDrive/Documents/MSDA/XTrain209.csv")

write.csv(X\_Test, "C:/Users/ntrei/OneDrive/Documents/MSDA/XTest209.csv")

write.csv(Y\_train, "C:/Users/ntrei/OneDrive/Documents/MSDA/YTrain209.csv")

write.csv(Y\_test, "C:/Users/ntrei/OneDrive/Documents/MSDA/YTest209.csv")

#Check dimensions of the split

prop.table(table(Train\_data$Churn)) \* 100

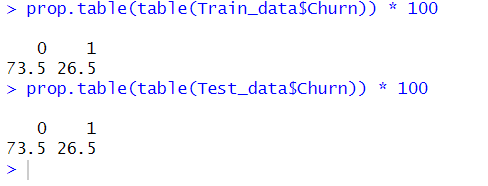
prop.table(table(Test\_data$Churn)) \* 100

#Feature scaling

Train\_data[-43] = scale(Train\_data[-43])

Test\_data[-43] = scale(Test\_data[-43])

See the attached split training and test data sets.



**2.** Output and intermediate calculations

####Fit Naive Bayes on Training data [In text citation: Rai, B. (2019)]

#library(naivebayes)

#NBmodel <- naive\_bayes(Churn ~ ., data = Train\_data)

library(e1071)

Classifier = naiveBayes(x = Train\_data[-43], y = Train\_data$Churn)

Classifier

##Make a prediction

#Pred <- predict(NBmodel, )

Y\_predict = predict(Classifier, newdata = Test\_data[-43])

Y\_predict

#Evaluate accuracy of the model

table(Y\_predict, Test\_data$Churn)

#Confusion Matrix

CM = table(Test\_data[, 43], Y\_predict)

#Visualization

plot(Y\_predict)

plot(CM)

#Define object to plot

class(Y\_predict)

rocobj <- roc(Test\_data$Churn, as.ordered(Y\_predict))

rocobj

#Create ROC plot

plot(rocobj, main="ROC curve")

#Define object to plot and calculate AUC

auc <- round(auc(Test\_data$Churn, as.ordered(Y\_predict)),4)

