EMIS 5357/7357 Spring 2019 Homework 2
Due Sunday, February 24, 2019 by 11:59pm
9 points

Assignment done by :

1. **Apoorva Jain(47527939)**

The data in hw2.csv represents fictitious data provided by a real consulting company about its workers' attrition (attrition means that some of the workers quit). **Attrition = 1 when "Yes"** (the worker quit.) The other columns headers are self-explanatory. The goal is to visualize the data (Part 1) and then build, refine and test a logistic regression to predict attrition as a function of all the others (Part 2).

For the plots, I prefer if you use ggplot2 in R. We will see how to use ggplot2 during the week of February 18. You can do Part 1 Q1, Q2 and Q3 and do all of Part 2 without having done the plots, and do the plots later.

Grading number: if two numbers, first one is for EMIS 5357, second one is for EMIS 7357. If one number, same number of points for both EMIS 5357 and 7357.

Part 1 (4 points):

1. (0.75pt/0.5pt) Tables:
   a. Compute a table showing the average attrition rate depending on
      i. business travel (hint: we are going to use tapply on MyData$Attrition, grouping by MyData$BusinessTravel and applying the function mean)

Answer:

Subcategory	Non-Travel	Travel Frequently	Travel Rarely
Mean	0.08	0.149568552253116	0.249097472924188

The above table computes the average value of attrition rate depending upon business travel.

ii. the department,

Answer:

Subcategory	Human Recourses	Research and Development	Sales
Mean	0.138397502601457	0.19047619047619	0.20627802690583

The above table computes the average value of attrition rate depending upon the department.

iii. education field,

Answer:

Subcategory	Human Recourses	Life Science	Marketing	Medical	Other	Technical
Mean	0.1341463	0.13577586	0.1468646	0.22012	0.2424	0.2592592

The above table computes the average value of attrition rate depending upon the education field.

iv. gender,

Answer:

Subcategory	Female	Male
Mean	0.147959183673469	0.170068027210884

The above table computes the average value of attrition rate depending upon the gender.

v. job role,

Answer:

Subcategory	Healthcare Re presentative	Human Resources	Laboratory Technical	Manager	Manufacturing Director
Mean	0.025	0.0490196078	0.0687022900	0.068965517	0.16095890410

Subcategory	Research Director	Research Scientist	Sales Executive	Sales Representative
Mean	0.1748466257	0.2307692307	0.23938223938	0.3975903614

The above table computes the average value of attrition rate depending upon the gender.

vi. marital status

Answer:

Subcategory	Divorced	Married	Single
Mean	0.100917431192661	0.12481426448737	0.25531914893617

Average Attrition rate for martial category is high for single as compared to Divorced and married sub category.

vii. and overtime.

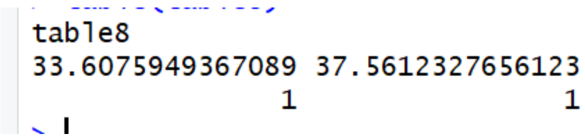
Answer:

Subcategory	No	Yes
Mean	0.104364326375712	0.305288461538462

Average Attrition rate for overtime category is high for those who do overtime as compared to those who don't do overtime.

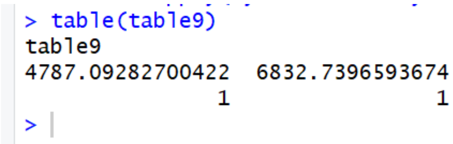
b.For each of those factors, comment on when the attrition rate is much higher from some values of the factors than others.

2. (0.75pt/0.5pt) More tables:
   a. Compute the average Age depending on whether Attrition = 0 or 1. (Hint: we use tapply again but now MyData$Attrition will be the second argument).



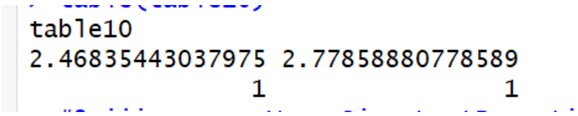
```
table8
33.6075949367089 37.5612327656123
               1                1
```

   b. Repeat for, instead of Age:
      i. Monthly Income,



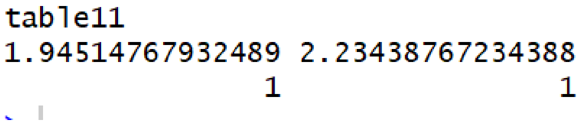
```
> table(table9)
table9
4787.09282700422  6832.7396593674
               1                1
>
```

      ii. JobSatisfaction,



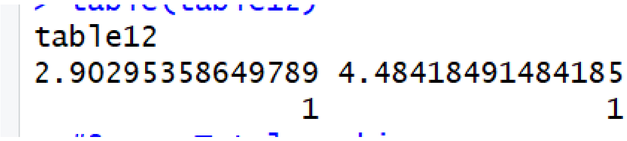
```
table10
2.46835443037975 2.77858880778589
               1                1
```

      iii. YearsSinceLastPromotion,



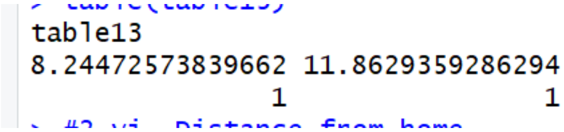
```
table11
1.94514767932489 2.23438767234388
               1                1
```

      iv. YearsInCurrentRole,



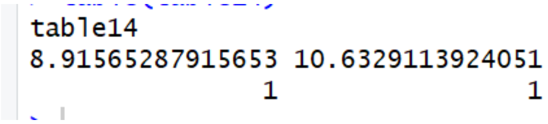
```
table12
2.90295358649789 4.48418491484185
               1                1
```

v. TotalWorkingYears,



```
> table(table13)
table13
8.24472573839662 11.8629359286294
               1                1
> #2 vi. Distance from home
```

vi. and DistanceFromHome.



```
> table(table14)
table14
8.91565287915653 10.6329113924051
               1                1
> |
```

c. Which of those categorical independent variables, if any, seem to have predictive power for attrition?

Answer:

I really think Monthly income and age are the two categorical variables which have high predictive power for attrition.

3. (0.5pt/0.5pt) Compute the correlation of the numerical independent variables. NOTE: Because the function "cor" works only for numerical variables, we have to get rid of the factors. You could do it by hand or duplicate a csv file where you delete the columns with factors and read that into R or you can copy and paste the following code, which finds the columns that have numeric data (they will be called "nums") and subset MyData into MyDataNum to keep only those, and then we compute correlation and view it in RStudio's viewer to make it easy to read this big table (next time you want to use this code, you just have to update the name of the data frame, which here is MyData):

```
nums = unlist(lapply(MyData,is.numeric))
MyDataNum = MyData[ , nums]
mycor= cor(MyDataNum)
View(mycor)
```

What independent variables are the most positively correlated with Attrition? The most negatively? Does it make sense?

Answer:

Independent variables which are **positive correlated** with Attrition are

a) Distance from Home

	Attrition
Distance from Home	0.077923583

b) Monthly Rate

	Attrition
Monthly Rate	0.015170213

c) Num Companies worked

	Attrition
Num Companies worked	0.043493739

a) Performance Rating

	Attrition
Performance Rating	0.002888752

Independent variables which are **negative correlated** with Attrition are

a) Age

	Attrition
Age	-0.159205007

b) Daily Rate

	Attrition
Daily Rate	-0.056651992

c) Education

	Attrition
Education	-0.031372820

d) Environmental Satisfaction

	Attrition
Environmental Satisfaction	-0.103368978

e) Hourly Rate

	Attrition
Hourly Rate	-0.006845550

f) Job Involvement

	Attrition
Hourly Rate	-0.130015957

g) Job level

	Attrition

Job level	-0.169104751

h) Job satisfaction

	Attrition
Job satisfaction	-0.103481126

i) Monthly income

	Attrition
Monthly income	-0.159839582

j) Percent Salary Hike

	Attrition
Percent Salary Hike	-0.013478202

k) Relationship satisfaction

	Attrition
Relationship satisfaction	-0.045872279

l) Stock option level

	Attrition
Stock option level	-0.137144919

m) Total working Year

	Attrition
Total working Year	-0.171063246

n) Training times last year

	Attrition
Training times last year	-0.059477799

o) Work life balance

	Attrition
Work life balance	-0.063939047

p) Years at company

	Attrition
Years at company	-0.134392214

q) Years in current role

	Attrition
Years in current role	-0.160545004

r) Years since last promotion

	Attrition
Years since last promotion	-0.033018775

s) Years with current manager.

