Data Mining D209

NVM2 Task 1 Classification Analysis

John-Eric Bonilla

Western Governors University

WGU Student ID #001464199

D209 Task 1

Part I: Research Question

A.1

The question we will answer with this data mining and analysis project is:

Utilizing the Naive Bayes algorithm with a few select variables can we predict churn?

A.2.

The goal of predicting churn is to reduce our churn ratio, extend client tenure and increase our business intelligence to guide our business model going forward.

Part II: Method Justification

B.1.

The chosen classification method for this analysis is Naive Bayes. This algorithm according to research from (Stecanella, 2017) is one of the most uncomplicated options and ofttimes a more efficient one. Despite the extraordinary progress in machine learning in recent years, Naive Bayes has demonstrated not only to be simple but also quick, precise, and dependable. Naive Bayes is a type of probabilistic algorithms and takes advantage of the probability theory and Bayes' Theorem to calculate a highly probable result. Probabilistic means that Naive Bayes evaluates the probability bases on a given variable, and then output the highest probable one. The way it reaches these probabilities is by utilizing Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

B.2.

Naive Bayes Assumptions

The original Naive Bayes supposition according to research from (Flatiron School Team, 2006) is that each attribute creates an independent and equal (nearly identical) stimulus on the conclusion. This is accepted as the i.i.d assumption. The expected outcome is to clearly identify the percentage of predicting accuracy.

B.3.

This project will be performed in the R programming language because according to research from Stat Analytica (admin, n.d.) R is not as conventional as other programming languages. R is specifically created for statistical and data reconfiguration. The library of R is expressly devised to make data analysis easier, more detailed, and accessible. using a variety of libraries in the mathematical, scientific, and research-based arenas. For the statistical functions of the project such as model building, sample splitting and analysis we will utilize the R package of caret which is short for "classification and regression training", caret contains functions to streamline the model training process for complex regression and classification problems. Also utilized for our statistical ventures is the naivebayes package which provides an efficient implementation of the popular Naive Bayes classifier algorithm. This package is exceptionally suitable for this analysis because we will be implementing this project under the Naïve Bayes theorem. Aimed at the required data wrangling, manipulation, cleaning, and feature engineering will be employ using the dplyr and plyr packages. And for the beautiful and explanatory graphics the packages of ggplot2 and knitr. Our final package will be InformationValue which is specifically designed to help generate plots such as the 'ROC' Curve in 'ggplot2', 'AUROC', 'IV', 'WOE' Calculation, 'KS Statistic', and to support accuracy development.

Part III: Data Preparation

C.1.

Data Processing goal

According to research by (Obaidat et al.), the single most critical preprocessing goal is to identify and then modify all null or NA values.

C.2.

Variable Identification

The initial attributes used to perform this analysis are first the outcome variable Churn which is a categorical datatype. The other categorical datatypes utilized for this analysis are Contract and Internet Service. Concerning our continuous attributes they will consist of Income, Yearly Equipment Failure, Monthly Charge, and Bandwidth Gigabytes.

C.3.

Data Preparation Steps

I. Import data

   i. Code (originaldata = read.csv("C:/Users/Admin/Desktop/churn_clean.csv"))

II. Reduce data into a subset data-frame

i. Code (dataset <- originaldata[, c("Churn","Contract", “Internet Service”, “Income”, “Yearly_equip_fail”, “MonthlyCharge", “Bandwidth_GB_Year")])

III. Encoding the target variable as factor

i. Code (dataset$Churn = factor(dataset$Churn))

IV. Revaluing target variable from Yes and No to 1 and 1

i. Code (dataset$Churn <- revalue(dataset$Churn, c("Yes"=0)))

ii. Code (dataset$Churn <- revalue(dataset$Churn, c("No"=1)))

C.4.

Clean Dataset Attached

I. Code used

a. Code (write.csv(dataset, "Task1CleanData.csv"))

Part IV: Analysis

Section D

D.1. Analysis

Training and testing datasets

- Code
  - #Data partition
  - set.seed(1234)
  - ind <- sample(2, nrow(dataset), replace = T, prob = c(0.8, 0.2))
  - train <- dataset[ind== 1,]
  - test <- dataset[ind ==2,]

I. Files attached

a. Code (write.csv(train, "Task1training_set.csv"))

b. Code (write.csv(test, "Task1test_set.csv"))

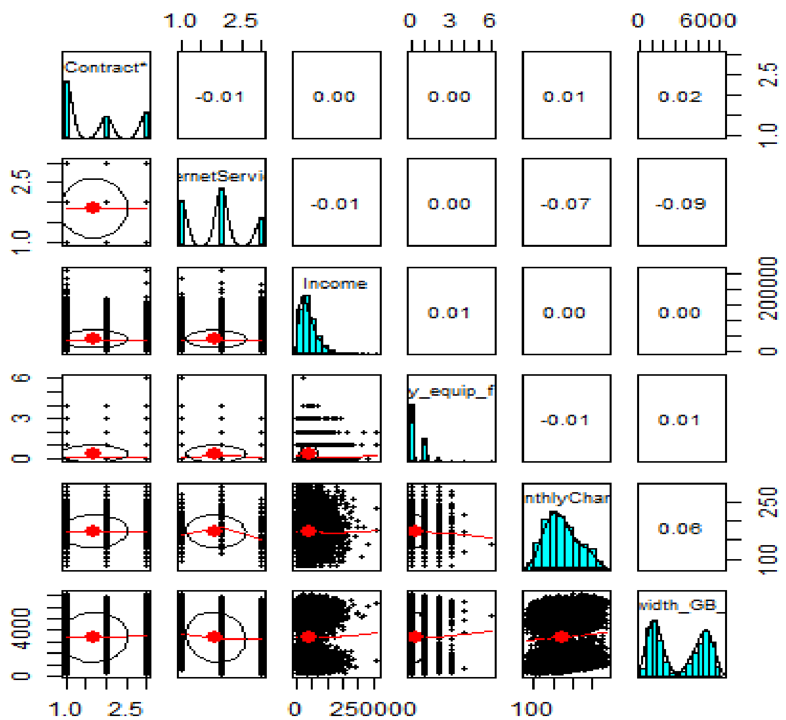
D.2.

Descriptions of analysis techniques utilized with screenshots

- According to research from (Papadopoulos et al.), the first step in the Naïve Bayes analytic process is to verify the chosen independent variables are not highly correlated.
  - Code

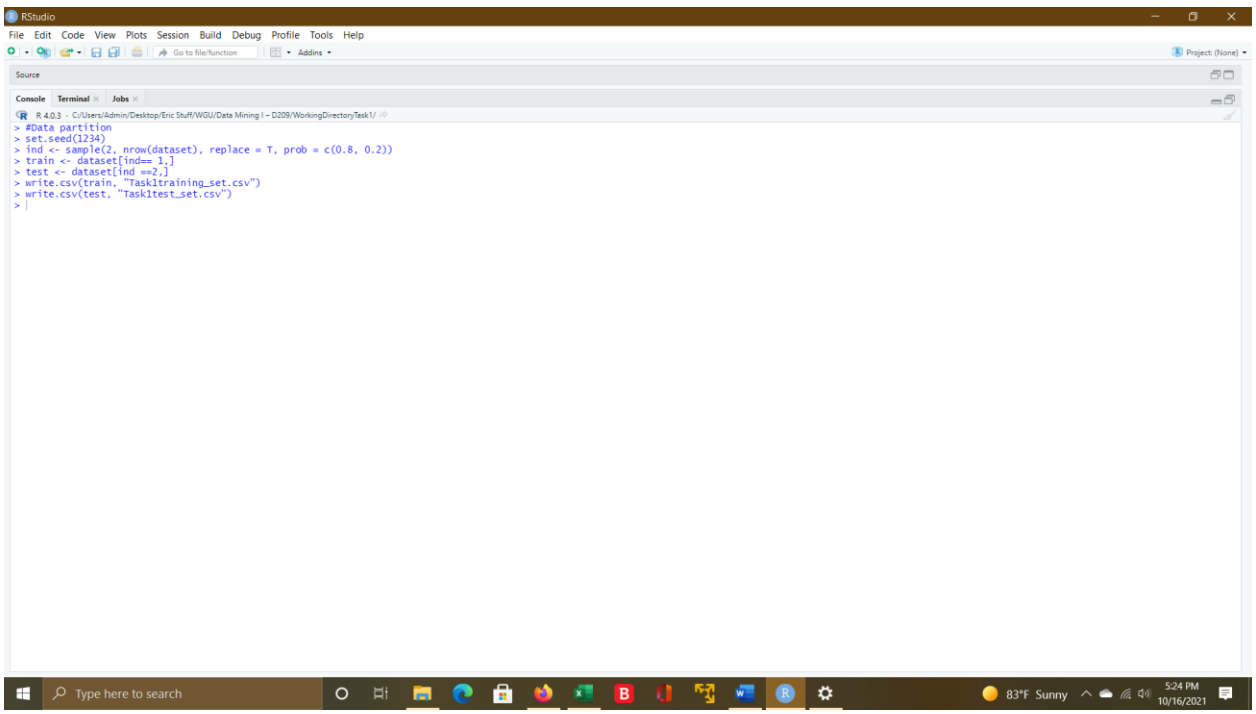
➢ #Checking correlation

➢ pairs.panels(dataset[-1])