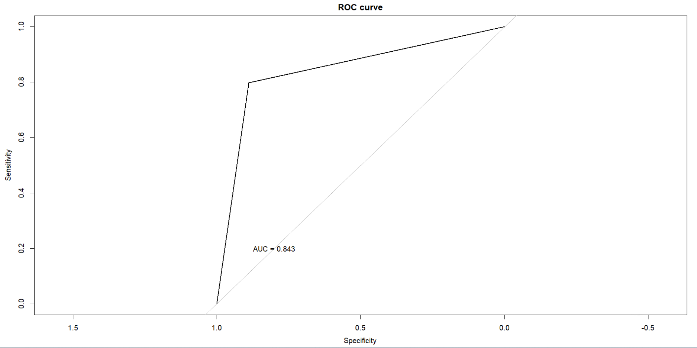
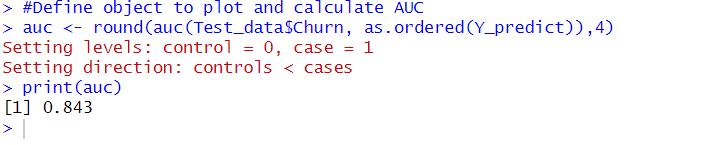
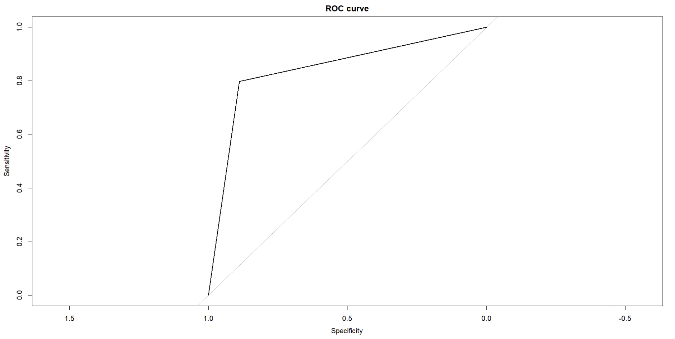
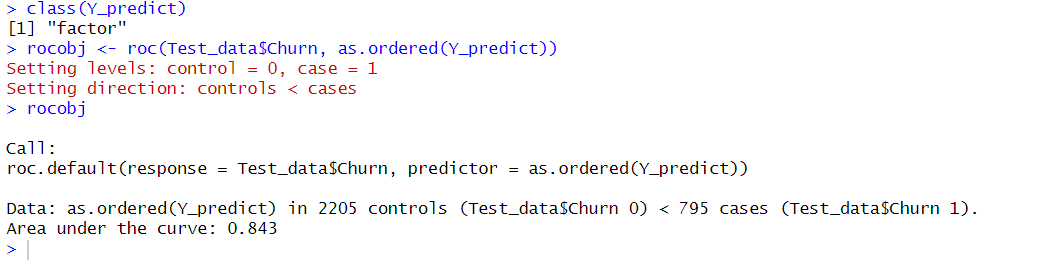
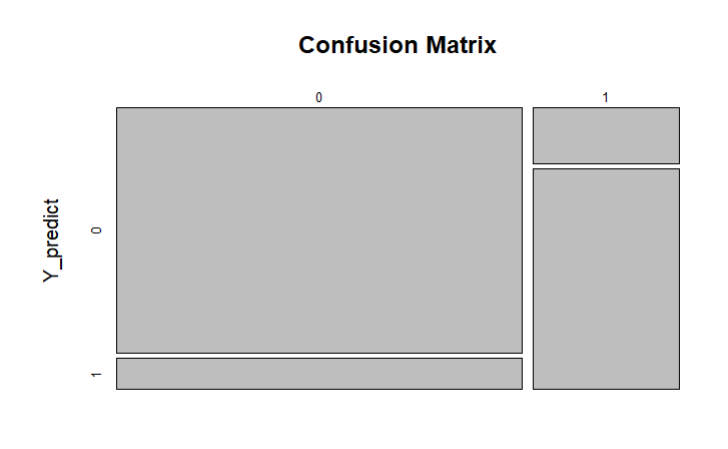
