DSBI Assignment #3

Your Name

# Set-up

Load packages.

rm(list = ls())  
library(readxl)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)

For this and the next assignments (Assignment 3 and 4), you will use the dataset from <https://www.ibm.com/communities/analytics/watson-analytics-blog/it-help-desk/>. (the link is now invalid.) This data is from an Information Technology (IT) department interested in improving the satisfaction of customers.  
To start their analysis, they constructed this data set of 100,000 closed tickets that were filed at their help desk. The dataset includes IT service requests and their feedback (result), and this firm wants to improve its IT service quality.

Load the dataset and save it as ithelp.

ithelp <- read\_excel("WA\_Fn-UseC\_-IT-Help-Desk.xlsx")

glimpse(ithelp)

## Rows: 100,000  
## Columns: 10  
## $ ticket <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, …  
## $ Requestor <dbl> 1929, 1587, 925, 413, 318, 858, 1978, 1209, 887, 17…  
## $ RequestorSeniority <chr> "1 - Junior", "2 - Regular", "2 - Regular", "4 - Ma…  
## $ ITOwner <dbl> 50, 15, 15, 22, 22, 38, 10, 1, 14, 46, 1, 50, 11, 2…  
## $ FiledAgainst <chr> "Systems", "Software", "Access/Login", "Systems", "…  
## $ TicketType <chr> "Issue", "Request", "Request", "Request", "Request"…  
## $ Severity <chr> "2 - Normal", "1 - Minor", "2 - Normal", "2 - Norma…  
## $ Priority <chr> "0 - Unassigned", "1 - Low", "0 - Unassigned", "0 -…  
## $ daysOpen <dbl> 3, 5, 0, 20, 1, 0, 9, 15, 6, 1, 7, 17, 10, 4, 7, 11…  
## $ Satisfaction <chr> "1 - Unsatisfied", "1 - Unsatisfied", "0 - Unknown"…

A total of 100,000 rows. Each row represent an individual request. The data includes:

* Requestor: employee who submitted the ticket
* RequestorSeniority: employee’s seniority within the company
* ITOwner: IT employee who serviced the ticket
* FileAgainst: functional area against which the ticket was filed (systems, software, hardware, access)
* TicketType: whether the ticket was a request for new services or an issue with existing services
* Severity: submitter-assigned severity of the ticket
* Priority: IT-assigned priority of the ticket
* daysOpen: number of days the ticket was open
* Satisfaction: satisfaction with the resolution of the ticket (reported by the submitter)

summary(ithelp)

## ticket Requestor RequestorSeniority ITOwner   
## Min. : 1 Min. : 1 Length:100000 Min. : 1.00   
## 1st Qu.: 25001 1st Qu.: 499 Class :character 1st Qu.:13.00   
## Median : 50000 Median : 999 Mode :character Median :26.00   
## Mean : 50000 Mean : 999 Mean :25.46   
## 3rd Qu.: 75000 3rd Qu.:1499 3rd Qu.:38.00   
## Max. :100000 Max. :2000 Max. :50.00   
## FiledAgainst TicketType Severity Priority   
## Length:100000 Length:100000 Length:100000 Length:100000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## daysOpen Satisfaction   
## Min. : 0.000 Length:100000   
## 1st Qu.: 1.000 Class :character   
## Median : 5.000 Mode :character   
## Mean : 6.843   
## 3rd Qu.:10.000   
## Max. :54.000

Some ‘character’ variables behave unexpectedly. Let’s convert all character variables into factors, which indicate categorical variables in R.

ithelp<-ithelp%>%  
 mutate(ticket=as.factor(ticket),  
 Requestor=as.factor(Requestor),  
 RequestorSeniority=as.factor(RequestorSeniority),  
 ITOwner=as.factor(ITOwner),  
 FiledAgainst=as.factor(FiledAgainst),  
 TicketType=as.factor(TicketType),  
 Severity=as.factor(Severity),  
 Priority=as.factor(Priority),  
 Satisfaction=as.factor(Satisfaction))

Let’s check the summary again.

summary(ithelp)