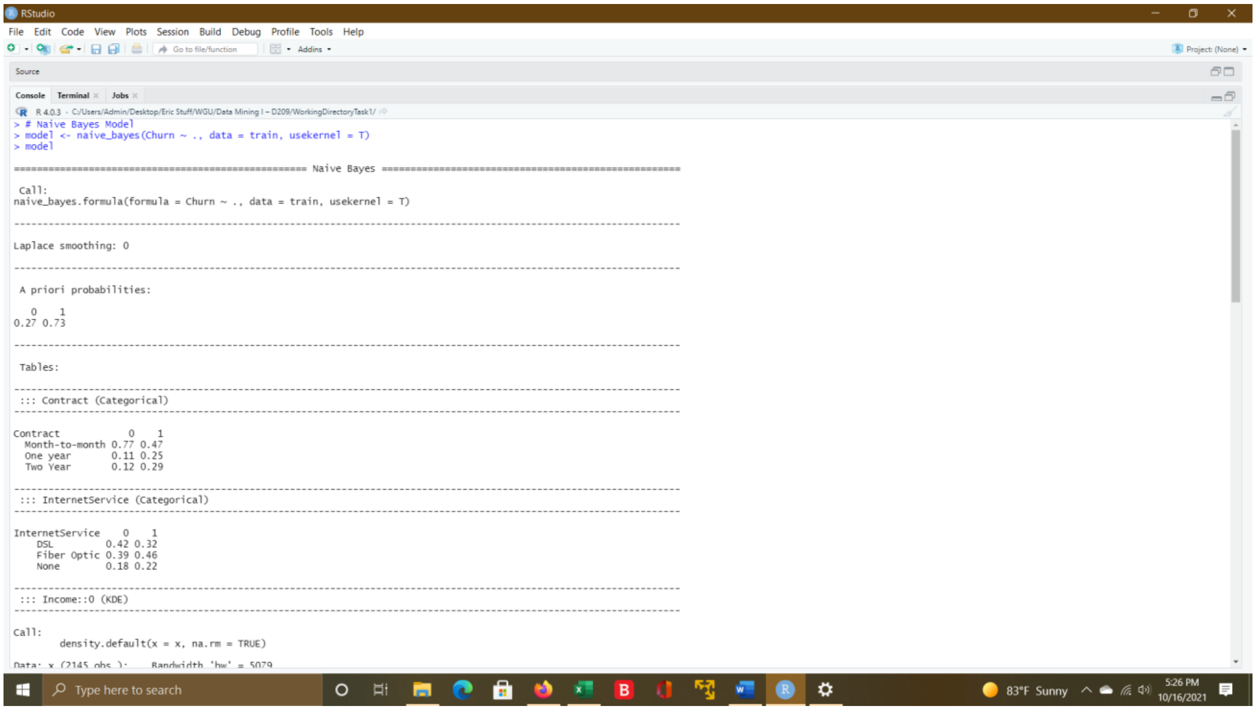


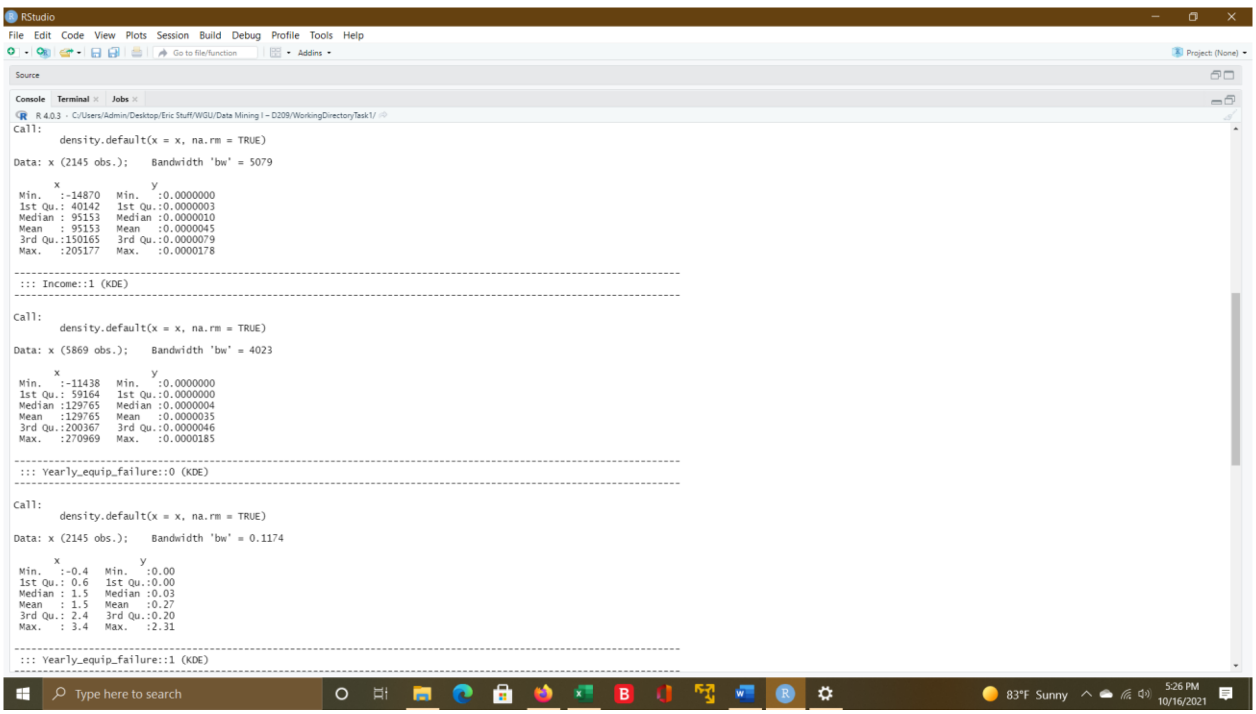
#Data partition

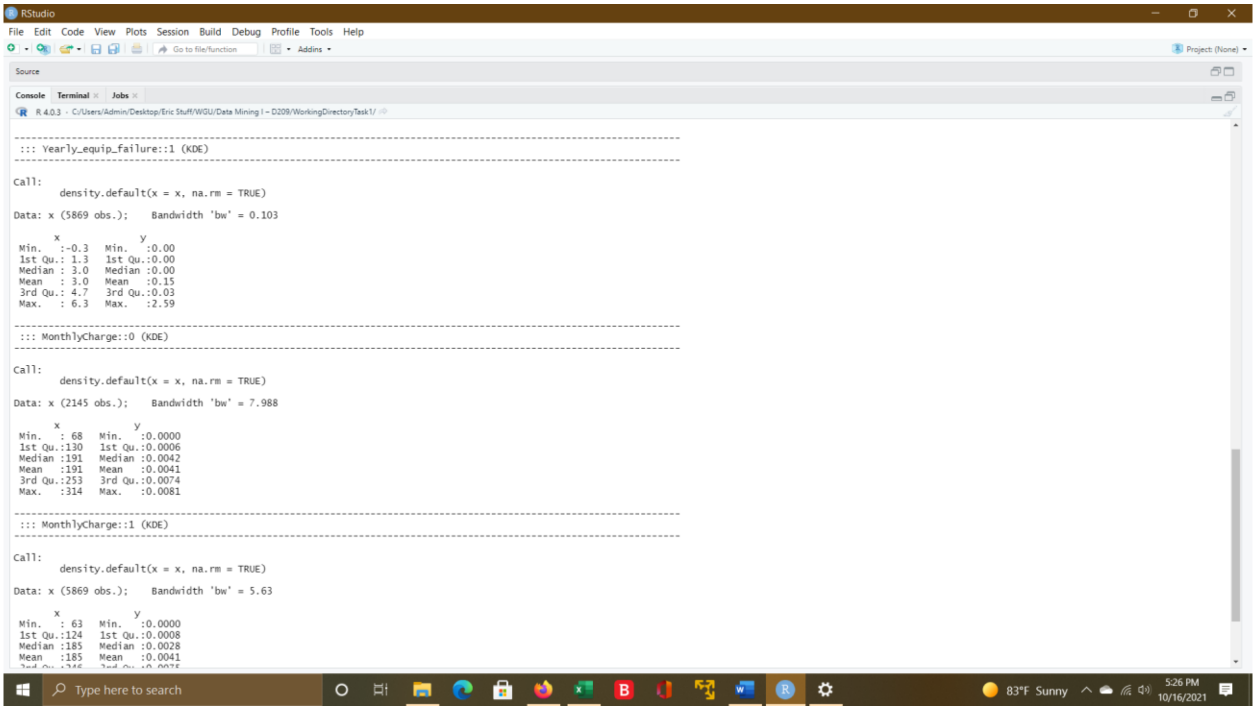
- ➢ Code
- ➢ set.seed(1234)
- ➢ ind <- sample(2, nrow(dataset), replace = T, prob = c(0.8, 0.2))
- ➢ train <- dataset[ind== 1,]
- ➢ test <- dataset[ind ==2,]
- ➢ write.csv(train, "Task1training_set.csv")
- ➢ write.csv(test, "Task1test_set.csv")