	Attrition
Years with current manager.	-0.156199316

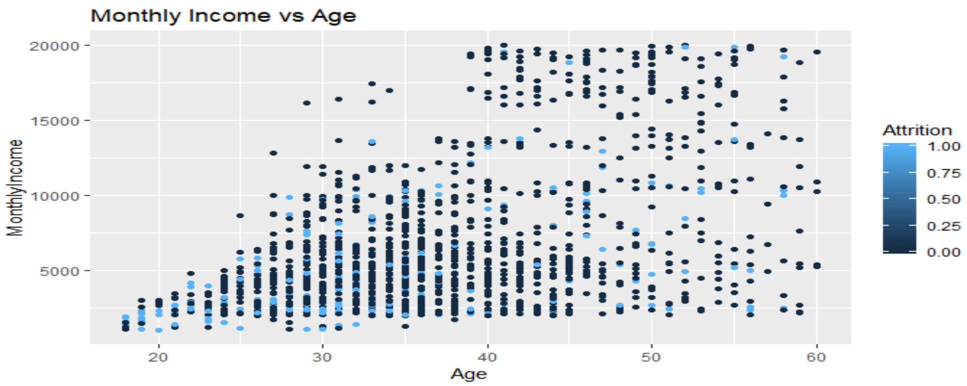
For Attrition there are less positive independent variables than negative independent variables.

Out of positive independent variable distance from home is most important as compared to monthly rate, Num Companies worked, and Performance Rating.

Out of negative independent variable Total Working Years is most important as compare to other negative independent variables.

4. (0.5pt/0.5pt) Plot, preferably using ggplot2, MonthlyIncome as a function of Age, color coded by Attrition (we want a scatterplot, so we use geom_point()). Comment if there is any pattern.

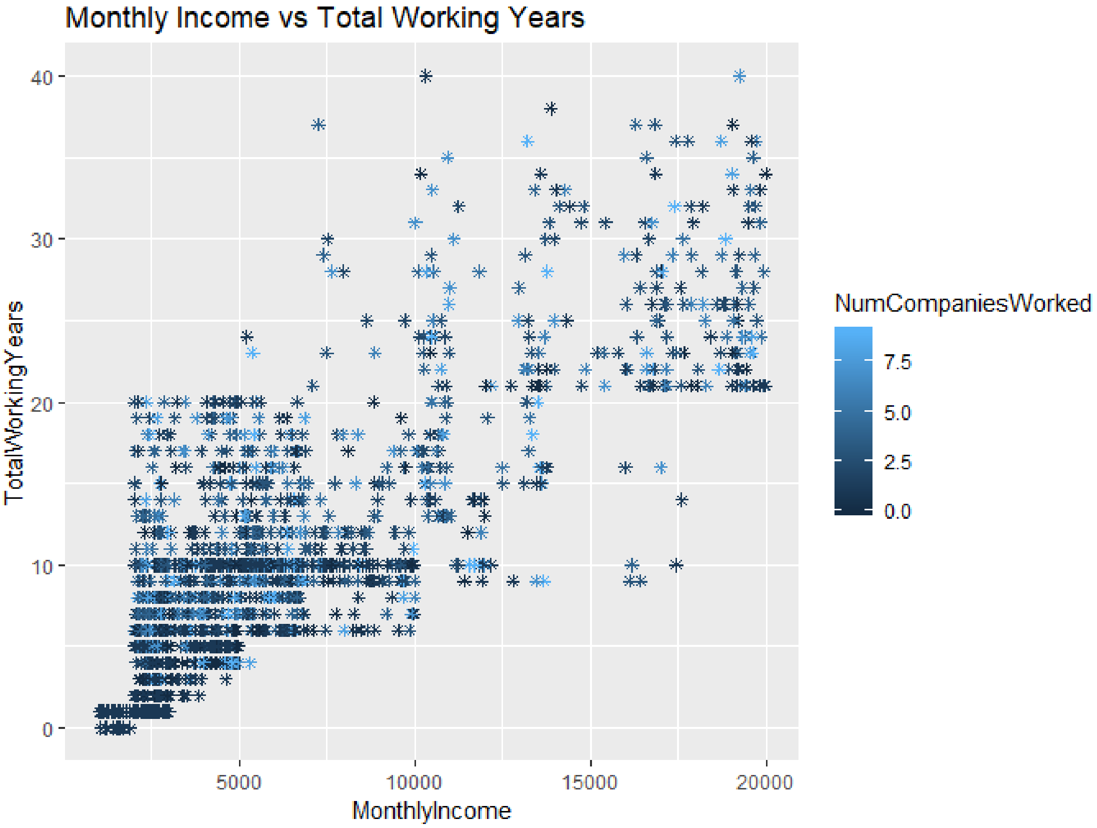
Answer:

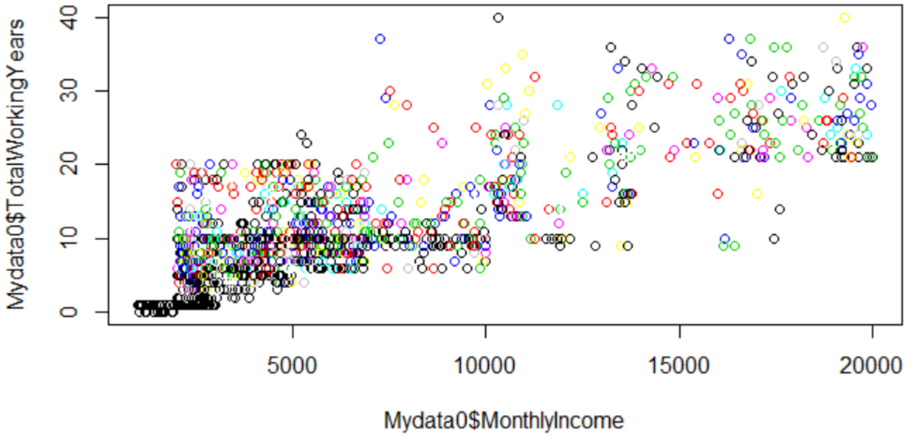


Attrition for monthly income of 5000 to 12500 , with age group of 20 to 40 attrition rate is rising and its more at the age of 40 years to 60 years that is going to be 12500 to 15000 and monthly income range from 15000 to 20000 from 40 to 60 years is moderate.

5. (0.75pt/0.5pt) Plot TotalWorkingYears as a function of MonthlyIncome color-coded by NumCompaniesWorked. Why do you think you don't have any workers with low MonthlyIncome who have worked more than 20 years?