## ticket Requestor RequestorSeniority ITOwner   
## 1 : 1 285 : 74 1 - Junior :20040 3 : 2084   
## 2 : 1 754 : 74 2 - Regular :41303 39 : 2080   
## 3 : 1 79 : 73 3 - Senior :18801 48 : 2072   
## 4 : 1 636 : 72 4 - Management:19856 35 : 2058   
## 5 : 1 1341 : 72 24 : 2045   
## 6 : 1 523 : 70 4 : 2040   
## (Other):99994 (Other):99565 (Other):87621   
## FiledAgainst TicketType Severity   
## Access/Login:29921 Issue :24926 0 - Unclassified: 367   
## Hardware : 9976 Request:75074 1 - Minor : 2317   
## Software :20068 2 - Normal :90912   
## Systems :40035 3 - Major : 4974   
## 4 - Critical : 1430   
##   
##   
## Priority daysOpen Satisfaction   
## 0 - Unassigned:30127 Min. : 0.000 0 - Unknown :30211   
## 1 - Low :17117 1st Qu.: 1.000 1 - Unsatisfied :21124   
## 2 - Medium :16258 Median : 5.000 2 - Satisfied :19602   
## 3 - High :36498 Mean : 6.843 3 - Highly satisfied:29063   
## 3rd Qu.:10.000   
## Max. :54.000   
##

It turns out that there are many cases with “Unknown” Satisfaction. Let’s exclude these cases from the analysis (Step 1: filter). Next, we may build a multi-class classification model (Unsatisfied, Satisfied, Highly satisfied), but let’s simplify it to a binary classifier and identify “Unsatisfied” cases, which are problematic. Create a new variable, “negative”, which indicates if a user’s feedback is negative (Step 2: mutate). The following code will do these jobs for you.

ithelp<-ithelp%>%  
 filter(Satisfaction!="0 - Unknown")%>%  
 mutate(negative=as.factor(ifelse(Satisfaction =="1 - Unsatisfied","Yes","No")))  
  
summary(ithelp)

## ticket Requestor RequestorSeniority ITOwner   
## 1 : 1 442 : 56 1 - Junior :13995 39 : 1470   
## 2 : 1 1489 : 55 2 - Regular :28963 24 : 1442   
## 5 : 1 98 : 53 3 - Senior :13027 19 : 1440   
## 9 : 1 248 : 53 4 - Management:13804 3 : 1437   
## 10 : 1 1391 : 53 48 : 1437   
## 12 : 1 1498 : 53 31 : 1435   
## (Other):69783 (Other):69466 (Other):61128   
## FiledAgainst TicketType Severity   
## Access/Login:20990 Issue :17454 0 - Unclassified: 263   
## Hardware : 6912 Request:52335 1 - Minor : 1620   
## Software :13972 2 - Normal :63459   
## Systems :27915 3 - Major : 3445   
## 4 - Critical : 1002   
##   
##   
## Priority daysOpen Satisfaction   
## 0 - Unassigned:20984 Min. : 0.00 0 - Unknown : 0   
## 1 - Low :12045 1st Qu.: 1.00 1 - Unsatisfied :21124   
## 2 - Medium :11409 Median : 5.00 2 - Satisfied :19602   
## 3 - High :25351 Mean : 6.83 3 - Highly satisfied:29063   
## 3rd Qu.:10.00   
## Max. :49.00   
##   
## negative   
## No :48665   
## Yes:21124   
##   
##   
##   
##   
##

# Q1. Classification tree model

library(rpart)  
library(rpart.plot)

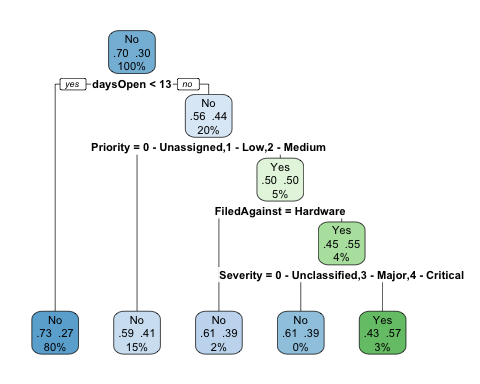
Build a model to predict “negative” with the following variables. Fill in the blank to complete the code below. \* RequestorSeniority \* FiledAgainst \* TicketType \* Severity \* Priority \* daysOpen

Set the control as control=rpart.control(cp=0.001).

ct\_model <- rpart(  
 negative ~ RequestorSeniority + FiledAgainst + TicketType + Severity + Priority + daysOpen, # model formula  
 data = ithelp, # dataset  
 method = "class", # classification tree  
 control = rpart.control(cp = 0.001) # tree control parameters  
)

Display the tree model with *rpart.plot* and answer the following questions based on this model result.

rpart.plot(ct\_model, type = 2, extra = 104, fallen.leaves = TRUE)



Question 1: Which one of the following characterizes the case that receive negative feedback (i.e. the probability of negative feedback greater than 50%)?