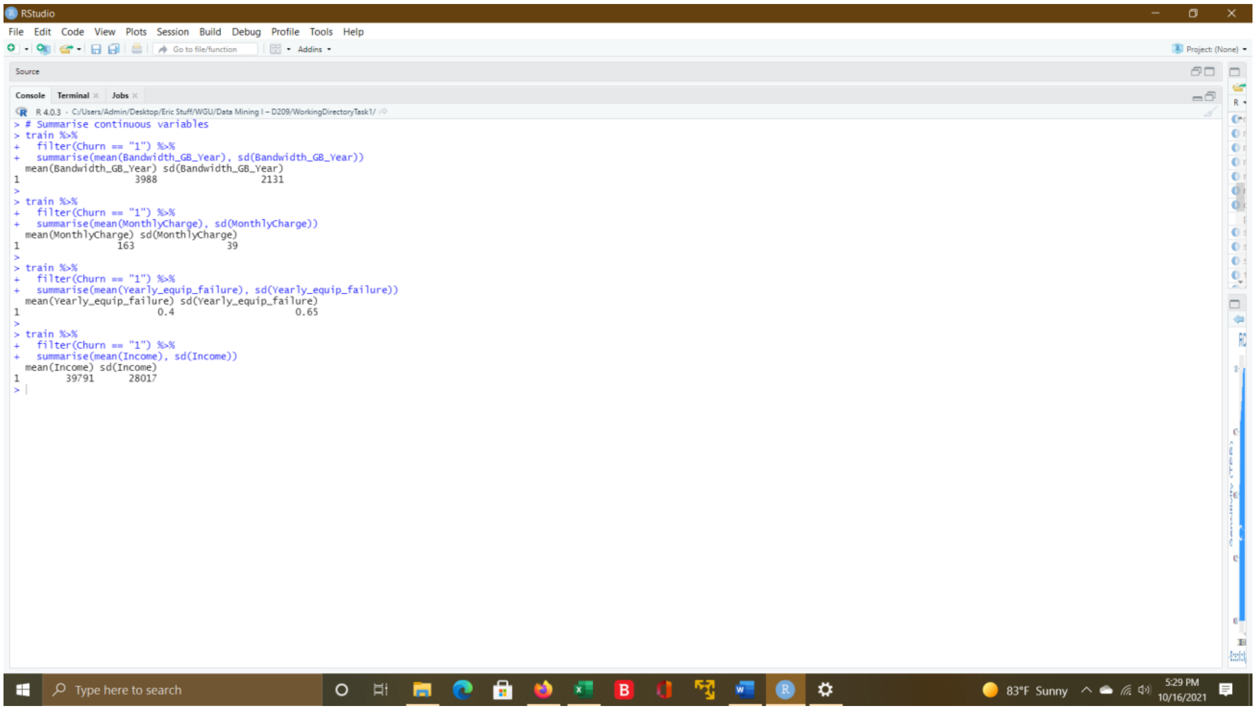
- ➢ # Naive Bayes Model
  - ○ model <- naive_bayes(Churn ~ ., data = train, usekernel = T)
  - ○ model


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Jobs
R 4.0.3 · C:/Users/Admin/Desktop/Eric Stuff/WGU/Data Mining I – D209/WorkingDirectoryTask1/
> # Naive Bayes Model
> model <- naive_bayes(Churn ~ ., data = train, usekernel = T)
> model
================================ Naive Bayes ================================
Call:
naive_bayes.formula(formula = Churn ~ ., data = train, usekernel = T)
Laplace smoothing: 0
A priori probabilities:
0 1
0.27 0.73
Tables:
::: Contract (Categorical)
Contract 0 1
Month-to-month 0.77 0.47
One year 0.11 0.25
Two Year 0.12 0.29
::: InternetService (Categorical)
InternetService 0 1
DSL 0.42 0.32
Fiber Optic 0.39 0.46
None 0.18 0.22
::: Income::0 (KDE)
Call:
density.default(x = x, na.rm = TRUE)
Type here to search
83°F Sunny
5:26 PM
10/16/2021