Answer:



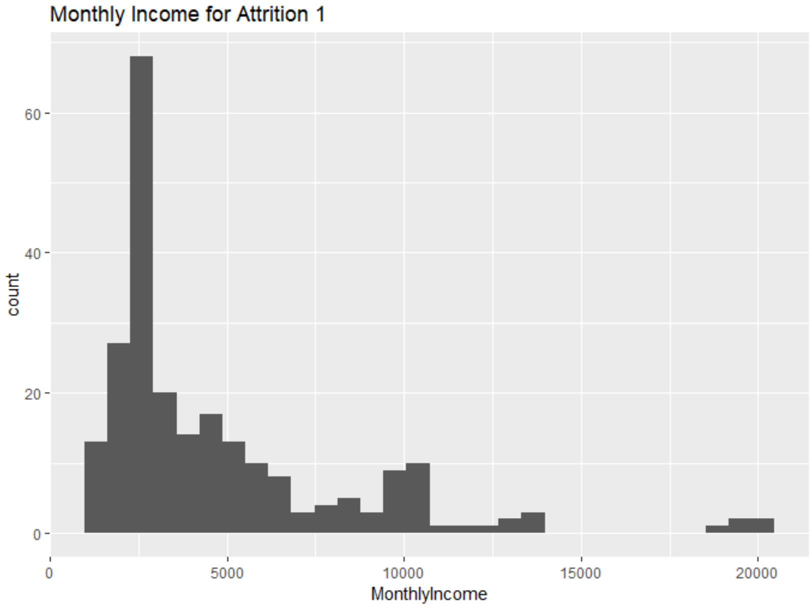


For comparison I have plotted the graph by two different ways that is shown above. I think the workers that have worked more than 20 years doesn't have any low income because they have high experience and the companies provide some privilege as token of appreciation for their hard-work and dedication.

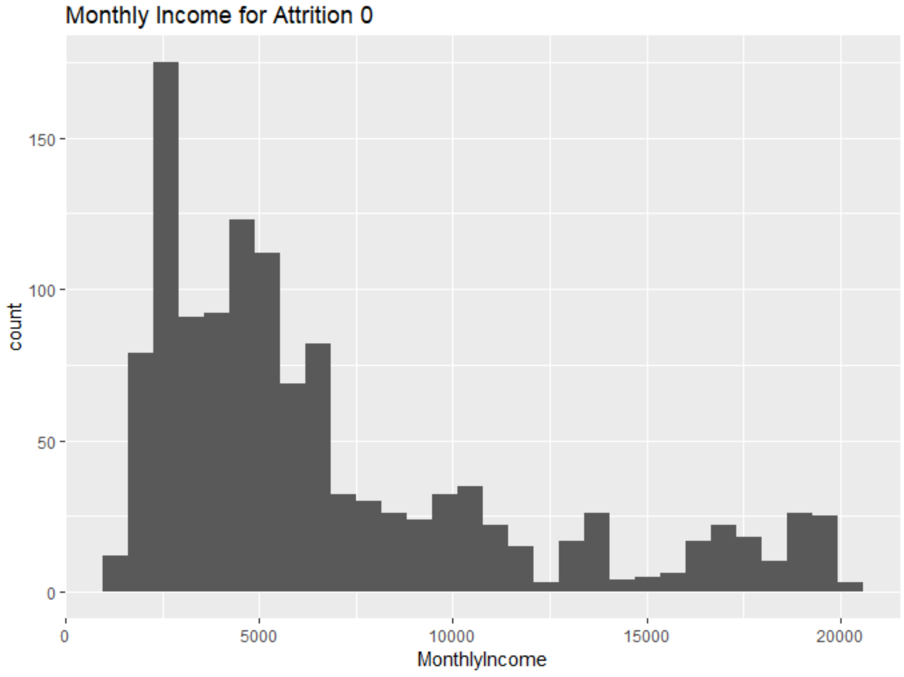
6. (0.75pt/0.5pt) Create two subsets of the data frame: one for when Attrition = 1 (called MyData1) and the other when Attrition = 0 (called MyData0). Plot a histogram of MonthlyIncome in each case (this uses geom_histogram).

Answer:

This plot shows monthly attrition equals to 1.



The plot below shows attrition equals to 0.



7. (EMIS 7357 only: 0.5pt) We notice that the y-axis of the graphs in the previous question don’t have the same scale, so it makes it hard to compare the two histograms. This is because each data frame has a different number of rows, so the count gets very different. Instead of count, we’d like to see probabilities in the vertical axis, so that they would

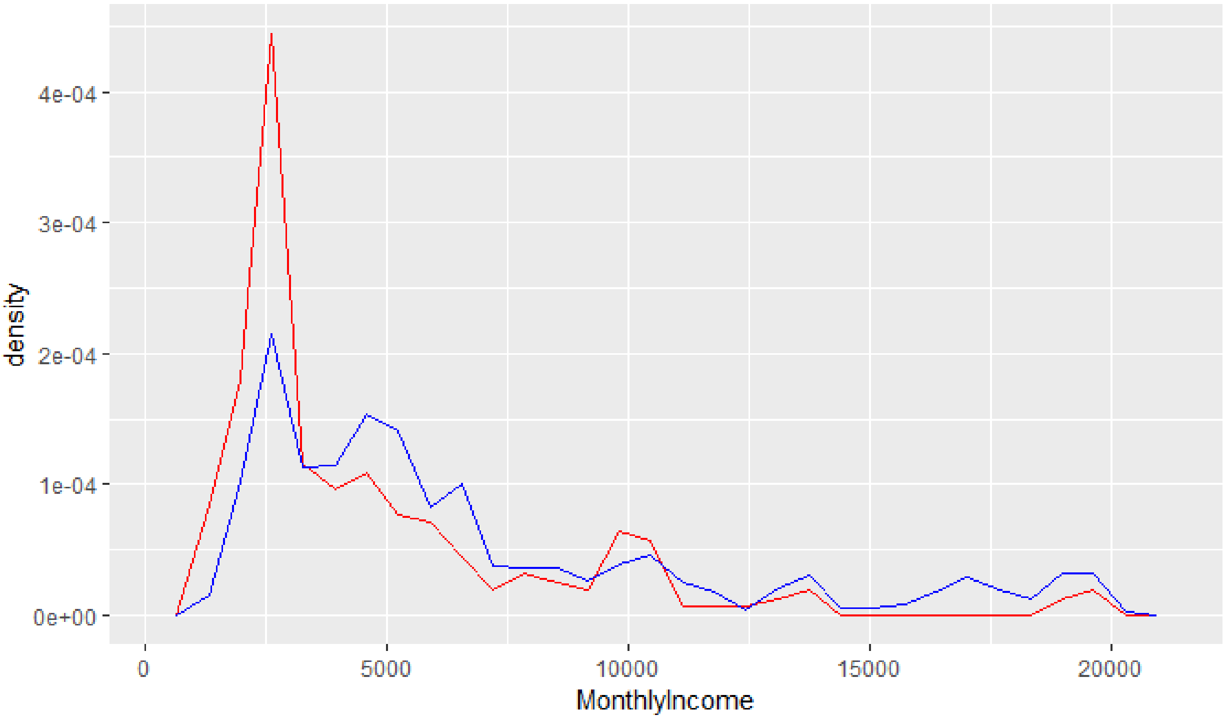
always have the same scale of [0,1]. In this case they’re called density and we’ll have to use geom_freqpoly instead of geom_histogram. Also, we’d prefer to have both plots on the same graph to compare them better. To do this, write the following code:

```
ggplot(MyData,aes(x=MonthlyIncome, y=..density..)) +
  geom_freqpoly(data=MyData1,color = "red") +
  geom_freqpoly(data=MyData0,color = "blue")
```

Comment on how MonthlyIncome could play a role in predicting attrition.

