Classification

Team: Abed Ahmed (ASA190005) & Dylan Kapustka (DLK190000)

Date: 09/25/2022

HomeWork 3 - Classification

Logistic Regression

Linear Models for Logistic regression answer classifer questions as opposed to "how much" questions in Linear Regression. Logisitc Regression is also known as classification because it uses a certain independant variable(s) **x** to classify a dependant variable **y** into some entity from a finite list. **Y** could be binary (Yes or NO), (Healthy not Healthy) or there could be multiple classifiers (different traffice signs).

Pros of Logistic Regression

- Easy To implement and train
- Makes no assumptions about distributions of classes
- Extensible to multiple classes
- Good accuracy for many simple datasets

Cons of Logistic Regression

- Constructs linear boundaries
- Major limitation is the assumption of linearity between dependant and independant variables.
- Can only be used to predict discrete functions
- Requires average or no multicollinearity between independent variables

The below chunk does the following

- Imports and cleans data
- Splits into 80/20 train/test
- Explores the data
- Plots two informative graphs

```
data <- read.csv(file="C:/Users/Abed/Documents/RegressionProject/companyData.csv",header=TRUE)
data <- na.omit(data)
data$Attrition <-  as.numeric(as.factor(data$Attrition))
for(i in 1:length(data$Attrition)){
  if(data$Attrition[i] == 1){
    data$Attrition[i] = 0
  }
  else data$Attrition[i] = 1
```

```
}
set.seed(2)
library(caTools)


split <- sample.split(data,SplitRatio=0.8)
train <- subset(data, split==TRUE)
test <- subset(data,split==FALSE)

names(train)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "Department"
##  [5] "DistanceFromHome"         "Education"
##  [7] "EducationField"           "EmployeeCount"
##  [9] "EmployeeID"               "Gender"
## [11] "JobLevel"                 "JobRole"
## [13] "MaritalStatus"            "MonthlyIncome"
## [15] "NumCompaniesWorked"       "Over18"
## [17] "PercentSalaryHike"        "StandardHours"
## [19] "StockOptionLevel"         "TotalWorkingYears"
## [21] "TrainingTimesLastYear"    "YearsAtCompany"
## [23] "YearsSinceLastPromotion" "YearsWithCurrManager"
```

```
head(train)
```

```
##   Age Attrition    BusinessTravel              Department DistanceFromHome
## 1  51         0     Travel_Rarely                   Sales                6
## 2  31         1 Travel_Frequently Research & Development               10
## 3  32         0 Travel_Frequently Research & Development               17
## 4  38         0        Non-Travel Research & Development                2
## 7  28         1     Travel_Rarely Research & Development               11
## 8  29         0     Travel_Rarely Research & Development               18
##   Education EducationField EmployeeCount EmployeeID Gender JobLevel
## 1         2  Life Sciences             1          1 Female        1
## 2         1  Life Sciences             1          2 Female        1
## 3         4          Other             1          3   Male        4
## 4         5  Life Sciences             1          4   Male        3
## 7         2        Medical             1          7   Male        2
## 8         3  Life Sciences             1          8   Male        2
##                      JobRole MaritalStatus MonthlyIncome NumCompaniesWorked
## 1 Healthcare Representative       Married        131160                  1
## 2        Research Scientist        Single         41890                  0
## 3           Sales Executive       Married        193280                  1
## 4           Human Resources       Married         83210                  3
## 7           Sales Executive        Single         58130                  2
## 8           Sales Executive       Married         31430                  2
##   Over18 PercentSalaryHike StandardHours StockOptionLevel TotalWorkingYears
## 1      Y               11             8               0                 1
## 2      Y               23             8               1                 6
## 3      Y               15             8               3                 5
## 4      Y               11             8               3                13
## 7      Y               20             8               1                 5
```

```
## 8      Y                22             8                3                10
##   TrainingTimesLastYear YearsAtCompany YearsSinceLastPromotion
## 1                     6             1                       0
## 2                     3             5                       1
## 3                     2             5                       0
## 4                     5             8                       7
## 7                     2             0                       0
## 8                     2             0                       0
##   YearsWithCurrManager
## 1                    0
## 2                    4
## 3                    3
## 4                    5
## 7                    0
## 8                    0
```

```
summary(train)
```

```
##       Age           Attrition      BusinessTravel      Department
##  Min.   :18.00   Min.   :0.0000   Length:3470        Length:3470
##  1st Qu.:30.00   1st Qu.:0.0000   Class :character   Class :character
##  Median :36.00   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :36.97   Mean   :0.1576
##  3rd Qu.:43.00   3rd Qu.:0.0000
##  Max.   :60.00   Max.   :1.0000
##  DistanceFromHome   Education     EducationField     EmployeeCount
##  Min.   : 1.000   Min.   :1.000   Length:3470        Min.   :1
##  1st Qu.: 2.000   1st Qu.:2.000   Class :character   1st Qu.:1
##  Median : 7.000   Median :3.000   Mode  :character   Median :1
##  Mean   : 9.235   Mean   :2.924                      Mean   :1
##  3rd Qu.:14.000   3rd Qu.:4.000                      3rd Qu.:1
##  Max.   :29.000   Max.   :5.000                      Max.   :1
##    EmployeeID      Gender             JobLevel       JobRole
##  Min.   :   1   Length:3470        Min.   :1.000   Length:3470
##  1st Qu.:1108   Class :character   1st Qu.:1.000   Class :character
##  Median :2210   Mode  :character   Median :2.000   Mode  :character
##  Mean   :2208                      Mean   :2.051
##  3rd Qu.:3310                      3rd Qu.:3.000
##  Max.   :4409                      Max.   :5.000
##  MaritalStatus      MonthlyIncome    NumCompaniesWorked    Over18
##  Length:3470        Min.   : 10090   Min.   :0.000      Length:3470
##  Class :character   1st Qu.: 29090   1st Qu.:1.000      Class :character
##  Mode  :character   Median : 49035   Median :2.000      Mode  :character
##                     Mean   : 65098   Mean   :2.704
##                     3rd Qu.: 83460   3rd Qu.:4.000
##                     Max.   :199990   Max.   :9.000
##  PercentSalaryHike StandardHours StockOptionLevel TotalWorkingYears
##  Min.   :11.0      Min.   :8     Min.   :0.0000   Min.   : 0.0
##  1st Qu.:12.0      1st Qu.:8     1st Qu.:0.0000   1st Qu.: 6.0
##  Median :14.0      Median :8     Median :1.0000   Median :10.0
##  Mean   :15.2      Mean   :8     Mean   :0.7963   Mean   :11.3
##  3rd Qu.:18.0      3rd Qu.:8     3rd Qu.:1.0000   3rd Qu.:15.0
##  Max.   :25.0      Max.   :8     Max.   :3.0000   Max.   :40.0
##  TrainingTimesLastYear YearsAtCompany   YearsSinceLastPromotion
```

```
##  Min.   :0.000         Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000         1st Qu.: 3.000   1st Qu.: 0.000
##  Median :3.000         Median : 5.000   Median : 1.000
##  Mean   :2.784         Mean   : 6.976   Mean   : 2.169
##  3rd Qu.:3.000         3rd Qu.: 9.000   3rd Qu.: 3.000
##  Max.   :6.000         Max.   :37.000   Max.   :15.000
##  YearsWithCurrManager
##  Min.   : 0.000
##  1st Qu.: 2.000
##  Median : 3.000
##  Mean   : 4.108
##  3rd Qu.: 7.000
##  Max.   :17.000
```