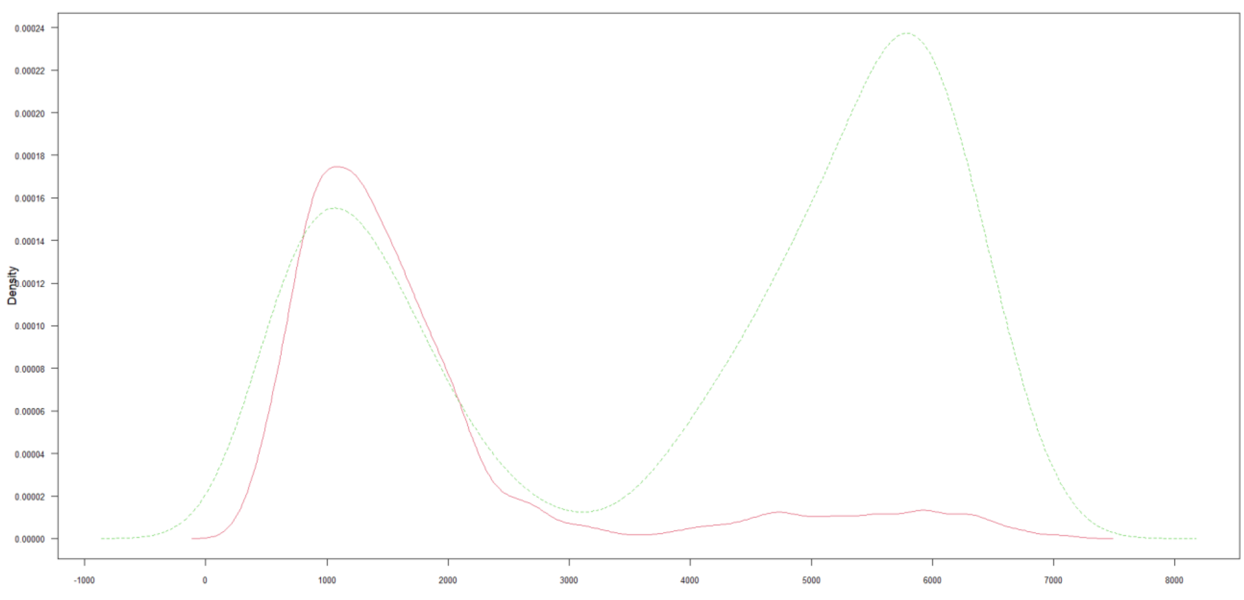

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Jobs
R 4.0.3 · C:/Users/Admin/Desktop/Eric Stuff/WGU/Data Mining I – D209/WorkingDirectoryTask1/
Call:
density.default(x = x, na.rm = TRUE)
Data: x (2145 obs.); Bandwidth 'bw' = 5079
x y
Min. :-14870 Min. :0.0000000
1st Qu.: 40142 1st Qu.:0.0000003
Median : 95153 Median :0.0000010
Mean : 95153 Mean :0.0000045
3rd Qu.:150165 3rd Qu.:0.0000079
Max. :205177 Max. :0.0000178
::: Income::1 (KDE)
Call:
density.default(x = x, na.rm = TRUE)
Data: x (5869 obs.); Bandwidth 'bw' = 4023
x y
Min. :-11438 Min. :0.0000000
1st Qu.: 59164 1st Qu.:0.0000000
Median :129765 Median :0.0000004
Mean :129765 Mean :0.0000035
3rd Qu.:200367 3rd Qu.:0.0000046
Max. :270969 Max. :0.0000185
::: Yearly_equip_failure::0 (KDE)
Call:
density.default(x = x, na.rm = TRUE)
Data: x (2145 obs.); Bandwidth 'bw' = 0.1174
x y
Min. :-0.4 Min. :0.00
1st Qu.: 0.6 1st Qu.:0.00
Median : 1.5 Median :0.03
Mean : 1.5 Mean :0.27
3rd Qu.: 2.4 3rd Qu.:0.20
Max. : 3.4 Max. :2.31
::: Yearly_equip_failure::1 (KDE)
Type here to search
83°F Sunny
5:26 PM
10/16/2021



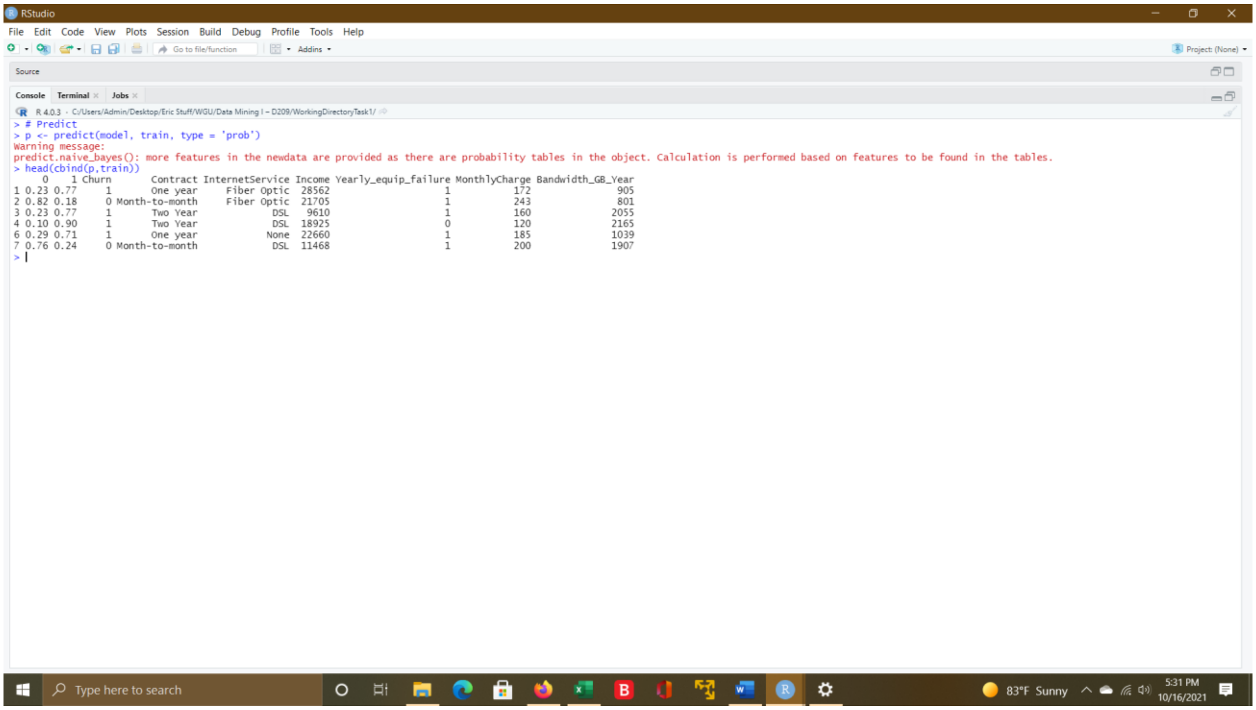
- # Summarise continuous variables
- train %>%
- filter(Churn == "1") %>%
- summarise(mean(Bandwidth_GB_Year), sd(Bandwidth_GB_Year))
- train %>%
- filter(Churn == "1") %>%
- summarise(mean(MonthlyCharge), sd(MonthlyCharge))
- train %>%
- filter(Churn == "1") %>%
- summarise(mean(Yearly_equip_failure), sd(Yearly_equip_failure))
- train %>%
- filter(Churn == "1") %>%
- summarise(mean(Income), sd(Income))



plot(model)



Plot model



Predict

p <- predict(model, train, type = 'prob')

head(cbind(p,train))

#placing prediction on p1 object and # Creating confusion matrix - train data

p1 <- predict(model, train)

(tab1 <- table(p1, train$Churn))

Summing correct predictions and preducing error rate

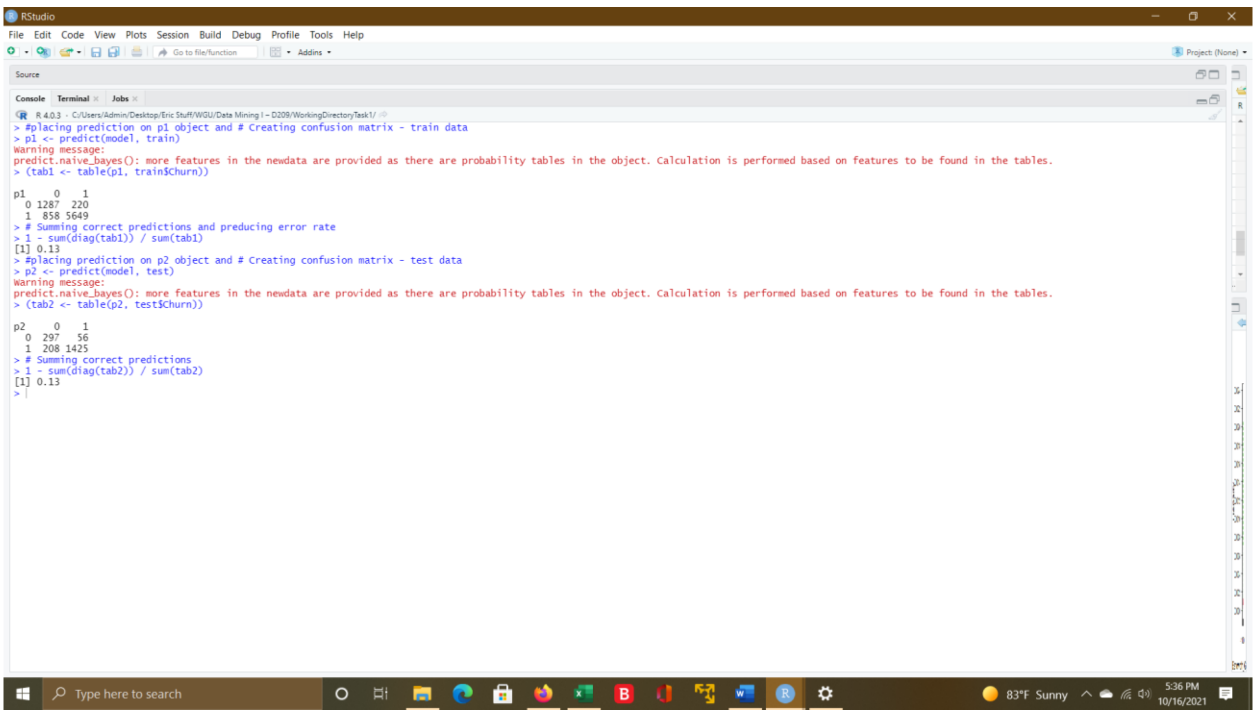
1 - sum(diag(tab1)) / sum(tab1)

#placing prediction on p2 object and # Creating confusion matrix - test data p2 <- predict(model, test)

(tab2 <- table(p2, test$Churn))

Summing correct predictions

1 - sum(diag(tab2)) / sum(tab2)



Comparing accurate scores against predictive scores and running the kolmogorov-smirnov statistic