Answer:

1 Comment on how Monthly Income could play a role in predicting attrition.



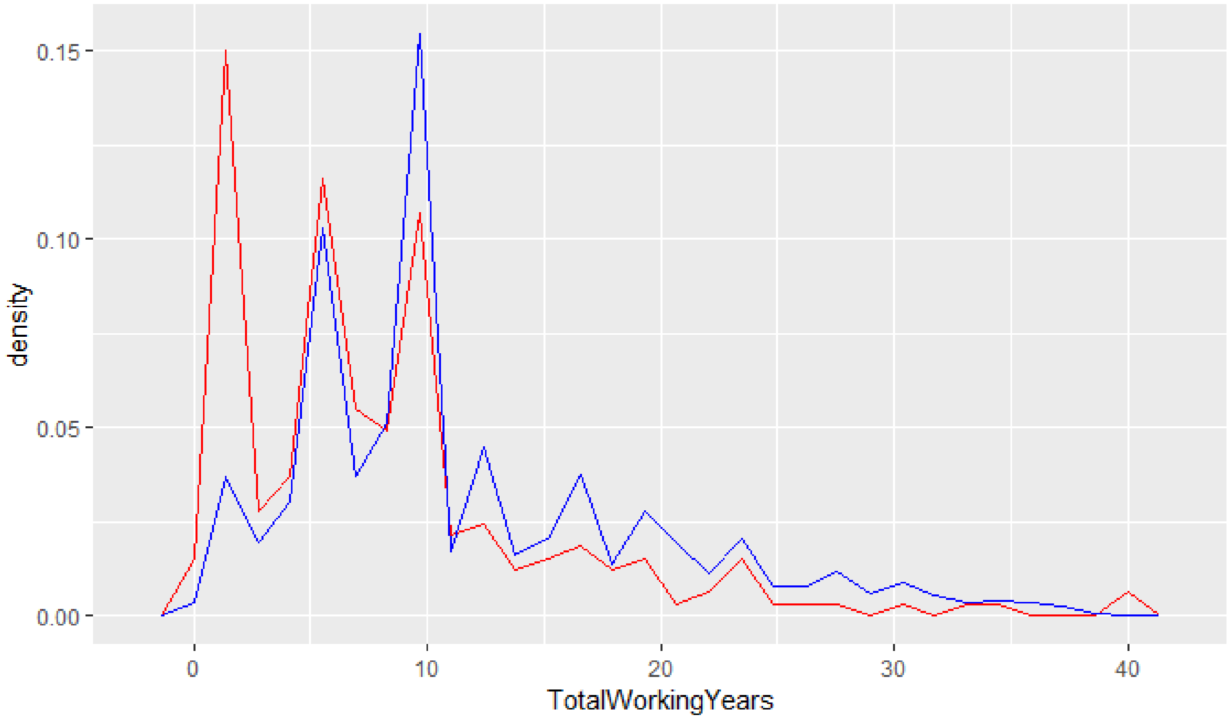
Monthly income plays vital role in predicting the income.
From Monthly Income from 0 to 5000 = Attrition is for monthly income 2500 is high
From Monthly Income from 5000 to 10000 = Attrition is decreases up to monthly income 7000 and rise, fall and again rise up to monthly income 10000.

From Monthly Income from 10000 to 15000 = Attrition rate is slightly rising upto approximately monthly income 13500 and fall for monthly income 14000 and after remaining nearly constant upto monthly income 15000.

From Monthly Income from 15000 to 20000 = Attrition rate is constant and low upto approximately 18000 monthly income and rise and fall and remains constant upto monthly income 20000.

8. (EMIS 7357 only: 0.5pt) Repeat the previous question for TotalWorkingYears instead of MonthlyIncome. Comment on how TotalWorkingYears could play a role in predicting attrition.

Answer:



Total Working Year from 0 to 10 = Attrition is high with highest density of 0.15 for first 0 to 2 year and decrease decreases with density of 0.03 from 2 to 3 and again rises from 3 to 4 with density of 0.04 and sharply increases up to density of 0.12 for years 4 to 6 and fall with different rate for year 6 to 7 & 7 to 8 and finally rises up to density 0.10 from years 8 to 10.

Total Working Year from 10 to 20: From year 10 up to 12 density is increase slightly and drop sharply and this process is repeated two times with final density 0.02 at the end of 20 years

Total Working Year from 20 to 30: For year 20 to 30 attrition rises upto 23 years and drops at different rate and finally increases at year 30

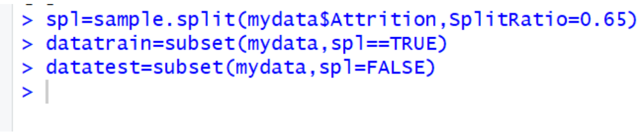
Total Working Year from 30 to 40: Attrition rate is fluctuating slightly from total working years from 30 to 40 with low density ranging from 0.01 to 0.03.

Part 2 (5 points):

1. (0.5pt) Separate the data frame into a training set (65% of data) and a testing set (35%), assigning instances randomly to either set. What is the proportion of attrition “yes” in your training set?

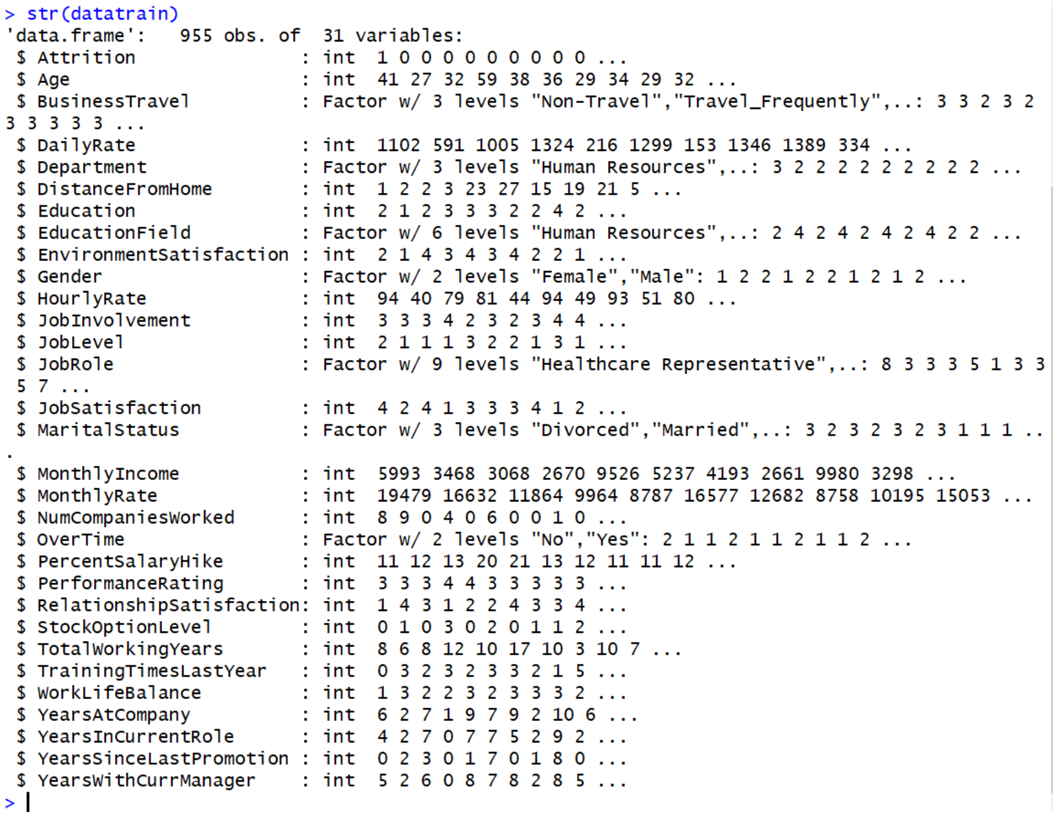
Answer.1:

Following are the commands for the splitting the data sets into training and testing sets:



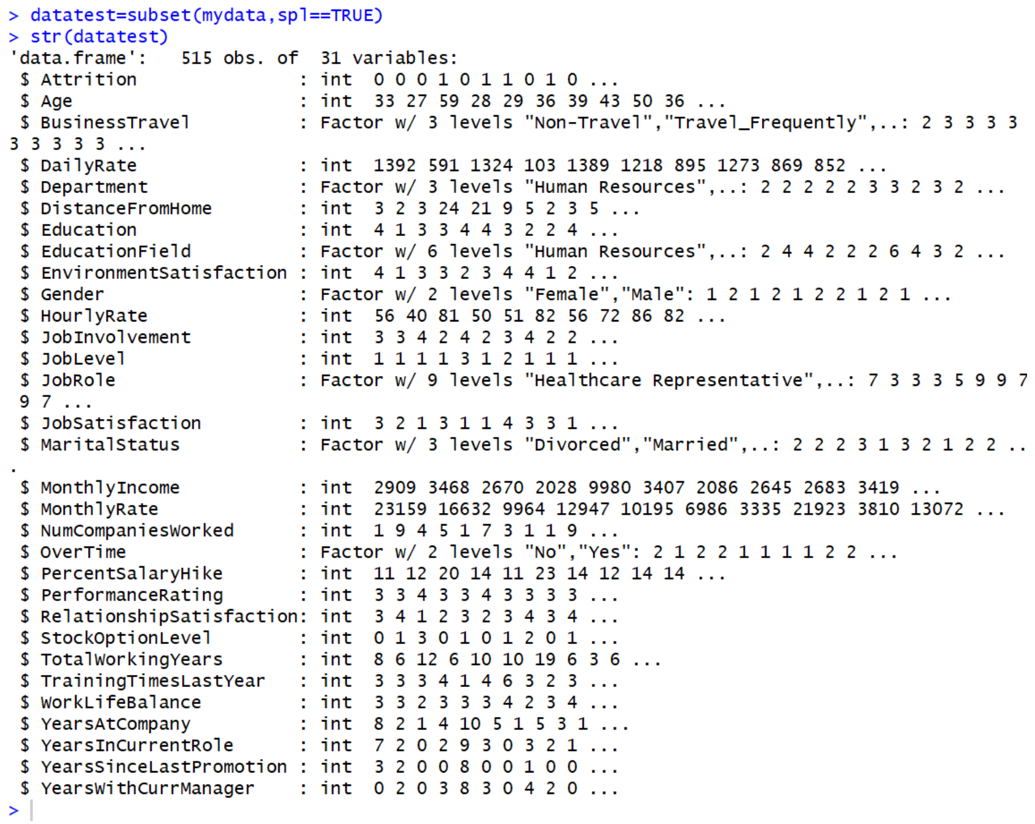
```
> spl=sample.split(mydata$Attrition,SplitRatio=0.65)
> datatrain=subset(mydata,spl==TRUE)
> datatest=subset(mydata,spl=FALSE)
> |
```

Output of training sets:



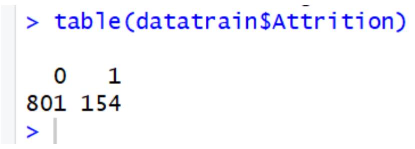
```
> str(datatrain)
'data.frame':	955 obs. of  31 variables:
 $ Attrition               : int  1 0 0 0 0 0 0 0 0 0 ...
 $ Age                     : int  41 27 32 59 38 36 29 34 29 32 ...
 $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2
3 3 3 3 3 ...
 $ DailyRate               : int  1102 591 1005 1324 216 1299 153 1346 1389 334 ...
 $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
 $ DistanceFromHome        : int  1 2 2 3 23 27 15 19 21 5 ...
 $ Education               : int  2 1 2 3 3 3 2 2 4 2 ...
 $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 2 4 2 4 2 4 2 2 ...
 $ EnvironmentSatisfaction : int  2 1 4 3 4 3 4 2 2 1 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 1 2 ...
 $ HourlyRate              : int  94 40 79 81 44 94 49 93 51 80 ...
 $ JobInvolvement          : int  3 3 3 4 2 3 2 3 4 4 ...
 $ JobLevel                : int  2 1 1 1 3 2 2 1 3 1 ...
 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 3 3 3 5 1 3 3
 5 7 ...
 $ JobSatisfaction         : int  4 2 4 1 3 3 3 4 1 2 ...
 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 3 2 3 1 1 1 ..
.
 $ MonthlyIncome           : int  5993 3468 3068 2670 9526 5237 4193 2661 9980 3298 ...
 $ MonthlyRate             : int  19479 16632 11864 9964 8787 16577 12682 8758 10195 15053 ...
 $ NumCompaniesWorked      : int  8 9 0 4 0 6 0 0 1 0 ...
 $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 1 1 2 ...
 $ PercentSalaryHike       : int  11 12 13 20 21 13 12 11 11 12 ...
 $ PerformanceRating       : int  3 3 3 4 4 3 3 3 3 3 ...
 $ RelationshipSatisfaction: int  1 4 3 1 2 2 4 3 3 4 ...
 $ StockOptionLevel        : int  0 1 0 3 0 2 0 1 1 2 ...
 $ TotalWorkingYears       : int  8 6 8 12 10 17 10 3 10 7 ...
 $ TrainingTimesLastYear   : int  0 3 2 3 2 3 3 2 1 5 ...
 $ WorkLifeBalance         : int  1 3 2 2 3 2 3 3 3 2 ...
 $ YearsAtCompany          : int  6 2 7 1 9 7 9 2 10 6 ...
 $ YearsInCurrentRole      : int  4 2 7 0 7 7 5 2 9 2 ...
 $ YearsSinceLastPromotion : int  0 2 3 0 1 7 0 1 8 0 ...
 $ YearsWithCurrManager    : int  5 2 6 0 8 7 8 2 8 5 ...
> |
```

Output of testing sets:



```
> datatest=subset(mydata,spl==TRUE)
> str(datatest)
'data.frame':	515 obs. of  31 variables:
 $ Attrition               : int  0 0 0 1 0 1 1 0 1 0 ...
 $ Age                     : int  33 27 59 28 29 36 39 43 50 36 ...
 $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 2 3 3 3 3
3 3 3 3 3 ...
 $ DailyRate               : int  1392 591 1324 103 1389 1218 895 1273 869 852 ...
 $ Department              : Factor w/ 3 levels "Human Resources",..: 2 2 2 2 2 3 3 2 3 2 ...
 $ DistanceFromHome        : int  3 2 3 24 21 9 5 2 3 5 ...
 $ Education               : int  4 1 3 3 4 4 3 2 2 4 ...
 $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 4 2 2 2 6 4 3 2 ...
 $ EnvironmentSatisfaction : int  4 1 3 3 2 3 4 4 1 2 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 2 2 1 2 1 ...
 $ HourlyRate              : int  56 40 81 50 51 82 56 72 86 82 ...
 $ JobInvolvement          : int  3 3 4 2 4 2 3 4 2 2 ...
 $ JobLevel                : int  1 1 1 1 3 1 2 1 1 1 ...
 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 7 3 3 3 5 9 9 7
 9 7 ...
 $ JobSatisfaction         : int  3 2 1 3 1 1 4 3 3 1 ...
 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 2 2 2 3 1 3 2 1 2 2 ..
.
 $ MonthlyIncome           : int  2909 3468 2670 2028 9980 3407 2086 2645 2683 3419 ...
 $ MonthlyRate             : int  23159 16632 9964 12947 10195 6986 3335 21923 3810 13072 ...
 $ NumCompaniesWorked      : int  1 9 4 5 1 7 3 1 1 9 ...
 $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 1 1 2 2 ...
 $ PercentSalaryHike       : int  11 12 20 14 11 23 14 12 14 14 ...
 $ PerformanceRating       : int  3 3 4 3 3 4 3 3 3 3 ...
 $ RelationshipSatisfaction: int  3 4 1 2 3 2 3 4 3 4 ...
 $ StockOptionLevel        : int  0 1 3 0 1 0 1 2 0 1 ...
 $ TotalWorkingYears       : int  8 6 12 6 10 10 19 6 3 6 ...
 $ TrainingTimesLastYear   : int  3 3 3 4 1 4 6 3 2 3 ...
 $ WorkLifeBalance         : int  3 3 2 3 3 3 4 2 3 4 ...
 $ YearsAtCompany          : int  8 2 1 4 10 5 1 5 3 1 ...
 $ YearsInCurrentRole      : int  7 2 0 2 9 3 0 3 2 1 ...
 $ YearsSinceLastPromotion : int  3 2 0 0 8 0 0 1 0 0 ...
 $ YearsWithCurrManager    : int  0 2 0 3 8 3 0 4 2 0 ...
> 
```

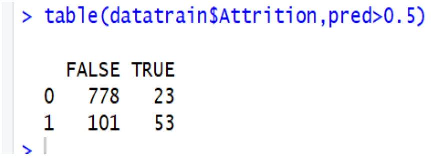
Well we have to calculate number of yes in the training set, so we applied the given command and obtained 154 'YES'. I can say that by looking at the output that says 1 in binary and its equivalent to 'YES', hence 154 counts of "YES".



```
> table(datatrain$Attrition)

  0   1 
801 154 
> 
```

2. (0.5pt) What does the baseline method predict on the training set? What is the accuracy of the baseline method?