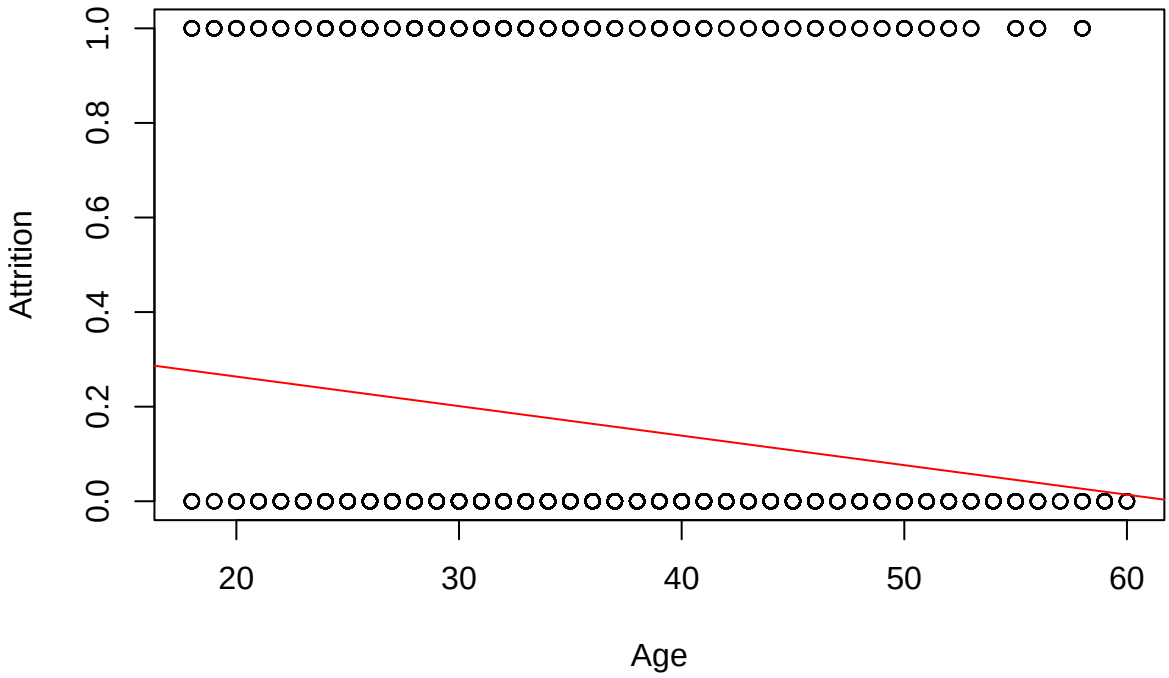
```
dim(train)
```

```
## [1] 3470   24
```