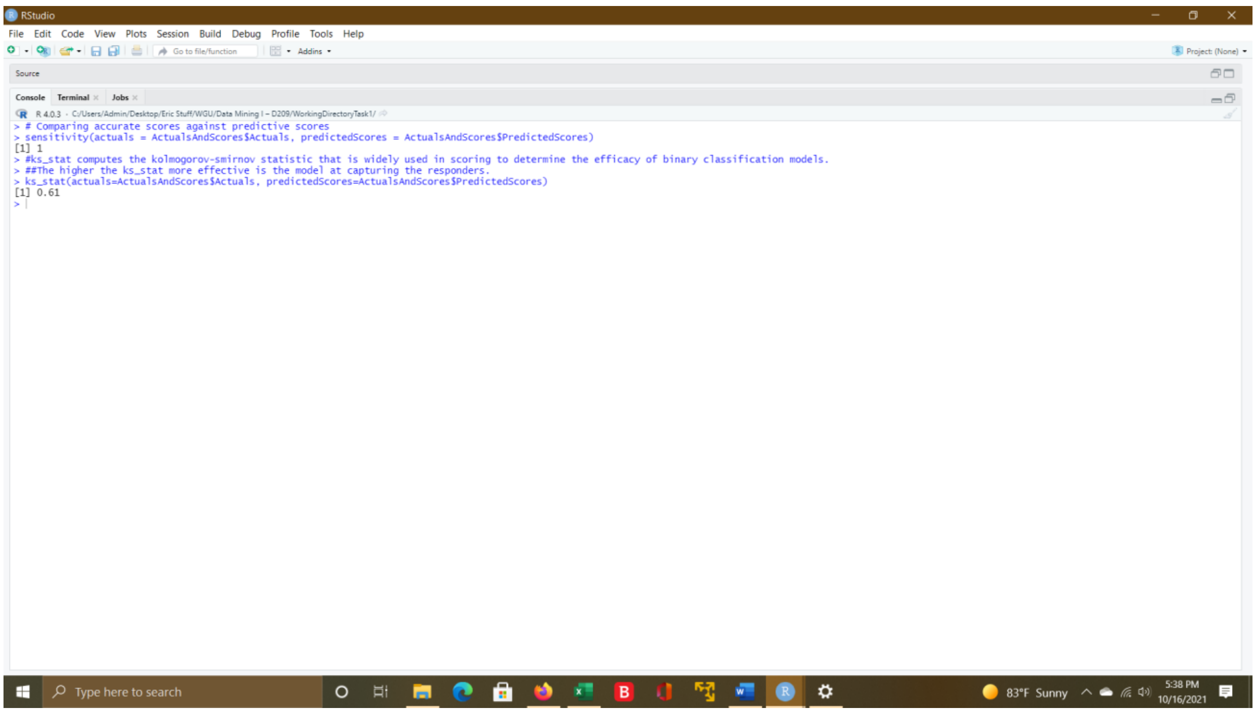
sensitivity(actuals = ActualsAndScores$Actuals, predictedScores = ActualsAndScores$PredictedScores)

#ks_stat computes the kolmogorov-smirnov statistic that is widely used in scoring to determine the efficacy of binary classification models.

##The higher the ks_stat more effective is the model at capturing the responders.

ks_stat(actuals=ActualsAndScores$Actuals,

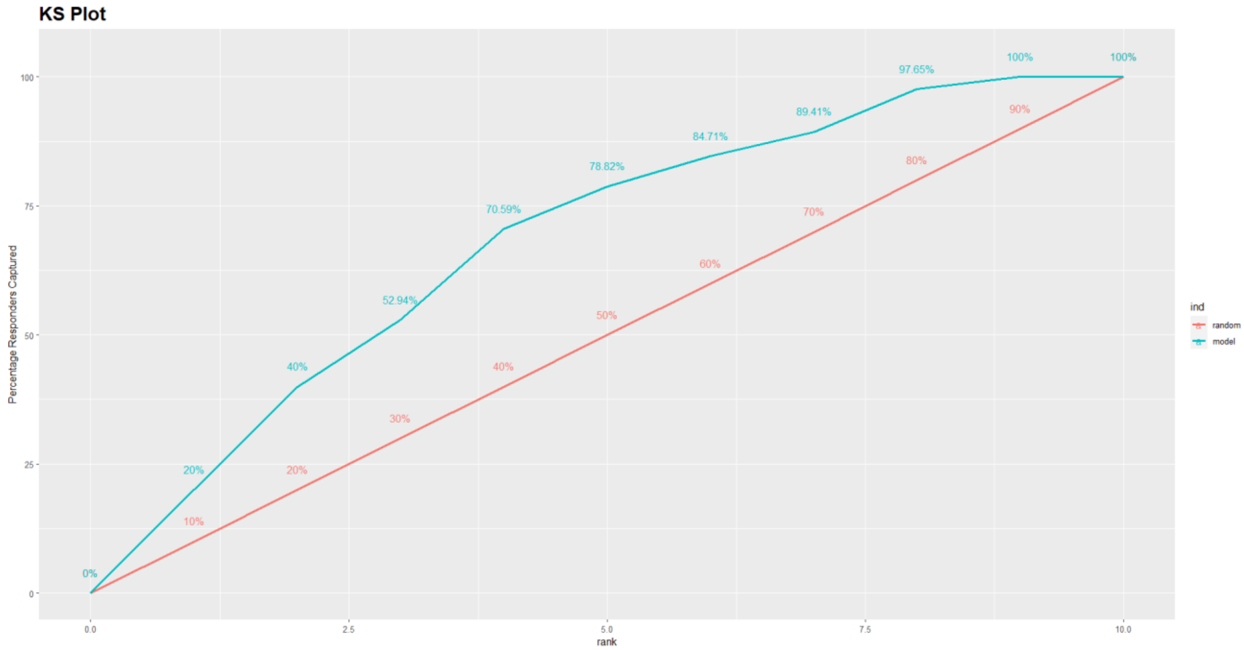
predictedScores=ActualsAndScores$PredictedScores)



Plotting the kolmogorov-smirnov statistic

ks_plot(actuals=ActualsAndScores$Actuals,

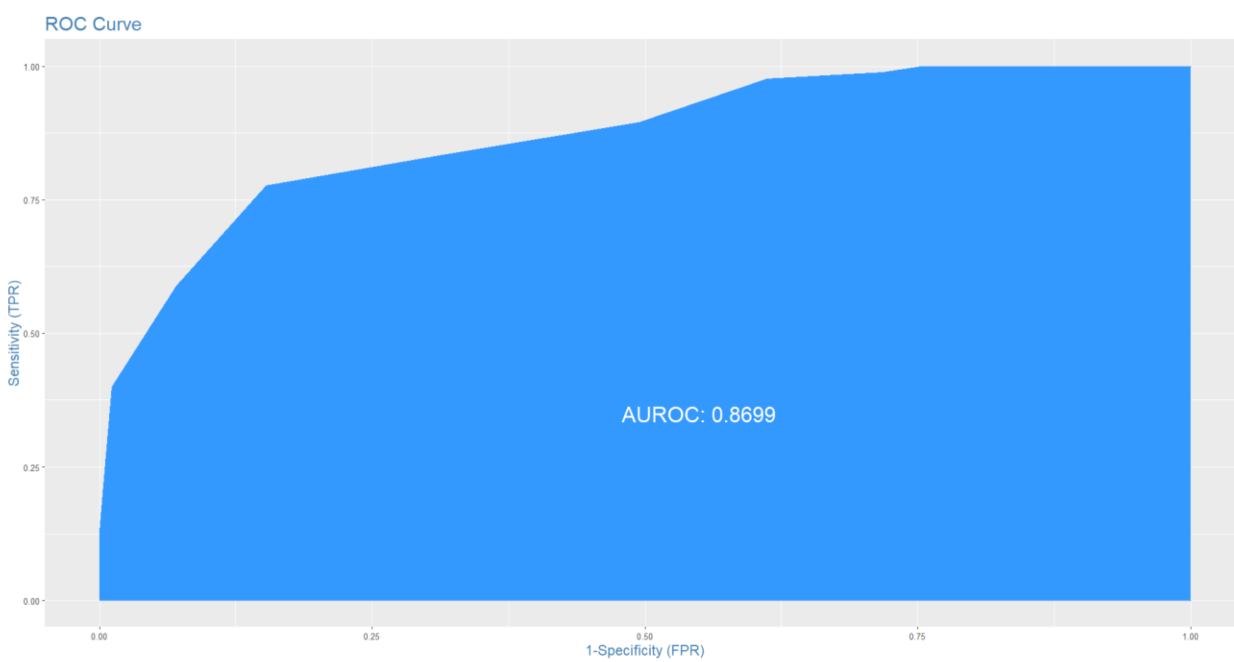
predictedScores=ActualsAndScores$PredictedScores)