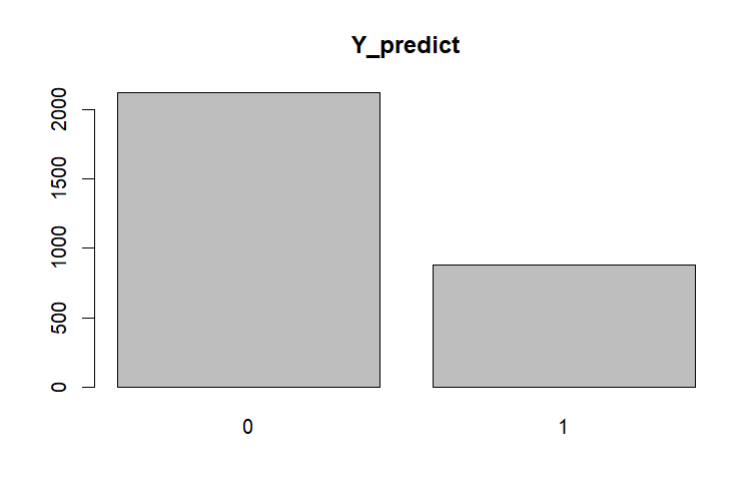
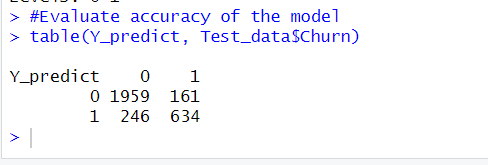
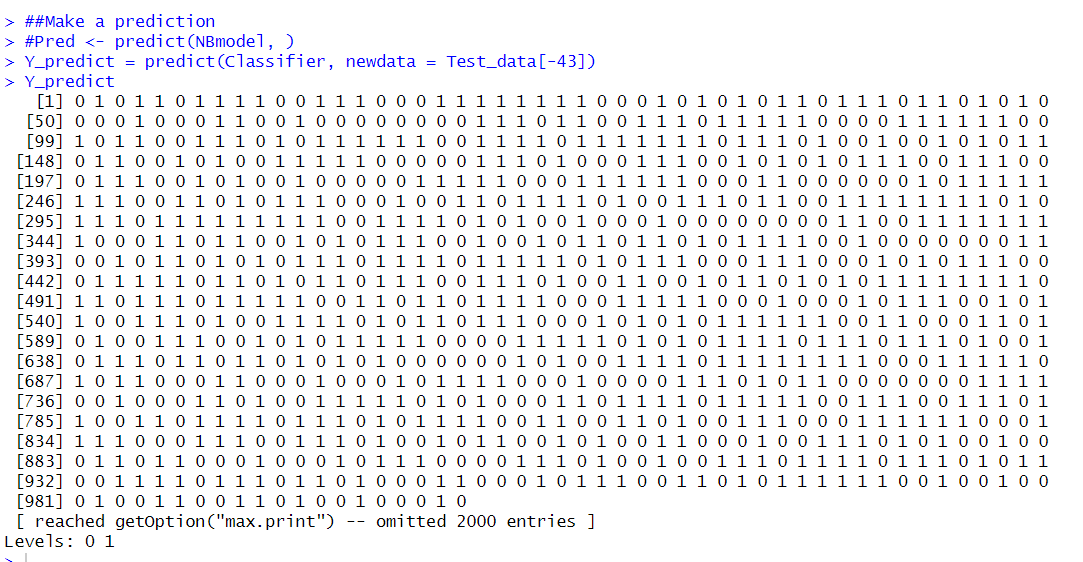
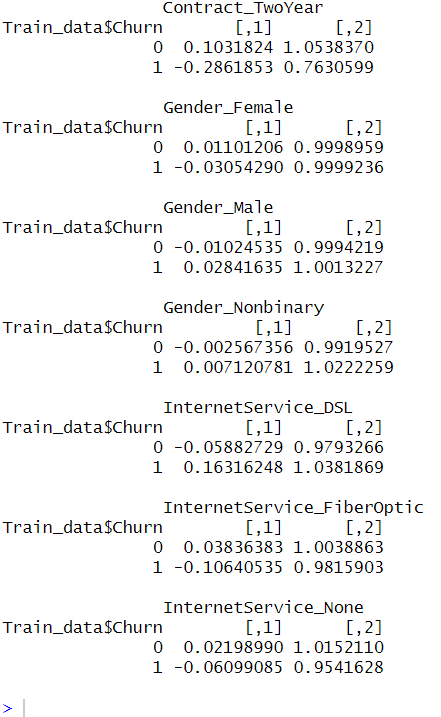
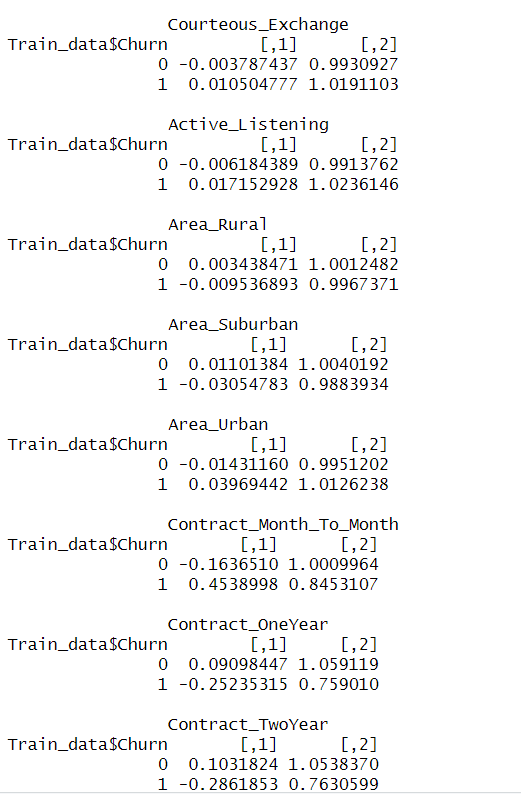