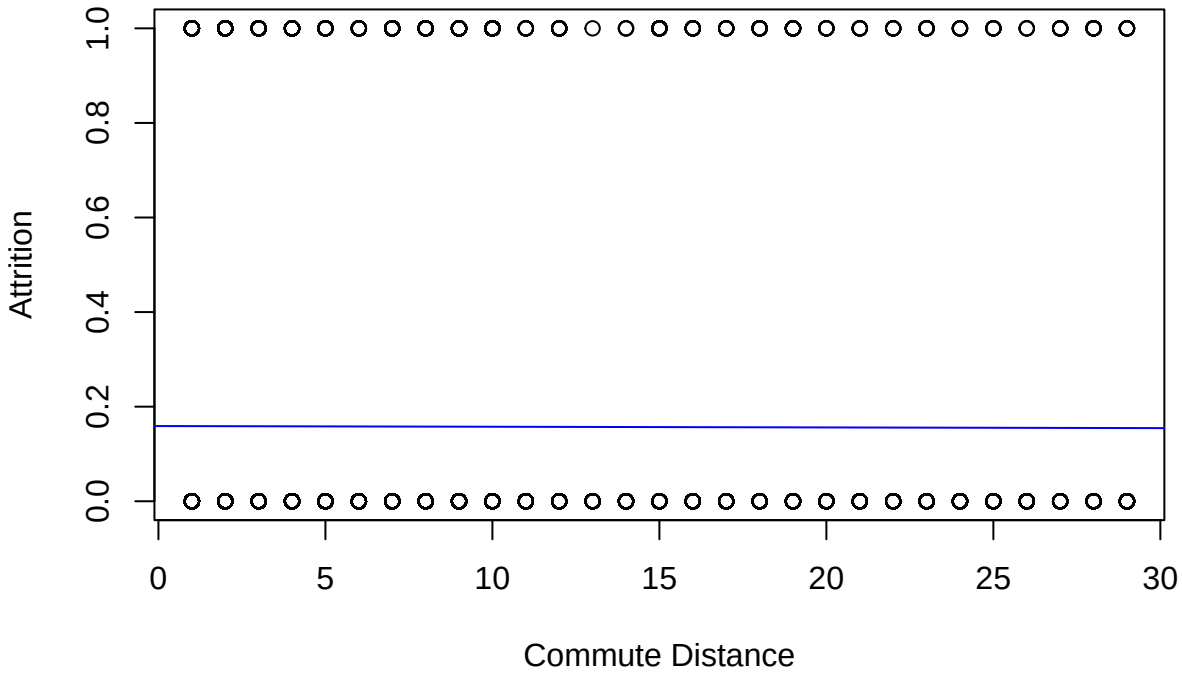
```
str(train)
```

```
## 'data.frame':    3470 obs. of  24 variables:
##  $ Age                     : int  51 31 32 38 28 29 31 25 45 36 ...
##  $ Attrition               : num  0 1 0 0 1 0 0 0 0 0 ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Frequently" "Non-Travel
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Research
##  $ DistanceFromHome        : int  6 10 17 2 11 18 1 7 17 28 ...
##  $ Education               : int  2 1 4 5 2 3 3 4 2 1 ...
##  $ EducationField          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeID              : int  1 2 3 4 7 8 9 10 11 12 ...
##  $ Gender                  : chr  "Female" "Female" "Male" "Male" ...
##  $ JobLevel                : int  1 1 4 3 2 2 3 4 2 1 ...
##  $ JobRole                 : chr  "Healthcare Representative" "Research Scientist" "Sales Executive"
##  $ MaritalStatus           : chr  "Married" "Single" "Married" "Married" ...
##  $ MonthlyIncome           : int  131160 41890 193280 83210 58130 31430 20440 134640 79910 33770 ...
##  $ NumCompaniesWorked      : int  1 0 1 3 2 2 0 1 0 0 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ PercentSalaryHike       : int  11 23 15 11 20 22 21 13 13 12 ...
##  $ StandardHours           : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ StockOptionLevel        : int  0 1 3 3 1 3 0 1 2 2 ...
##  $ TotalWorkingYears       : int  1 6 5 13 5 10 10 6 21 16 ...
##  $ TrainingTimesLastYear   : int  6 3 2 5 2 2 2 2 2 2 ...
##  $ YearsAtCompany          : int  1 5 5 8 0 0 9 6 20 15 ...
##  $ YearsSinceLastPromotion: int  0 1 0 7 0 0 7 1 4 10 ...
##  $ YearsWithCurrManager    : int  0 4 3 5 0 0 8 5 10 11 ...
```

```
plot(train$Attrition~train$Age,xlab="Age",ylab="Attrition")
abline(lm(train$Attrition~train$Age),col="red")
```




```
plot(train$Attrition~train$DistanceFromHome,xlab="Commute Distance",ylab="Attrition")
abline(lm(train$Attrition~train$DistanceFromHome),col="blue")
```



This chunk will

- Build a simple linear model of the data
- outputs the summary

```
lm <- glm(Attrition ~ Age + DistanceFromHome + TotalWorkingYears, data = train, family = 'binomial')
summary(lm)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + DistanceFromHome + TotalWorkingYears,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8979  -0.6345  -0.5461  -0.3682   2.8030
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.278398   0.216434  -1.286  0.19834
## Age               -0.023377   0.007177  -3.257  0.00112 **
## DistanceFromHome  -0.001186   0.005873  -0.202  0.84000
## TotalWorkingYears -0.056146   0.009969  -5.632 1.78e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3024  on 3469  degrees of freedom
## Residual deviance: 2901  on 3466  degrees of freedom
## AIC: 2909
##
## Number of Fisher Scoring iterations: 5
```

Summary of Logistic Regression Model

The summary function in R has outputted a number of things.