Plotting the ROC chart

plotROC(actuals=ActualsAndScores$Actuals,

predictedScores=ActualsAndScores$PredictedScores)

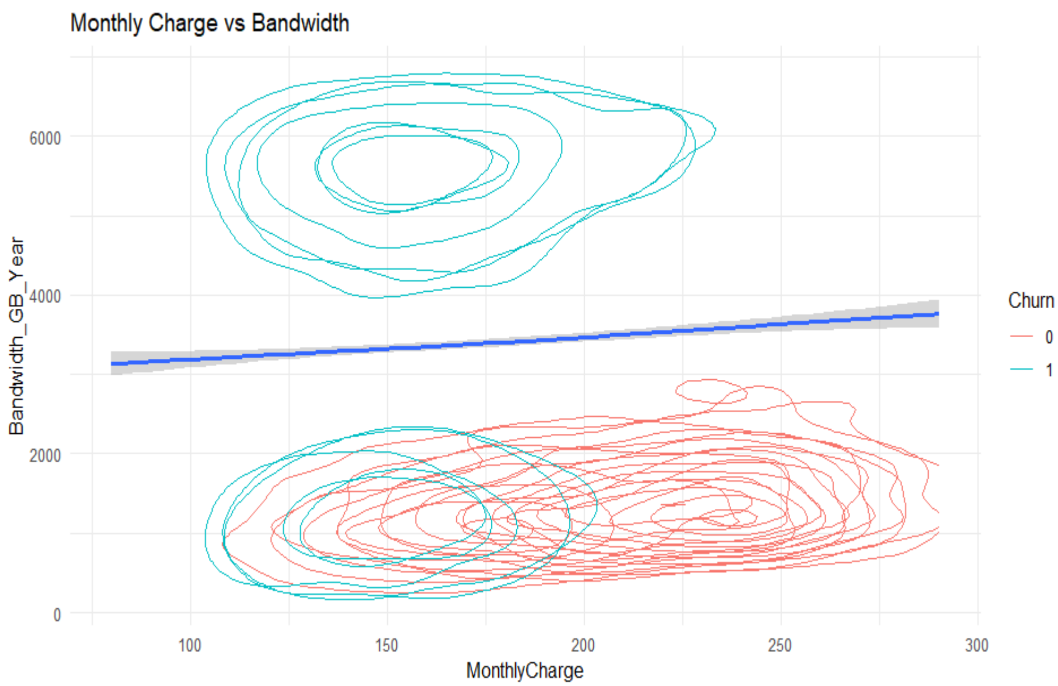


Graphs to support analysis recommendations

#Conclusion visuals

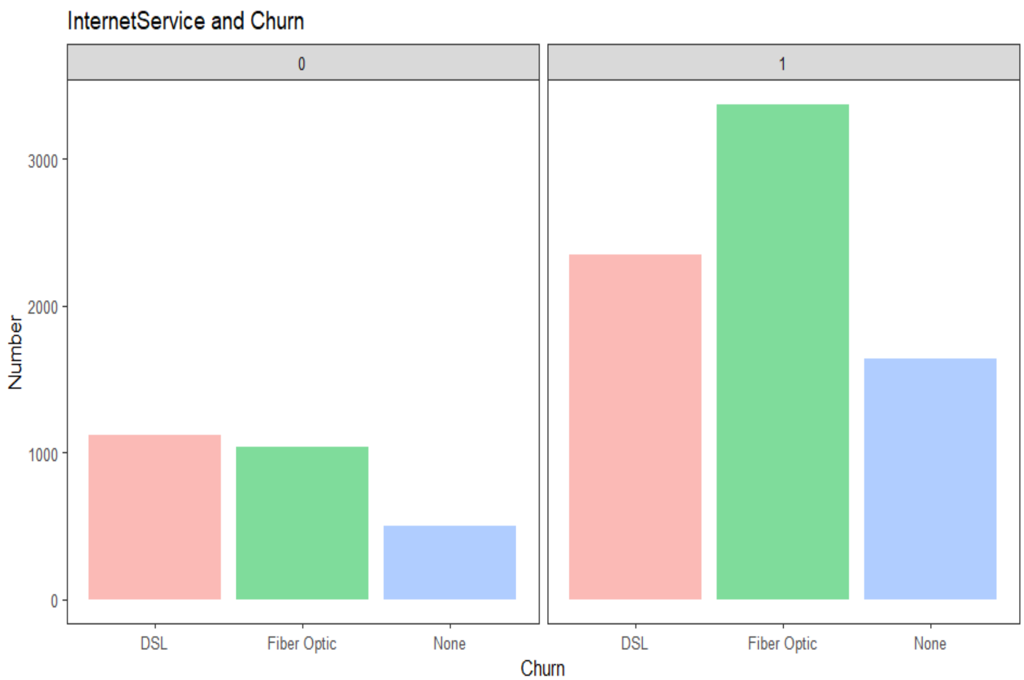
#plot demonstrating the low risk of churn with high bandwidth consumption and a monthly payment around $160 per month

```
dataset %>%
  ggplot(aes(MonthlyCharge, Bandwidth_GB_Year))+
  geom_density_2d(aes(color = Churn,
              Size = Contract)) +
  geom_smooth()+
  labs(x="MonthlyCharge",
     y="Bandwidth_GB_Year",
     title = "Monthly Charge vs Bandwidth")+
  theme_minimal()
```



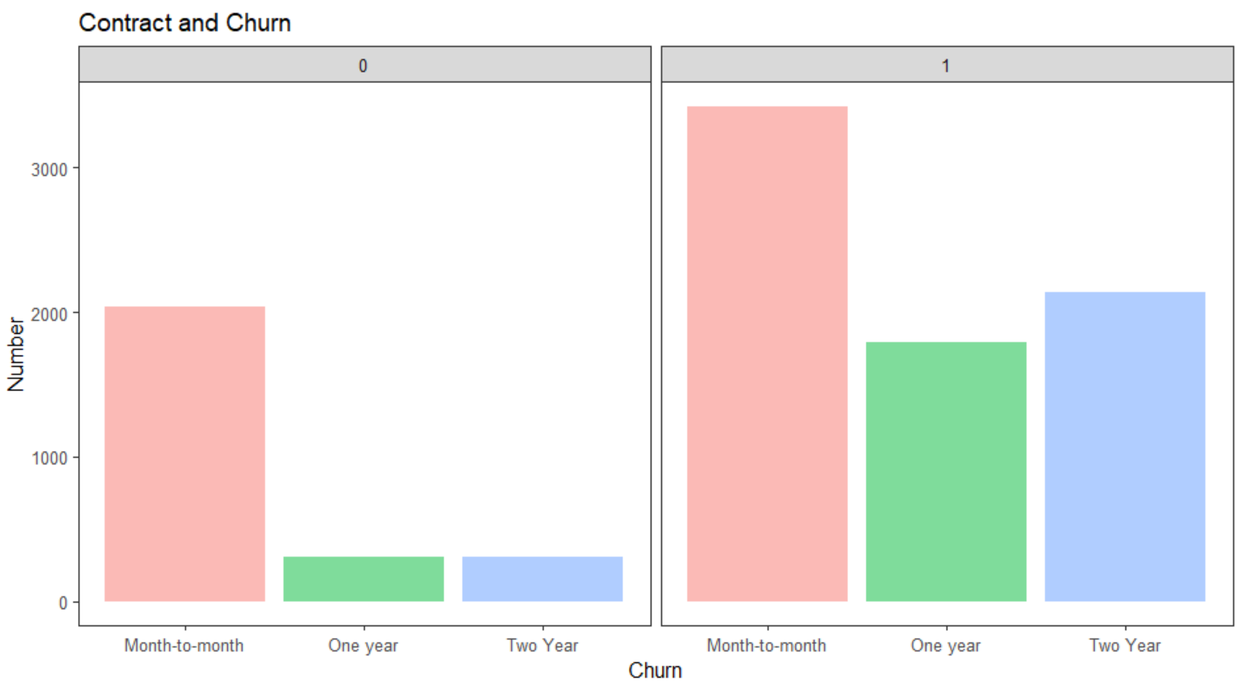

```
# Visual showing the chun ratio of those client with fiber optic internet service

dataset  %>%

  filter(Churn %in% c("0", "1")) %>%

  ggplot(aes(InternetService))+

  geom_bar(aes(fill = InternetService), alpha = 0.5)+

  facet_wrap(~Churn)+

  theme_bw()+

  theme(panel.grid.major = element_blank(),

        panel.grid.minor = element_blank(),
```

```
legend.position = "none")+
labs(title = "InternetService and Churn",
     x = "Churn",
     y = "Number")
```




```
dataset %>%
  filter(Churn %in% c("0", "1")) %>%
  ggplot(aes(Contract))+
  geom_bar(aes(fill = Contract), alpha = 0.5)+
  facet_wrap(~Churn)+
```

```
theme_bw()+

theme(panel.grid.major = element_blank(),

      panel.grid.minor = element_blank(),

      legend.position = "none")+

labs(title = "Contract and Churn",

     x = "Churn",

     y = "Number")
```



D.3.