Answer.2:

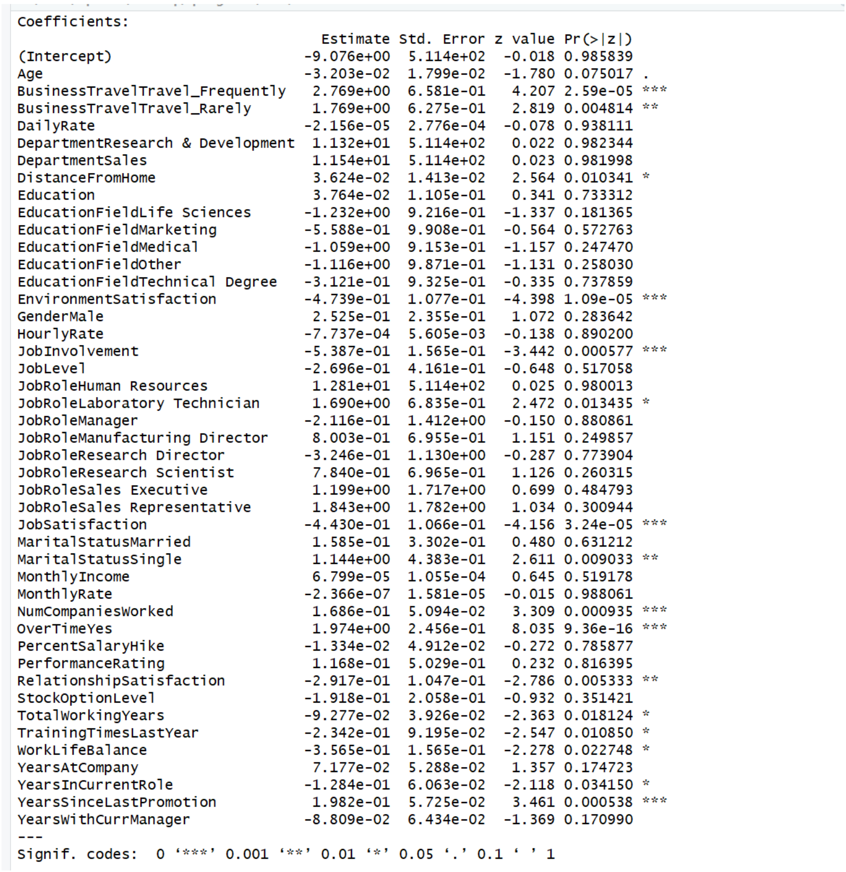


```
> table(datatrain$Attrition,pred>0.5)

   FALSE TRUE
  0   778   23
  1   101   53
>
```

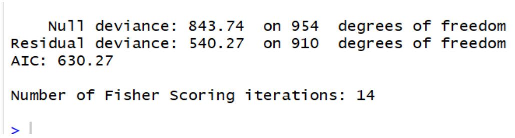
Well here I have assumed the threshold anonymously 0.5 for the calculations. So the baseline method predicts that the True negative is equivalent to 778 which directly means that the worker didn't quit.
Accuracy by the baseline method : 778+23/(778+23+101+53) = 801/955=83.8%.

3. (0.5pt) Run a logistic regression trying to predict attrition using all the independent variables provided. Show your output. NOTE: to avoid having to type a lot of names of independent variables, use simply . which means "all of them" (for instance: glm(y ~ . , data=DataTrain, family = "binomial"))