- A **formula** that shows modelling Attrition as a function of Age, Commute Distance, and Total working years.
- **Residuals** that show the difference between what the model predicted and the actual value of **y**
- **Coefficients**
- The **Estimates** where the intercept tells us the value when all other features are 0. For the other features, the estimates give us the expected change in the response due to a unit change in the feature.
- **Standard Error** which allows us to construct marginal confidence intervals for the estimate of that particular feature.
- **z-value** is the regression coefficient divided by standard error. If the z-value is too big in magnitude, it indicates that the corresponding true regression coefficient is not 0 and the corresponding X-variable matters.
- **p-value** The Pr(>|z|) column represents the p-value associated with the value in the z value column. If the p-value is less than a certain significance level then this indicates that the predictor variable has a statistically significant relationship with the response variable in the model.
- The **null deviance** tells us how well the response variable can be predicted by a model with only an intercept term.
- The **residual deviance** tells us how well the response variable can be predicted by a model with p predictor variables. The lower the value, the better the model is able to predict the value of the response variable.
- The Akaike information criterion **(AIC)** is a metric that is used to compare the fit of several regression models.

Next, we will build a Naive Bayes Model and output the model learned

```
library(e1071)
nb <- naiveBayes(Attrition ~ Age + DistanceFromHome + TotalWorkingYears, data = train, family = 'binomia
nb
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace, family = "binomial")
##
## A-priori probabilities:
## Y
##         0         1
```

```
## 0.8423631 0.1576369
##
## Conditional probabilities:
##    Age
## Y       [,1]     [,2]
##   0 37.58878 8.933010
##   1 33.65631 9.564451
##
##    DistanceFromHome
## Y       [,1]     [,2]
##   0 9.245980 8.163334
##   1 9.173675 7.824132
##
##    TotalWorkingYears
## Y        [,1]     [,2]
##   0 11.858365 7.793962
##   1  8.325411 6.835063
```

Summary of Naive Bayes ouput

- Probability of attrition is **15.76%** and Probability of of retaining employee is **0.842**
- For the other quantative predictors (Age, Commute distance, and Total years worked), we get the conditional probability and the means and STDEV of the predictors