1. Not closed (i.e. resolved) within 12 days (daysOpen=>13 days)
2. Priority is High
3. Not filed against Hardware (i.e. filed against Access/Login, Software, or Systems)
4. Severity is either minor or normal
5. All conditions above should be met.

Question 2: What is the probability that a case receives negative feedback if the case is resolved within 12 days? (1) 27% (2) 41% (3) 80% (4) Cannot know.

# Logit Regression Model

Build a logit regression model to predict “negative” with the following variables. Fill in the blank to complete the code below. \* RequestorSeniority \* FiledAgainst \* TicketType \* Severity \* Priority \* daysOpen

# Build a logistic regression model  
logit\_model <- glm(  
 negative ~ RequestorSeniority + FiledAgainst + TicketType + Severity + Priority + daysOpen, # model formula  
 data = ithelp, # dataset  
 family = binomial # logistic regression  
)  
  
# View summary of the model  
summary(logit\_model)

##   
## Call:  
## glm(formula = negative ~ RequestorSeniority + FiledAgainst +   
## TicketType + Severity + Priority + daysOpen, family = binomial,   
## data = ithelp)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.165590 0.131668 -1.258 0.208524   
## RequestorSeniority2 - Regular 0.014075 0.023388 0.602 0.547322   
## RequestorSeniority3 - Senior -0.060720 0.029196 -2.080 0.037552 \*   
## RequestorSeniority4 - Management -0.020905 0.029468 -0.709 0.478064   
## FiledAgainstHardware -1.544872 0.045352 -34.064 < 2e-16 \*\*\*  
## FiledAgainstSoftware -0.494240 0.026275 -18.811 < 2e-16 \*\*\*  
## FiledAgainstSystems -0.778581 0.026288 -29.618 < 2e-16 \*\*\*  
## TicketTypeRequest -0.399541 0.020834 -19.177 < 2e-16 \*\*\*  
## Severity1 - Minor 0.163199 0.138826 1.176 0.239769   
## Severity2 - Normal -0.397980 0.129497 -3.073 0.002117 \*\*   
## Severity3 - Major -0.997003 0.135929 -7.335 2.22e-13 \*\*\*  
## Severity4 - Critical -0.974058 0.150456 -6.474 9.54e-11 \*\*\*  
## Priority1 - Low -0.097741 0.026535 -3.683 0.000230 \*\*\*  
## Priority2 - Medium 0.041825 0.025866 1.617 0.105890   
## Priority3 - High 0.075306 0.021471 3.507 0.000453 \*\*\*  
## daysOpen 0.084106 0.001716 49.027 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 85578 on 69788 degrees of freedom  
## Residual deviance: 82753 on 69773 degrees of freedom  
## AIC: 82785  
##   
## Number of Fisher Scoring iterations: 4

Question 3: Which of the following is NOT an implication that you can draw from this logit regression result?

1. An increase in the daysOpen is associated with a higher chance of negative feedback.
2. A request with high priority is more likely to receive negative feedback.
3. Compared to Junior, a request from Senior is more likely to receive negative feedback.
4. All of the above

Question 4. Save the predicted probability from a logit model by creating a new variable *logit\_pred* in the dataset. What is the probability that the ticket number 1404 receives negative feedback?

# Save predictions  
ithelp$logit\_pred <- predict(logit\_model, type = "response")  
  
# Find ticket 1404  
ithelp %>%  
 filter(ticket == 1404) %>%  
 select(logit\_pred)

## # A tibble: 1 × 1  
## logit\_pred  
## <dbl>  
## 1 0.235

1. 18%
2. 23%
3. 53%
4. 118%

Question 5. The same as the previous assignments, your last task is creating a report. Change the author name on the top of this R markdown file to yours. Compile this R markdown file into a Word document or html and submit it through the course Canvas. Your report should contain *all codes and results* of the assignment questions and answers.