Answer.3 :



```
Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -9.076e+00  5.114e+02  -0.018 0.985839    
Age                              -3.203e-02  1.799e-02  -1.780 0.075017 .  
BusinessTravelTravel_Frequently   2.769e+00  6.581e-01   4.207 2.59e-05 ***
BusinessTravelTravel_Rarely       1.769e+00  6.275e-01   2.819 0.004814 ** 
DailyRate                        -2.156e-05  2.776e-04  -0.078 0.938111    
DepartmentResearch & Development  1.132e+01  5.114e+02   0.022 0.982344    
DepartmentSales                   1.154e+01  5.114e+02   0.023 0.981998    
DistanceFromHome                  3.624e-02  1.413e-02   2.564 0.010341 *  
Education                         3.764e-02  1.105e-01   0.341 0.733312    
EducationFieldLife Sciences      -1.232e+00  9.216e-01  -1.337 0.181365    
EducationFieldMarketing          -5.588e-01  9.908e-01  -0.564 0.572763    
EducationFieldMedical            -1.059e+00  9.153e-01  -1.157 0.247470    
EducationFieldOther              -1.116e+00  9.871e-01  -1.131 0.258030    
EducationFieldTechnical Degree   -3.121e-01  9.325e-01  -0.335 0.737859    
EnvironmentSatisfaction          -4.739e-01  1.077e-01  -4.398 1.09e-05 ***
GenderMale                        2.525e-01  2.355e-01   1.072 0.283642    
HourlyRate                       -7.737e-04  5.605e-03  -0.138 0.890200    
JobInvolvement                   -5.387e-01  1.565e-01  -3.442 0.000577 ***
JobLevel                         -2.696e-01  4.161e-01  -0.648 0.517058    
JobRoleHuman Resources            1.281e+01  5.114e+02   0.025 0.980013    
JobRoleLaboratory Technician      1.690e+00  6.835e-01   2.472 0.013435 *  
JobRoleManager                   -2.116e-01  1.412e+00  -0.150 0.880861    
JobRoleManufacturing Director     8.003e-01  6.955e-01   1.151 0.249857    
JobRoleResearch Director         -3.246e-01  1.130e+00  -0.287 0.773904    
JobRoleResearch Scientist         7.840e-01  6.965e-01   1.126 0.260315    
JobRoleSales Executive            1.199e+00  1.717e+00   0.699 0.484793    
JobRoleSales Representative       1.843e+00  1.782e+00   1.034 0.300944    
JobSatisfaction                  -4.430e-01  1.066e-01  -4.156 3.24e-05 ***
MaritalStatusMarried              1.585e-01  3.302e-01   0.480 0.631212    
MaritalStatusSingle               1.144e+00  4.383e-01   2.611 0.009033 ** 
MonthlyIncome                     6.799e-05  1.055e-04   0.645 0.519178    
MonthlyRate                      -2.366e-07  1.581e-05  -0.015 0.988061    
NumCompaniesWorked                1.686e-01  5.094e-02   3.309 0.000935 ***
OverTimeYes                       1.974e+00  2.456e-01   8.035 9.36e-16 ***
PercentSalaryHike                -1.334e-02  4.912e-02  -0.272 0.785877    
PerformanceRating                 1.168e-01  5.029e-01   0.232 0.816395    
RelationshipSatisfaction         -2.917e-01  1.047e-01  -2.786 0.005333 ** 
StockOptionLevel                 -1.918e-01  2.058e-01  -0.932 0.351421    
TotalWorkingYears                -9.277e-02  3.926e-02  -2.363 0.018124 *  
TrainingTimesLastYear            -2.342e-01  9.195e-02  -2.547 0.010850 *  
WorkLifeBalance                  -3.565e-01  1.565e-01  -2.278 0.022748 *  
YearsAtCompany                    7.177e-02  5.288e-02   1.357 0.174723    
YearsInCurrentRole               -1.284e-01  6.063e-02  -2.118 0.034150 *  
YearsSinceLastPromotion           1.982e-01  5.725e-02   3.461 0.000538 ***
YearsWithCurrManager             -8.809e-02  6.434e-02  -1.369 0.170990    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



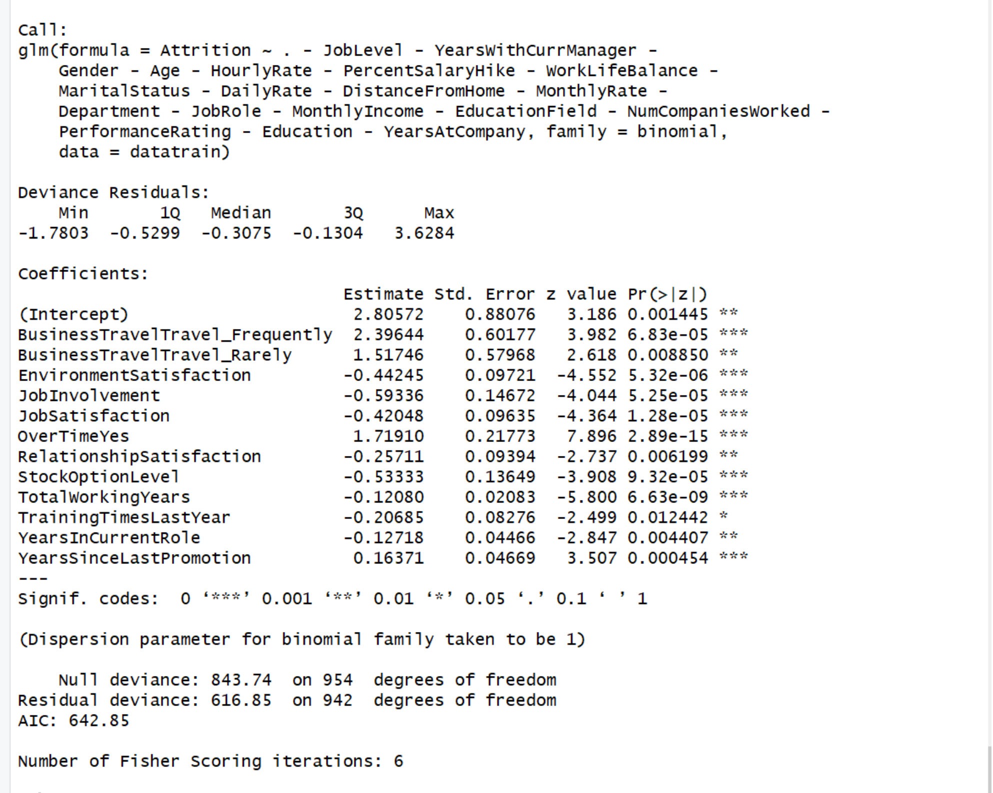
```
    Null deviance: 843.74  on 954  degrees of freedom
Residual deviance: 540.27  on 910  degrees of freedom
AIC: 630.27

Number of Fisher Scoring iterations: 14

>
```

The above shown is the output where AIC is 630.27 which is just used when comparing with another model. We have Business Travel, Environment Satisfaction, Job involvement, Job satisfaction, Over time and Years since last promotions variable which are very significant in the model.

4. (1.5pt) Refining the regression: Refine your logistic regression as much as you can. (Keep the . from the previous question and remove independent variables by using – in front of their name, for instance glm(y~ . – BusinessTravel,data=DataTrain,family=”binomial”)) NOTE: given how many independent variables you have to remove, it is ok here to remove a few at a time before rerunning the regression instead of doing it one at a time, but explain which ones you picked and why you stopped. Show your final output in your report.

Answer. 4:

The output after the refining the model is shown below. I refined the model 18 times before getting this output, I stopped for the given model because here each and every variable that is under consideration have p-value which is nearer to 0 and that tends to be more significant or have a high significance level. This was the main reason why I selected this model.
If I want to compare the model so I can do that with the previous model before refining which have an AIC of 630.27 and after refining the model I have an AIC of 642.85, so Definitely I will choose the model which have higher value of AIC.

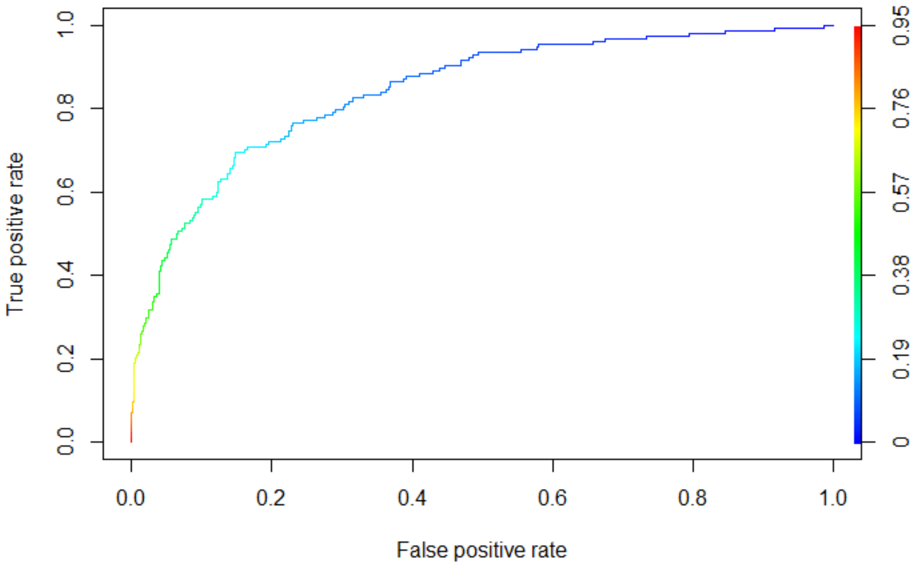


```
Call:
glm(formula = Attrition ~ . - JobLevel - YearsWithCurrManager -
    Gender - Age - HourlyRate - PercentSalaryHike - WorkLifeBalance -
    MaritalStatus - DailyRate - DistanceFromHome - MonthlyRate -
    Department - JobRole - MonthlyIncome - EducationField - NumCompaniesWorked -
    PerformanceRating - Education - YearsAtCompany, family = binomial,
    data = datatrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7803  -0.5299  -0.3075  -0.1304   3.6284

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      2.80572    0.88076   3.186 0.001445 **
BusinessTravelTravel_Frequently  2.39644    0.60177   3.982 6.83e-05 ***
BusinessTravelTravel_Rarely      1.51746    0.57968   2.618 0.008850 **
EnvironmentSatisfaction         -0.44245    0.09721  -4.552 5.32e-06 ***
JobInvolvement                  -0.59336    0.14672  -4.044 5.25e-05 ***
JobSatisfaction                 -0.42048    0.09635  -4.364 1.28e-05 ***
OverTimeYes                      1.71910    0.21773   7.896 2.89e-15 ***
RelationshipSatisfaction        -0.25711    0.09394  -2.737 0.006199 **
StockOptionLevel                -0.53333    0.13649  -3.908 9.32e-05 ***
TotalWorkingYears               -0.12080    0.02083  -5.800 6.63e-09 ***
TrainingTimesLastYear           -0.20685    0.08276  -2.499 0.012442 *
YearsInCurrentRole              -0.12718    0.04466  -2.847 0.004407 **
YearsSinceLastPromotion          0.16371    0.04669   3.507 0.000454 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 843.74  on 954  degrees of freedom
Residual deviance: 616.85  on 942  degrees of freedom
AIC: 642.85