Next, we use the models to make predictions on the tesa data

```
lrProb <- predict(lm,newdata=test,type="response")
pred1 <- ifelse(lrProb>0.5,1,0)
acc <- mean(pred1==test$Attrition)
print(paste("accuracy of the logistic regression model is: ", acc))
```

```
## [1] "accuracy of the logistic regression model is:  0.826754385964912"
```

```
table(lrProb,test$Attrition)
```

```
##
## lrProb               0 1
##   0.0196747474749168 0 2
##   0.0201842117310576 0 3
##   0.0237989750974774 2 0
##   0.0251328988186297 2 0
##   0.0255133282383638 1 0
##   0.0262800905887119 2 0
##   0.0268451697606327 1 0
##   0.027313722854317  1 0
##   0.0305605077060366 1 0
##   0.0314361859260527 2 0
##   0.0335374301846455 1 0
##   0.0338056134735913 0 1
##   0.035331152999618  1 0
```

```
##   0.0362871491145278 1 0
##   0.0364897477185058 2 0
##   0.0365817881766101 1 0
##   0.037077860306109  3 0
##   0.0374180057462336 2 0
##   0.0380191854326714 1 0
##   0.0387451443724018 1 0
##   0.0388687444025147 1 0
##   0.0423781609279834 1 0
##   0.0424782820764203 1 0
##   0.0431019802018461 1 0
##   0.043866899487563  1 0
##   0.0445307458617559 1 0
##   0.0454265615115556 1 0
##   0.046040927990176  2 0
##   0.0468688842429162 3 0
##   0.0481411705941239 1 0
##   0.0481480622080093 2 0
##   0.050435681460544  2 0
##   0.0504411771313706 3 0
##   0.0506633008998626 1 0
##   0.0507774749681813 2 0
##   0.0510431724041132 1 0
##   0.0511744884123255 1 0
##   0.0522069327253742 2 0
##   0.0530128570139783 1 0
##   0.053423463115573  2 0
##   0.0537316948150905 1 0
##   0.0549991761214667 2 0
##   0.055055993460077  1 0
##   0.0570880398960325 0 1
##   0.0571499760457419 2 0
##   0.05773587377907   1 0
##   0.058307738602549  3 1
##   0.0588829387119669 1 0
##   0.0589486774267685 2 0
##   0.0590815695017177 1 0
##   0.0593425670613802 1 0
##   0.0600026378343849 2 0
##   0.0608607266456526 2 0
##   0.0624164200839295 2 0
##   0.0626911680565773 2 0
##   0.0629560693450957 2 0
##   0.0635179567732999 1 0
##   0.0640789377351073 1 0
##   0.0644151186026537 0 1
##   0.0652870863777708 1 0
##   0.0654354534114316 2 0
##   0.0654952813012843 1 0
##   0.0661456947874021 2 0
##   0.0679636457069567 2 0
##   0.0680328918965789 1 0
##   0.0696994891692144 2 0
##   0.069764335137101  0 1
```

```
##   0.0703824227533001 2 0
##   0.0706933612095959 2 0
##   0.0716964943849785 2 0
##   0.0721613482781574 1 0
##   0.0725303888678388 2 0
##   0.0738232892852166 2 0
##   0.0766783910108642 0 1
##   0.0771381996542579 2 0
##   0.077335271898682  0 1
##   0.0774173711780877 2 0
##   0.0776635874178791 1 0
##   0.0784314221975067 1 0
##   0.0790157078481954 2 0
##   0.0798513014813555 1 0
##   0.0798760225057711 1 0
##   0.0802879708789941 2 0
##   0.0808148504524251 2 0
##   0.0809911766382788 2 0
##   0.0827482830099489 2 0
##   0.0829028664654625 1 0
##   0.0832641639532484 3 0
##   0.0835547116337771 2 0
##   0.0838549951607379 2 0
##   0.0841010684649598 2 0
##   0.0842883353921935 1 0
##   0.084292763749085  0 1
##   0.0845400031458909 2 0
##   0.0846406999754533 0 1
##   0.0851845555189314 3 0
##   0.0852769974236928 1 0
##   0.0854621541704739 0 1
##   0.0857069169222676 3 0
##   0.086368580872211  1 0
##   0.0875834636850922 2 0
##   0.087612087908699  2 0
##   0.0879557414703801 3 0
##   0.0879632030879228 1 0
##   0.0880359650988629 3 0
##   0.0880583701145092 2 0
##   0.0883657853294664 2 0
##   0.0899708009104114 1 0
##   0.0901205136980572 2 0
##   0.0910041010333169 1 0
##   0.0910974788611291 2 0
##   0.0911099312481073 2 0
##   0.091490926382021  1 0
##   0.0915895267602934 2 0
##   0.0927615471248947 2 0
##   0.0942514128712332 1 0
##   0.0942563107911188 2 0
##   0.094760584533843  0 2
##   0.0948315939238967 2 0
##   0.0951811469131924 2 0
##   0.095333673237987  3 0
```

```
##   0.0965597354908345 2 0
##   0.0967749279929711 1 0
##   0.0994916913361021 1 0
##   0.0999038838902488 3 0
##   0.100322772153195  1 0
##   0.100828616028573  2 0
##   0.101245724697699  1 0
##   0.10307744206866   1 0
##   0.103082746447114  1 0
##   0.104062836441797  3 0
##   0.105133068091606  1 0
##   0.105611865385404  1 0
##   0.105934078790532  0 2
##   0.106620678355096  0 2
##   0.106794845446299  1 0
##   0.106826939459254  2 0
##   0.106940122091024  2 0
##   0.108585422873048  1 0
##   0.109738417466346  1 0
##   0.110581465977017  2 0
##   0.111253346468342  1 0
##   0.111488029843231  3 0
##   0.111846565720916  1 0
##   0.112043159708934  1 0
##   0.112399699418009  1 0
##   0.112779407208402  2 0
##   0.113124297908737  1 0
##   0.113243306631851  0 2
##   0.113508327470137  1 0
##   0.113901123188762  1 0
##   0.114309889829685  1 0
##   0.114775649197953  2 0
##   0.115275946510998  0 1
##   0.116194630680995  1 0
##   0.116210075043936  2 0
##   0.116560408803367  1 0
##   0.117071140660661  2 0
##   0.117172260155532  0 2
##   0.11753105928264   2 0
##   0.118989277667975  2 0
##   0.119129407822048  1 0
##   0.119487352517766  2 0
##   0.119612152464823  1 0
##   0.120626783117093  1 0
##   0.121335336538084  0 1
##   0.122049803191588  1 0
##   0.122095824412995  2 0
##   0.12272268153675   1 0
##   0.123000620413018  2 0
##   0.123106147365973  3 0
##   0.123224148708451  3 0
##   0.123454234405665  1 0
##   0.123625252681192  2 0
##   0.12407697101352   1 0
```

```