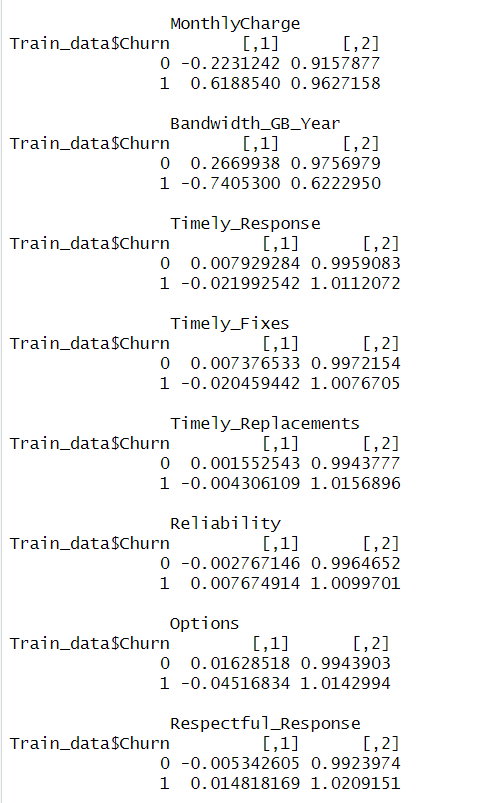
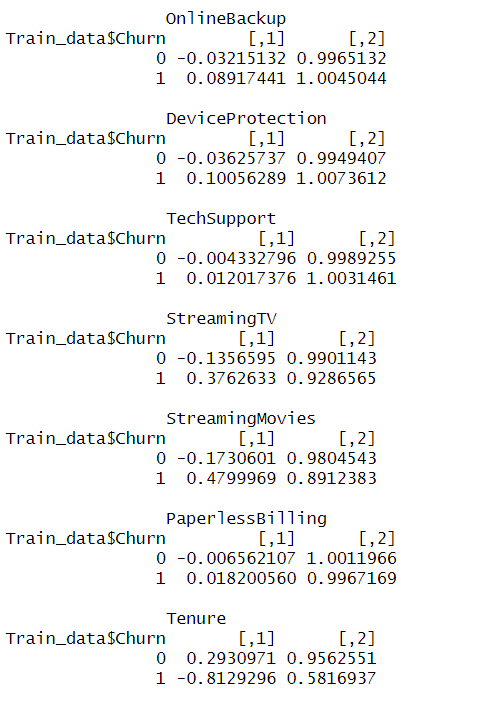
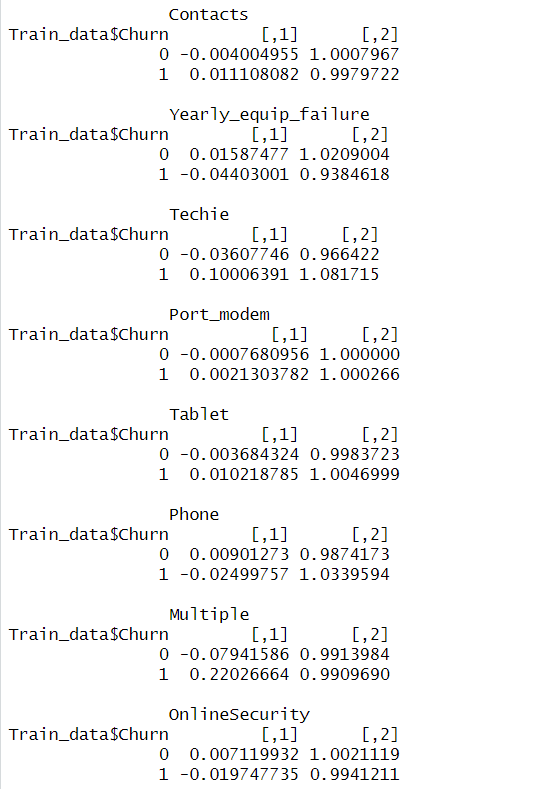
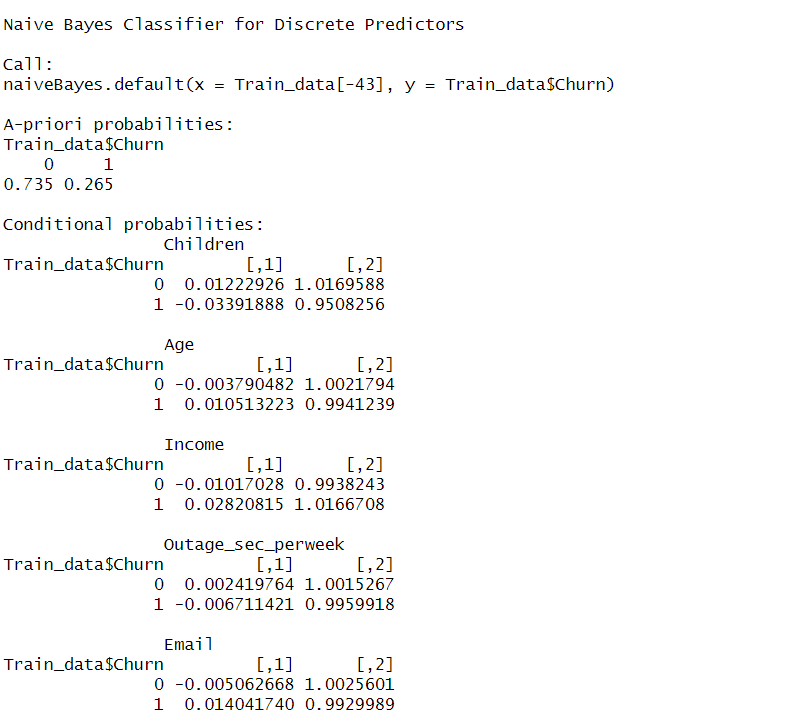
print(auc)