Number of Fisher Scoring iterations: 6
```

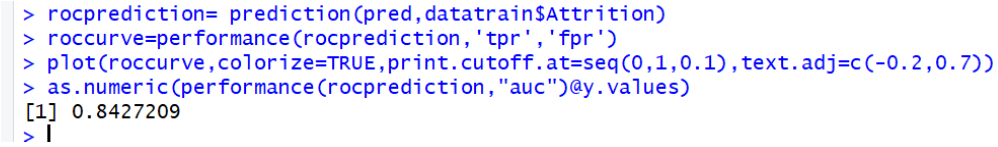
5. (1pt) Here we evaluate the performance of our (refined) model on the training set.
    a. Compute the ROC curve and compute the AUC.

    <u>Answer.5:</u>

    The ROC curve computed is shown below:



Area under the curve is 0.8427209 of the above ROC curve.



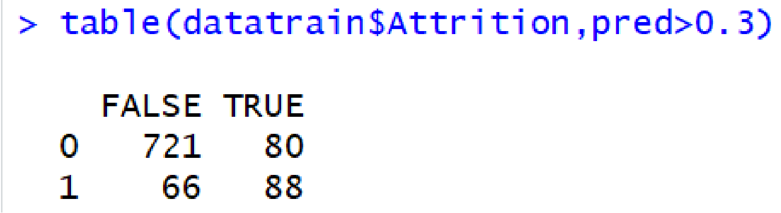
```
> rocprediction= prediction(pred,datatrain$Attrition)
> roccurve=performance(rocprediction,'tpr','fpr')
> plot(roccurve,colorize=TRUE,print.cutoff.at=seq(0,1,0.1),text.adj=c(-0.2,0.7))
> as.numeric(performance(rocprediction,"auc")@y.values)
[1] 0.8427209
> |
```

*note: curve is computed assuming the threshold value t=0.5

b. Argue for a choice of threshold t. For that threshold t, compute the confusion matrix.

Answer:

Here I am going to choose the threshold t=0.3, as I look at the computed confusion matrix here the true negative is more than false negative and true positive is more than false positive. This says our prediction for tn=721 which means 721 people don’t leave the job and tp=88 which means 88 people left the job. So this is the reason I choose threshold = 0.3 which gives our predicted value true than giving it false. So everything below 0.3 is unlikely and above is likely. Tr



```
> table(datatrain$Attrition,pred>0.3)

   FALSE TRUE
 0   721   80
 1    66   88
```

c. Then compute accuracy, true positive rate, true negative rate.

Answer:

Assuming t=0.3 as stated in the above part.
Accuracy= 721+88/721+66+88+80 = 84.71%
True positive rate =88/88+66 = 57.14%
Specificity = 721/721+80 = 90%
True negative rate = 1- specificity = 10%

d. Compare the accuracy you obtained with that of the baseline method.

Answer:

*note: t=0.3

Accuracy= 84.71%
Baseline method accuracy= 801/955= 83.8%
We have predicted a higher accuracy than the baseline method, which is great as it says 84.71% of our prediction is true, 721 are likely not to leave whereas 88 are likely to leave.

e. Are you satisfied with your model so far? Why or why not?

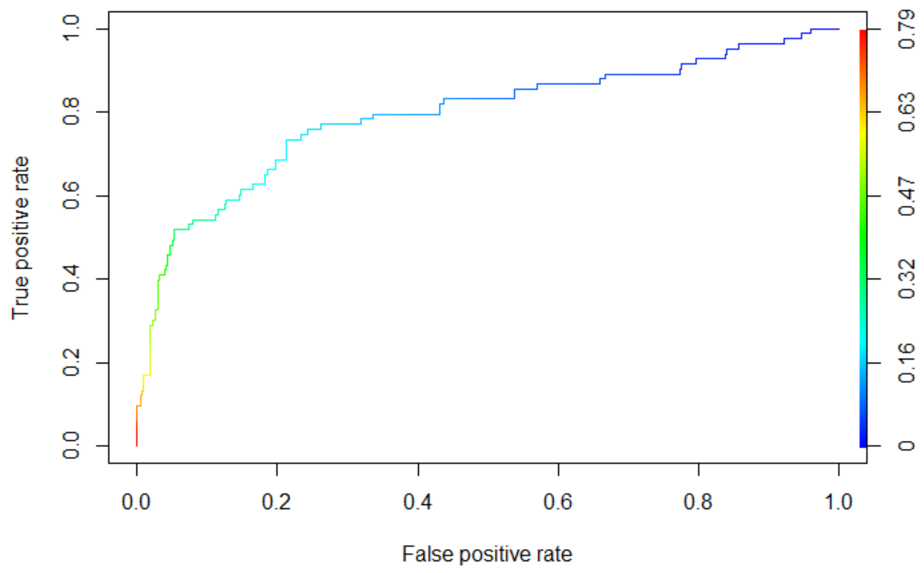
Answer:

Yes, I am satisfied with my model seeing the accuracy and the significance level of the model after refining it. I can say that my prediction of 721 people not leaving is higher than any of my prediction going wrong.

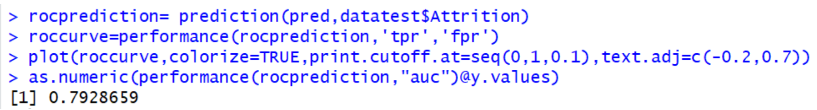
6. (1pt) Test your model on the testing set.
   a. Compute the ROC curve and compute the AUC.

Answer.6:

Below shown is the ROC curve:



Computed AUC as shown below:



```
> rocprediction= prediction(pred,datatest$Attrition)
> roccurve=performance(rocprediction,'tpr','fpr')
> plot(roccurve,colorize=TRUE,print.cutoff.at=seq(0,1,0.1),text.adj=c(-0.2,0.7))
> as.numeric(performance(rocprediction,"auc")@y.values)
[1] 0.7928659
```

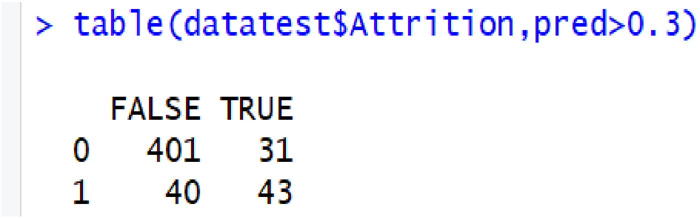
AUC is equivalent to 0.7928659 on testing sets.

b. For the same threshold as earlier, compute the confusion matrix.

Answer :

*note : t=0.3

I assumed t=0.3 while I was modelling for the training sets. Now for the testing sets the confusion matrix is as follows:



```
> table(datatest$Attrition,pred>0.3)

   FALSE TRUE
 0   401   31
 1    40   43
```

c. Then compute accuracy, true positive rate, true negative rate.

Answer:

*note : t=0.3

Accuracy : 401+43/401+40+43+31= 86.21 %
True positive rate: 43/43+40 = 51.80 %
Specificity : 401/401+31= 92.82%
True negative rate: 1-specificity = 1-92.82% = 7.17%

d. Compute the accuracy of the baseline method on the testing set and compare the two accuracies.

Answer:

*note: t=0.3

Accuracy = 401+43/401+40+43+31= 86.21%
Baseline method = 401+31/401+31+40+43= 83.88%

Over here our predicted is higher then the baseline method accuracy which seems to be great, which is one of the most important factor to be considered to prove your model is good.

e. Are you satisfied with your model?

Answer:

Well yes, I am satisfied with my model as we have great accuracy predicted on the testing model.