##   0.125134274650282  1 0
##   0.125261752064667  1 0
##   0.125268038608796  1 0
##   0.125521811983125  0 1
##   0.125674841971857  0 1
##   0.126529495157316  1 0
##   0.126693866351723  1 0
##   0.127148802860385  1 0
##   0.127303527638044  1 0
##   0.127868866119817  1 0
##   0.12794682336365   1 0
##   0.128647448931942  1 0
##   0.129789365172911  1 0
##   0.129954883682557  1 0
##   0.130498570478574  1 0
##   0.13207635520657   2 0
##   0.132501919204606  1 0
##   0.133174098145174  2 0
##   0.133609295929366  0 1
##   0.133962573521006  1 0
##   0.134222974612165  3 0
##   0.134233783572529  1 0
##   0.134676059493154  0 2
##   0.134935094390048  2 0
##   0.135300584929446  2 0
##   0.136267784763276  0 1
##   0.136414154731841  1 0
##   0.136418353204882  1 0
##   0.136682774525267  1 0
##   0.136693750297875  1 0
##   0.136855701553236  2 0
##   0.137118264448656  3 0
##   0.137283212780487  2 0
##   0.137687177677184  2 0
##   0.137770175219559  2 0
##   0.138052105798     1 0
##   0.138522894763534  3 0
##   0.138529741448388  1 0
##   0.138646485167519  2 0
##   0.138765908853775  0 2
##   0.139622567221482  0 2
##   0.140218311642768  1 0
##   0.140345790458682  3 0
##   0.140629505408799  3 0
##   0.140743419560582  2 0
##   0.141048657998346  3 0
##   0.14118108885421   2 0
##   0.141938047810443  1 0
##   0.142595056857223  0 1
##   0.143478466689553  2 0
##   0.14375868149384   1 0
##   0.144043745540586  2 0
##   0.145352872129854  2 0
##   0.1461923949189    1 0
```

```
##   0.146660216975809  3 0
##   0.146950047139399  2 0
##   0.146957239294138  2 0
##   0.147098738902021  2 0
##   0.147361432933658  3 0
##   0.14770197359828   1 0
##   0.14783234823335   2 0
##   0.14886919253586   2 0
##   0.149150821033275  1 0
##   0.149459297171076  2 0
##   0.14976099342633   1 0
##   0.150044011604822  1 0
##   0.150517429229141  3 0
##   0.150669093102677  1 0
##   0.150695681263852  0 2
##   0.150808970711731  0 2
##   0.150948955100389  1 0
##   0.151062401086912  2 0
##   0.151253123616251  2 0
##   0.152129852685211  1 0
##   0.152896101853734  2 0
##   0.152935061110595  3 0
##   0.153037674429586  0 2
##   0.153179354388906  2 0
##   0.153222996666376  3 0
##   0.15349465931936   3 0
##   0.153530918872615  6 0
##   0.153685069847293  1 0
##   0.153692532244178  2 0
##   0.153815133913765  0 1
##   0.154084822464671  0 1
##   0.154246912651362  2 0
##   0.154600479988441  2 0
##   0.15489848796235   1 0
##   0.154910664006942  2 0
##   0.155189442873323  1 0
##   0.155344953466741  1 0
##   0.15598806572227   1 0
##   0.156124414596986  1 0
##   0.156288253935792  2 0
##   0.15659361946936   1 0
##   0.156750276068249  3 0
##   0.156757859732676  0 2
##   0.15747849022829   1 0
##   0.157648217682527  0 1
##   0.15812114399254   1 0
##   0.158141163994468  0 1
##   0.158753504009064  2 0
##   0.159050337228033  2 0
##   0.159070452647775  1 0
##   0.159078127387513  0 2
##   0.159229119279451  3 0
##   0.159249253028328  3 0
##   0.159387914177988  1 0
```

```
##   0.159526673153788  3 0
##   0.159668446135323  2 0
##   0.159705888950929  1 0
##   0.159892974360235  1 0
##   0.160482811599162  2 0
##   0.160764902089706  1 0
##   0.160978113875381  0 1
##   0.162061405914024  2 0
##   0.162383683935875  1 0
##   0.162868069403914  1 0
##   0.163166237448395  1 0
##   0.163340906591729  0 1
##   0.164132058517425  1 0
##   0.164248603370101  1 0
##   0.164778847352447  1 0
##   0.16493418891128   1 0
##   0.16508473577709   1 0
##   0.165097556103462  2 0
##   0.165398999309795  2 0
##   0.166041844249247  2 0
##   0.166218980279953  1 0
##   0.166560794104391  0 1
##   0.166911147770889  2 0
##   0.167034230113833  1 0
##   0.167249135754628  1 0
##   0.167385379629144  1 0
##   0.167550687580579  1 0
##   0.167868701580754  1 0
##   0.168000366849722  3 0
##   0.168179170910093  1 0
##   0.168511168236356  2 0
##   0.168664320611994  2 0
##   0.168685408035376  1 0
##   0.168796483328857  1 0
##   0.168843687917586  0 2
##   0.168864793211355  1 0
##   0.169510295529505  1 0
##   0.169831270095193  0 3
##   0.169977282380936  2 0
##   0.170019720507145  1 0
##   0.170312102537543  1 0
##   0.170466556138763  0 1
##   0.170647446626224  2 2
##   0.170815315234449  2 2
##   0.171427097783931  2 0
##   0.171474886269313  0 1
##   0.171643400384186  2 0
##   0.172002238719404  2 0
##   0.172657242806201  1 1
##   0.172826676486615  1 0
##   0.172982935068205  2 0
##   0.17334398781331   0 1
##   0.17351395407885   1 0
##   0.174650635934888  2 0
```

```
##   0.17469399528072   1 0
##   0.174773088002811  0 2
##   0.174794778595524  2 0
##   0.174821612441555  2 0
##   0.174865005241385  2 0
##   0.175357078982576  0 2
##   0.176008652929463  6 0
##   0.176022142744353  3 0
##   0.176194175277839  0 1
##   0.176511595791564  0 1
##   0.176856544663856  1 0
##   0.176891980573178  1 0
##   0.177007309089296  2 0
##   0.177388532022606  2 0
##   0.177427223561134  1 0
##   0.177729616418956  0 1
##   0.178219607229463  2 0
##   0.178374467103756  1 0
##   0.178393322447564  1 0
##   0.178923691891661  2 0
##   0.179089508929134  2 0
##   0.179599336467536  0 2
##   0.179613041494861  1 0
##   0.179774102339737  1 0
##   0.179949000972439  1 0
##   0.180636156947019  0 2
##   0.181339164329274  2 0
##   0.181817767059445  2 0
##   0.181867816348581  3 0
##   0.182207056602225  2 0
##   0.18233363625352   1 0
##   0.183078173714349  2 0
##   0.183083503905279  4 0
##   0.183929945058487  2 0
##   0.184135923849978  0 2
##   0.184478440856496  3 0
##   0.184523694600022  1 0
##   0.18555111381786   1 0
##   0.185895663459002  1 0
##   0.186240706591775  1 0
##   0.186420466587438  2 0
##   0.186757546454335  4 0
##   0.187826230098312  2 0
##   0.187992968126115  2 0
##   0.188035566859998  2 0
##   0.188406656905504  1 0
##   0.189459483774551  1 0
##   0.1901138966146    2 0
##   0.190554352177054  2 0
##   0.190685378583075  2 0
##   0.190722943309864  3 0
##   0.191456034105527  3 0
##   0.193222267885448  1 0
##   0.193606717261302  6 0
```