All code applied

D209 Task 1

```
# Classification template

library(ggplot2)

library(caret)

library(naivebayes)

library(dplyr)

library(knitr)

library(plyr)

library(InformationValue)

#SET WORKING DIRECTORY

setwd("C:/Users/Admin/Desktop/Eric Stuff/WGU/Data Mining I –
D209/WorkingDirectoryTask1")

## Read in original data

originaldata <- read.csv("C:/Users/Admin/Desktop/churn_clean.csv")

str(originaldata)

#subset data to analysis
```

```
dataset <- originaldata[, c("Churn","Contract", "InternetService", "Income",

"Yearly_equip_failure","MonthlyCharge","Bandwidth_GB_Year" )]

#Check analysis subset

summary(dataset)

#Revalue Yes and No to 1 and 0 respectivly

dataset$Churn <- revalue(dataset$Churn, c("Yes"=0))

dataset$Churn <- revalue(dataset$Churn, c("No"=1))

head(dataset)

# Convert column from character to factor

df$Churn <- as.factor(df$Churn)

df$Contract <- as.factor(df$Contract)

df$InternetService <- as.factor(df$InternetService)

#Check subset

print(dataset)

#Export cleaned data set for assignment # C.4.

write.csv(dataset, "Task1CleanData.csv")
```

```
#Checking correlation

pairs.panels(dataset[-1])

#Data partition

set.seed(1234)

ind <- sample(2, nrow(dataset), replace = T, prob = c(0.8, 0.2))

train <- dataset[ind== 1,]

test <- dataset[ind ==2,]

write.csv(train, "Task1training_set.csv")

write.csv(test, "Task1test_set.csv")

# Naive Bayes Model

model <- naive_bayes(Churn ~ ., data = train, usekernel = T)

model

# Summarise continuous variables

train %>%

  filter(Churn == "1") %>%

  summarise(mean(Bandwidth_GB_Year), sd(Bandwidth_GB_Year))
```

```
train %>%
  filter(Churn == "1") %>%
  summarise(mean(MonthlyCharge), sd(MonthlyCharge))
train %>%
  filter(Churn == "1") %>%
  summarise(mean(Yearly_equip_failure), sd(Yearly_equip_failure))
train %>%
  filter(Churn == "1") %>%
  summarise(mean(Income), sd(Income))
plot(model)
# Predict
p <- predict(model, train, type = 'prob')
head(cbind(p,train))
#placing prediction on p1 object and # Creating confusion matrix - train data
p1 <- predict(model, train)
(tab1 <- table(p1, train$Churn))
```

Summing correct predictions and preducing error rate

1 - sum(diag(tab1)) / sum(tab1)

#placing prediction on p2 object and # Creating confusion matrix - test data

p2 <- predict(model, test)

(tab2 <- table(p2, test$Churn))

Summing correct predictions

1 - sum(diag(tab2)) / sum(tab2)

Comparing accurate scores against predictive scores

sensitivity(actuals = ActualsAndScores$Actuals, predictedScores = ActualsAndScores$PredictedScores)

#ks_stat computes the kolmogorov-smirnov statistic that is widely used in scoring to determine the efficacy of binary classification models.

##The higher the ks_stat more effective is the model at capturing the responders.

ks_stat(actuals=ActualsAndScores$Actuals, predictedScores=ActualsAndScores$PredictedScores)

Plotting the kolmogorov-smirnov statistic

```
ks_plot(actuals=ActualsAndScores$Actuals,
predictedScores=ActualsAndScores$PredictedScores)

### Code =for future analysis

#Weight of Evidence (WoE) and Information Value (IV) of a variable in respect to a binary
outcome.

##options(scipen = 999, digits = 2)

##WOETable(X=SimData$X.Cat, Y=SimData$Y.Binary)

# Produce ROC chart

plotROC(actuals=ActualsAndScores$Actuals,
predictedScores=ActualsAndScores$PredictedScores)

#Conclusion visuals

#plot demonstrating the low risk of churn  with high bandwidth consumption and a monthly
payment  around $160 per month

dataset %>%

  ggplot(aes(MonthlyCharge, Bandwidth_GB_Year))+

  geom_density_2d(aes(color = Churn,

                Size = Contract)) +
```

```
  geom_smooth()+
  labs(x="MonthlyCharge",
       y="Bandwidth_GB_Year",
       title = "Monthly Charge vs Bandwidth")+
  theme_minimal()
# Visual showing the chun ratio of those client with fiber optic internet service
dataset  %>%
  filter(Churn %in% c("0", "1")) %>%
  ggplot(aes(InternetService))+
  geom_bar(aes(fill = InternetService), alpha = 0.5)+
  facet_wrap(~Churn)+
  theme_bw()+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "none")+
  labs(title = "InternetService and Churn",
```

```
      x = "Churn",
      y = "Number")
dataset  %>%
  filter(Churn %in% c("0", "1")) %>%
  ggplot(aes(Contract))+
  geom_bar(aes(fill = Contract), alpha = 0.5)+
  facet_wrap(~Churn)+
  theme_bw()+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "none")+
  labs(title = "Contract and Churn",
      x = "Churn",
      y = "Number")
```

END CODE

Part V: Data Summary and Implications

Summary and Implications

E. Summarize your data analysis by doing the following:

E.1.

The accuracy as demonstrated by this Naïve Bayes analysis substantiates that there is an approximate 86.9% ratio for accurate predictions, with an error ratio of approximately 13%. Both metrics are clearly visible in the ROC curve visual, the confusion matrix, the kolmogorov-smirnov statistic, and the demonstrated sensitivity metric.

E.2.

Results and Implications

The results are convincing that the independent attributes utilized in conjunction with the outcome variable and the Naïve Bayes algorithm produce vary optimistic classification prediction results. Leading to conclusive implications that we as an organization can utilize this business intelligence to strengthen our current market position. Pleas see suggestions and recommendation in section E.4.

E.3.

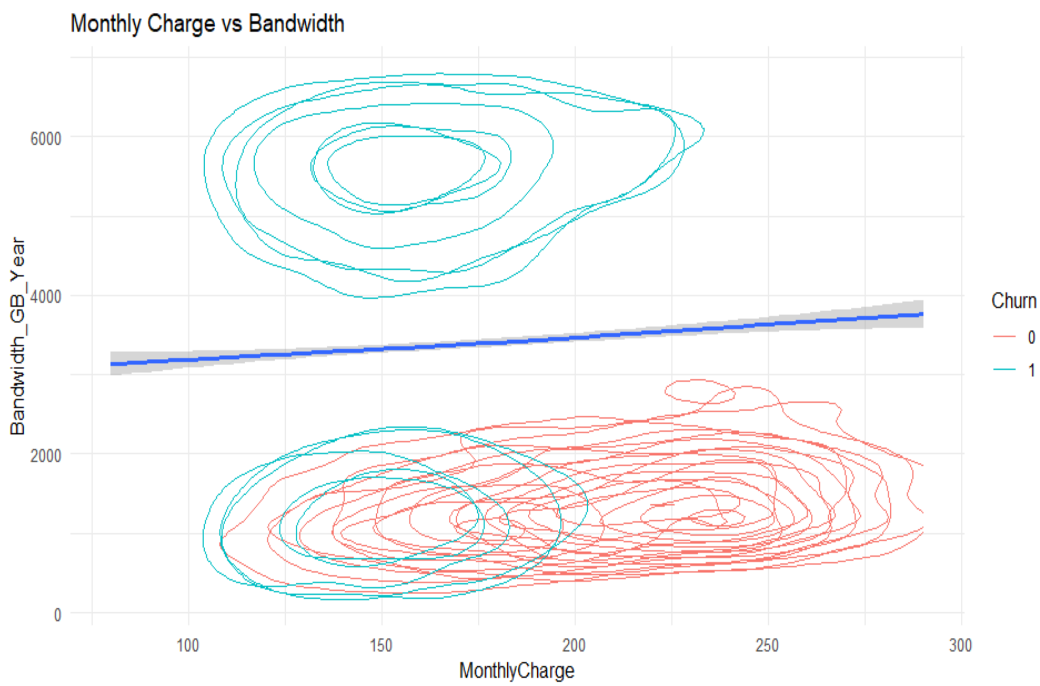
Analysis Limitations

There are several limitations concerning this analysis. But the primary constraint plaguing this analysis and the contemporary telecommunications industry is the perpetually changing ecosystem. This analysis yields very persuasive business intelligence that now has the potential to provide substantial results. But with the rapid industry innovations that could all change overnight, what today is revolutionary tomorrow is outdated.

