## By Yashvardhan Rathi

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.1

## -- Attaching packages ----------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.5.1

## Warning: package 'tibble' was built under R version 3.5.1

## Warning: package 'tidyr' was built under R version 3.5.1

## Warning: package 'readr' was built under R version 3.5.1

## Warning: package 'purrr' was built under R version 3.5.1

## Warning: package 'dplyr' was built under R version 3.5.1

## Warning: package 'stringr' was built under R version 3.5.1

## Warning: package 'forcats' was built under R version 3.5.1

## -- Conflicts -------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(gains)  
library(leaps)

## Warning: package 'leaps' was built under R version 3.5.1

library(caret)

## Warning: package 'caret' was built under R version 3.5.1

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(ggplot2)

### Read the Dataset

syst\_admin.df <- read.csv("SystemAdministrators.csv")

## Question 1

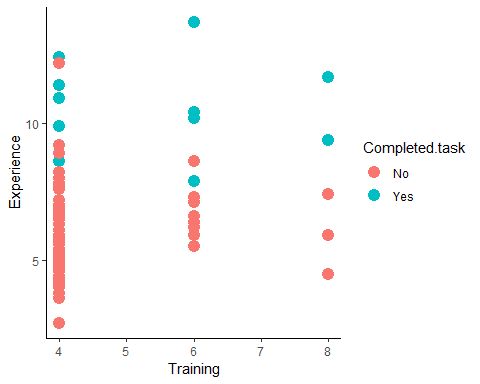
### Convert ‘Completed Task’ to categorical variable

factor(syst\_admin.df$Completed.task)

## [1] Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes Yes No No   
## [18] No No No No No No No No No No No No No No No No No   
## [35] No No No No No No No No No No No No No No No No No   
## [52] No No No No No No No No No No No No No No No No No   
## [69] No No No No No No No   
## Levels: No Yes

### Scatterplot using ggplot2

ggplot(syst\_admin.df) +  
 geom\_point(aes(x = Training, y = Experience, color = Completed.task), size = 4) + theme\_classic()



*As seen from the output we notice that when predictor ‘Experience’ has a higher value it is more likely that the task will be completed. Hence ‘Experience’ is a potentially useful predictor to classify task completion.*

## Question 2

### Logistic Regression Model

logit\_reg<- glm(Completed.task ~ ., data = syst\_admin.df, family = 'binomial')  
logit\_reg

##   
## Call: glm(formula = Completed.task ~ ., family = "binomial", data = syst\_admin.df)  
##   
## Coefficients:  
## (Intercept) Experience Training   
## -10.9813 1.1269 0.1805   
##   
## Degrees of Freedom: 74 Total (i.e. Null); 72 Residual  
## Null Deviance: 75.06   
## Residual Deviance: 35.71 AIC: 41.71

options(scipen=999)  
summary(logit\_reg)

##   
## Call:  
## glm(formula = Completed.task ~ ., family = "binomial", data = syst\_admin.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.65306 -0.34959 -0.17479 -0.08196 2.21813   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -10.9813 2.8919 -3.797 0.000146 \*\*\*  
## Experience 1.1269 0.2909 3.874 0.000107 \*\*\*  
## Training 0.1805 0.3386 0.533 0.593970   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 75.060 on 74 degrees of freedom  
## Residual deviance: 35.713 on 72 degrees of freedom  
## AIC: 41.713  
##   
## Number of Fisher Scoring iterations: 6

### Evaluate Performance of the Logit Model

logit\_reg\_pred <- predict(logit\_reg, syst\_admin.df, type = "response")  
  
t(t(head(logit\_reg\_pred, 15)))

## [,1]  
## 1 0.88332273  
## 2 0.71040418  
## 3 0.86078581  
## 4 0.99609157  
## 5 0.74191004  
## 6 0.97621610  
## 7 0.26982801  
## 8 0.44285501  
## 9 0.83152106  
## 10 0.93006865  
## 11 0.36177639  
## 12 0.52709747  
## 13 0.97461443  
## 14 0.15517272  
## 15 0.08543037

### Generate confusion matrix

conf\_table <- table(syst\_admin.df$Completed.task , logit\_reg\_pred > 0.5)  
conf\_table

##   
## FALSE TRUE  
## No 58 2  
## Yes 5 10

### Percentage of programmers incorrectly classified as failing to complete task.

Sens <- (conf\_table[2,1]/(conf\_table[2,1] + conf\_table[2,2]))  
Sens

## [1] 0.3333333

Sens\_percentage <- Sens\*100  
Sens\_percentage

## [1] 33.33333

*The percentage of programmers incorrectly classified as failing to complete task is 33.3333333.*

## Question 3

### Logistic Regression equation

### Assigning values which we got with p=0.6 and x\_2 = 6

### We need to find the value of X1. Reubstituting the values in the equation for X1:

LHS <- log(0.6/(1-0.6))  
X1 <- (LHS - (0.1805 \* 6) + 10.9813)/1.1269  
X1

## [1] 9.14346

*The number of years of experience a programmer with 6 years of training needs so that his or her estimated probability of completing the task exceeds 0.6 is 9.14346 years.*