```
##   0.193986175281799  1 0
##   0.194000718545037  1 0
##   0.194357214026203  2 0
##   0.195100906334151  1 0
##   0.195473560145022  2 0
##   0.195607151634909  0 1
##   0.195846752554252  1 0
##   0.196009850640587  2 0
##   0.196033550769907  1 0
##   0.196901649154899  0 1
##   0.197479466964607  1 0
##   0.199744039660749  3 0
##   0.200054627217905  0 1
##   0.200410287181894  0 2
##   0.200512411492755  2 0
##   0.201234646003388  2 0
##   0.201249597004522  2 2
##   0.202356297101864  0 1
##   0.202660646846837  2 0
##   0.202852304710405  1 0
##   0.203122880248693  2 0
##   0.203452459020815  1 0
##   0.203491906454017  1 0
##   0.203506982433702  3 0
##   0.2048008049834    1 0
##   0.205211789666523  1 0
##   0.205395874758265  2 0
##   0.20577734059933   2 0
##   0.206713841650642  0 1
##   0.206933016687407  2 0
##   0.207502072764531  0 1
##   0.208078155679346  1 0
##   0.208674364100132  1 0
##   0.209418296742066  0 2
##   0.209755401469602  1 0
##   0.210401492131421  2 0
##   0.210755228512826  1 0
##   0.21119048330189   1 0
##   0.211529677472219  0 1
##   0.213330519262359  2 0
##   0.214311440805398  2 0
##   0.214910982051954  1 0
##   0.215054306668833  1 0
##   0.21511109962147   1 0
##   0.215264212677778  2 0
##   0.216685332562026  1 0
##   0.217264102701172  2 0
##   0.217837916490317  0 1
##   0.217869613570381  0 3
##   0.218258093176781  1 0
##   0.218881501644021  1 0
##   0.219084284691614  2 0
##   0.219094100553091  1 0
##   0.220262015145183  1 0
```

```
##   0.220669552301646  2 0
##   0.22067941911813   2 0
##   0.221061613342682  2 0
##   0.221486247286938  1 0
##   0.221674710475701  1 0
##   0.222289022741805  1 0
##   0.222315021740333  1 0
##   0.222699239188075  5 0
##   0.222904549859355  1 0
##   0.223119940176712  0 2
##   0.223462910316602  2 0
##   0.224286968210966  0 1
##   0.225468347890138  2 0
##   0.225494610925547  2 0
##   0.226755777126695  0 1
##   0.226808517471744  2 0
##   0.22717182879118   2 1
##   0.228214314439611  2 0
##   0.228632250954762  2 0
##   0.228841421057593  2 0
##   0.229270302795349  1 0
##   0.230056150834843  2 0
##   0.230459957274377  1 0
##   0.231302132263297  2 0
##   0.232156683536121  1 0
##   0.232173264159375  2 0
##   0.232579664140163  0 1
##   0.232781109153781  2 0
##   0.233171618890053  2 0
##   0.233188249986307  1 0
##   0.233781281828496  1 0
##   0.235085044200285  1 0
##   0.235287993895991  4 0
##   0.235501394128403  1 0
##   0.236152747077709  0 2
##   0.236553605240484  2 2
##   0.236982124398137  2 0
##   0.237834306871643  2 0
##   0.237851167382115  2 0
##   0.238479679890645  2 0
##   0.239874024342294  1 0
##   0.240134687685857  1 0
##   0.240151659890044  1 0
##   0.240179102898477  0 3
##   0.240368084786842  2 0
##   0.242087213600572  1 0
##   0.242740451566874  2 0
##   0.243842418647563  2 0
##   0.24447108714127   0 1
##   0.247929280450784  1 0
##   0.248399777908062  2 0
##   0.249258131602868  2 0
##   0.249656540140678  1 0
##   0.249896148352843  1 0
```

```
##   0.25033422379851   1 0
##   0.25055679930707   1 0
##   0.251002345368612  0 2
##   0.251179533182605  1 0
##   0.252789790965282  0 1
##   0.252950237976793  0 1
##   0.2551863920434    2 0
##   0.256099785878487  1 0
##   0.256778024576941  1 0
##   0.260122061339847  2 0
##   0.261646493942763  2 0
##   0.262316243263745  1 0
##   0.263494208083458  0 2
##   0.265801606627573  1 0
##   0.266033054145144  0 1
##   0.266264630108071  1 0
##   0.266698755394665  1 0
##   0.267590843869931  0 1
##   0.268754309237199  2 0
##   0.269453922216684  0 1
##   0.270856591685557  0 1
##   0.273844503836662  2 3
##   0.274249340793767  2 0
##   0.275498010978276  1 0
##   0.276178484617274  1 0
##   0.277754539883319  1 0
##   0.278517566472303  0 1
##   0.279471580790912  0 1
##   0.280188402746882  0 1
##   0.280636546871103  1 0
##   0.283170978215161  0 1
##   0.286844722732407  1 0
##   0.28729926393099   1 0
##   0.287959156042627  0 2
##   0.288252256915392  2 0
##   0.288495571839058  0 2
##   0.292600153962958  1 0
##   0.293267211519331  1 0
##   0.294231827983905  0 2
##   0.294301690921543  0 5
##   0.295730642583659  0 2
##   0.297462166553735  2 0
##   0.297690557494091  2 0
##   0.302100777854188  0 1
##   0.304355338769774  0 5
##   0.305289045536217  1 0
##   0.308567856621793  0 3
##   0.309327269226804  0 2
##   0.310777122914331  0 1
##   0.314088227563271  2 0
##   0.314343719113155  1 0
##   0.316115196041537  1 0
##   0.325527027319278  0 2
##   0.331468154572941  0 2
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

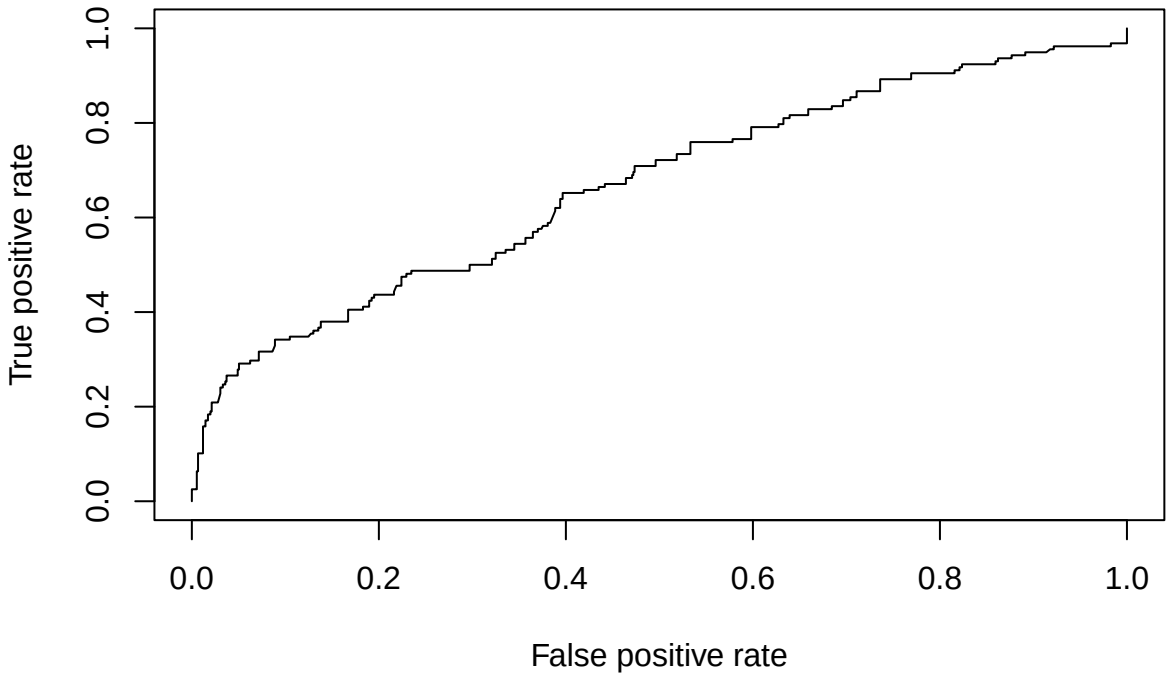
```
## Loading required package: lattice
```