E.4. Recommendation and suggestions

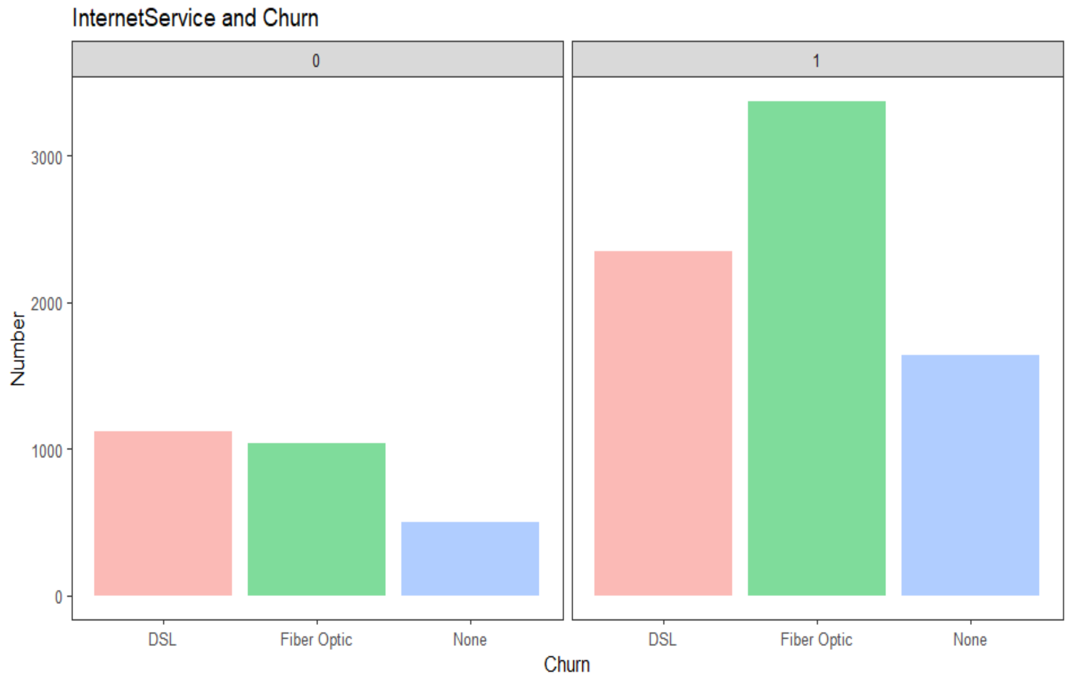
Reasoning Point #1

As the chart below exhibits by the blue circle located in the upper hemisphere, the least susceptible to churn are those clients who consume a high amount of bandwidth (approx.6000 gigabytes per year) and retain approximately $160 as their monthly payment.



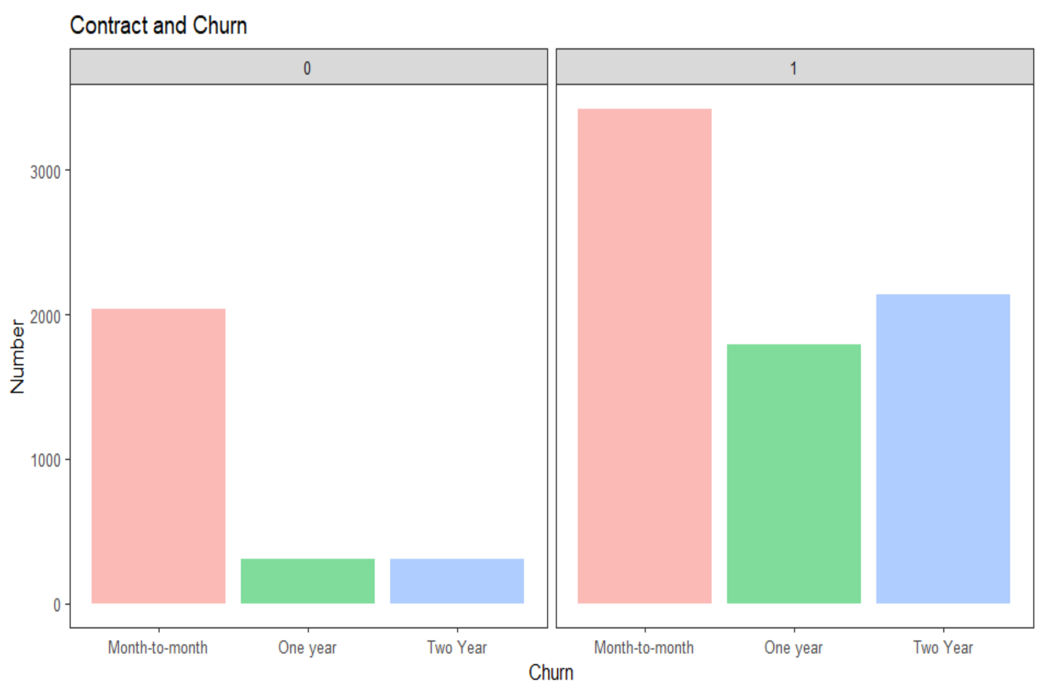
Reasoning Point #2

As the next chart exhibits with a high percentage of no-churn clients also happen to utilize fiber optic for their internet service.



Reasoning Point #3

As the final chart below demonstrations although month-to-month contracts have both the highest churn and no churn percentages the two-year contracts have a definite no churn attribute considering all clients from both categories.



Recommended course of action

This Naïve Bayes analysis has clearly brought to my notice with strong statistically significant evidence what I have sought to established with the 3 easy to comprehend reasoning points above.

Suggestion

Offer a plan the provides high bandwidth capacity, fiber optic internet service for approximately $160 per month with a two-year contract commitment.

This should lead us to retain clients far longer than our average, developing our business control and prospering our position in the telecommunications arena.

Part VI: Demonstration

F. Panopto video recording attached

References

admin. (n.d.). *Top Reasons For Why Should You Use R for Data Science*. Retrieved October 17, 2021, from https://statanalytica.com/blog/top-5-reasons-to-use-r-language-for-data-science/

Data-Preprocessing Technique - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. Retrieved October 17, 2021, from https://www.sciencedirect.com/topics/computer-science/data-preprocessing-technique

Flatiron School Team, F. S. (2006, February 22). *Dsc 2 22 06 Naive Bayes Assumptions - Learn.co*. Learn - A Platform for Education. https://portal.flatironschool.com/lessons/dsc-2-22-06-naive-bayes-assumptions

Obaidat, Mohammad S., et al. *Simulation and Modeling Methodologies, Technologies and Applications: International Conference, SIMULTECH 2013, Reykjavik, Iceland, July 29-31, 2013: Revised Selected Papers*. Springer, 2015.

Papadopoulos, Alecos. "Correlation and Naive Bayes." *Data Science Stack Exchange*, 1 Mar. 1964, https://datascience.stackexchange.com/questions/9087/correlation-and-naive-bayes.

Stecanella, B. S. (2017, May 25). *A practical explanation of a Naive Bayes classifier*. MonkeyLearn Blog. https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

Timbers, T. (2021, October 3). *Chapter 6 Classification II: evaluation & tuning | Data Science: A First Introduction*. Github. https://ubc-dsci.github.io/introduction-to-datascience/classification2.html