#Creat ROC plot with AUC

plot(rocobj, main="ROC curve")

text(0.8, 0.2, paste("AUC =", auc))

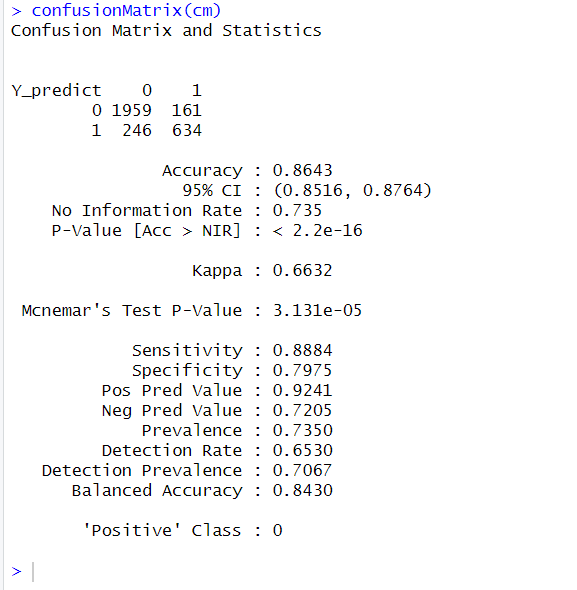
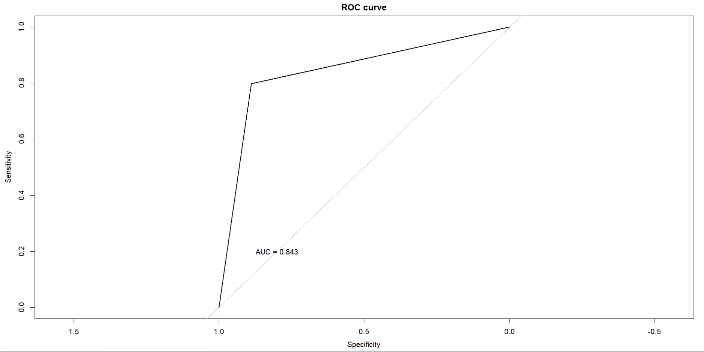
No intermediate calculations were performed in this analysis.



**3.** See the attached code.

**E. Part V: Data Summary and Implications**

**1.** The accuracy score is 0.8643 which means it is 86% percent accurate and the AUC area under the curve is 0.843 which is a good score. As you can see on the ROC receiver operator characteristics curve below, the line is close to 1. 1 is where a model has predictions 100% correct. Therefore 0.843 or 84% is a good score.



**2.** The result and implication of the analysis is that Churn the dependent variable has a good probability of being predicted by the independent variables in the analysis.

**3.** A limitation of this analysis is that Naive Bayes assumes all independent variables are independent, but this may not always be the case. For example, two independent variables such as income and children or other variables may have a relationship.

**4.** The recommended course of action is to further analyze the variables that customers who churn have in common to further narrow down the root cause of churn. This would result in a greater likelihood of success in retaining customers who typically would churn.

**F.** **Part VI: Demonstration**

**1.** See attached Panopto recording.

**G.** Sources

Elleh, F. (2023). *D209 Task 1: Expectations and Data Preprocessing- Python.*

[D209 Task 1: Expectations and Data Preprocessing - Python (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7329050b-3de4-412a-9ff4-b0790009d02b)

Rai, B. (2019). *Naive Bayes Classification with R.*

[Naive Bayes Classification with R | Example with Steps (youtube.com)](https://www.youtube.com/watch?v=RLjSQdcg8AM)