```
library(ROCR)
confusionMatrix(as.factor(pred1), reference = as.factor(test$Attrition))
```

```
## Warning in confusionMatrix.default(as.factor(pred1), reference =
## as.factor(test$Attrition)): Levels are not in the same order for reference and
## data. Refactoring data to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 754 158
##          1   0   0
##
##                Accuracy : 0.8268
##                  95% CI : (0.8006, 0.8508)
##     No Information Rate : 0.8268
##     P-Value [Acc > NIR] : 0.5212
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.8268
##          Neg Pred Value :    NaN
##              Prevalence : 0.8268
##          Detection Rate : 0.8268
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

```
pr <- prediction(lrProb,test$Attrition)
prf <- performance(pr,measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
nbProb <- predict(nb,newdata=test,type="class")
pred2 <- ifelse(nbProb>0.5,1,0)
```

```
## Warning in Ops.factor(nbProb, 0.5): '>' not meaningful for factors
```

```
acc2 = mean(nbProb==test$Attrition)
print(paste("accuracy of the Naive Bayes model is: ", acc2))
```

```
## [1] "accuracy of the Naive Bayes model is:  0.826754385964912"
```

```
table(nbProb,test$Attrition)
```

```
##
## nbProb   0   1
##      0 754 158
##      1   0   0
```

Analysis of models

Advantages of Naive Bayes

- Simple to Implenebt. Conditional probabilities are easy to evaluate

- Very Fast
- If conditional indepenedence assumption holds, model will yield great results

disadvantages of Naive Bayes

- Conditional independance assumption does not always hold
- Zero Probability Problem: Canencounter words in the test data for a particular class that are not present in the training data, we might end up with zero class probabilities.

Advantages of Logistic Regression

- Logistic Regression is one of the simplest machine learning algorithms
- This algorithm allows models to be updated easily to reflect new data
- Logistic Regression outputs well-calibrated probabilities along with classification results. This is an advantage over models that only give the final classification as results.

disdvantages of Logistic Regression

- Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set
- Non linear problems can't be solved with logistic regression since it has a linear decision surface.

Ultimately, the two models produced similar results. I would say this was the case because the data was complete and rather easy and straightforward to interpret, the target variable was binary as well.

Benefits of metrics

Confusion matrix:

- Gives information about errors made by the classifer and the types of errors
- Reflects how a classification model is disorganizerd

ROC Curves

- Allows us to see the tru positive rate against the false positive rate at various thresholds