Homework 3

Ghandilyan Lilit

October 8, 2017

(1 point) Read the file Attrition.csv into R. The general objective of this analysis will be predicting whether certain employees will quit their job at this company or not (which is reflected in the Attrition variable). First, check whether the variables have the right data types and make appropriate corrections if needed.

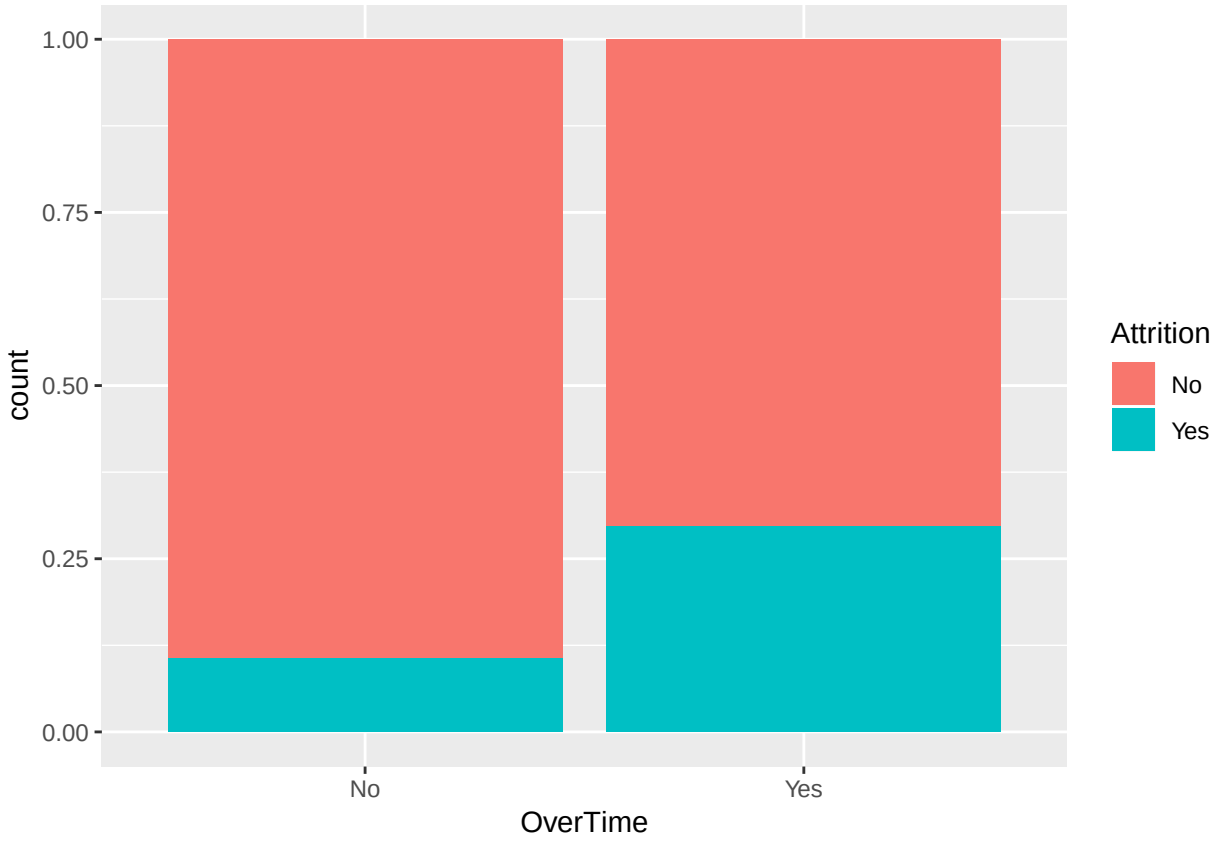
```
df <- read.csv("Attrition2.csv")
levels = c(1,2,3,4)
lab = c("low", "medium", "high", "very high")
df$EnvironmentSatisfaction <- factor(df$EnvironmentSatisfaction, levels, lab)
df$JobLevel <- as.factor(df$JobLevel)
df$StockOptionLevel <- as.factor(df$StockOptionLevel)
summary(df)
```

```
##  Attrition   DailyRate      EnvironmentSatisfaction JobLevel OverTime
##  No :987   Min.   : 102.0   low      :231           1:431    No :841
##  Yes:190   1st Qu.: 465.0   medium   :225           2:414    Yes:336
##            Median : 804.0   high     :367           3:183
##            Mean   : 801.8   very high:354           4: 93
##            3rd Qu.:1162.0                          5: 56
##            Max.   :1499.0
##  StockOptionLevel TotalWorkingYears YearsInCurrentRole
##  0:502            Min.   : 0.0      Min.   : 0.000
##  1:475            1st Qu.: 6.0      1st Qu.: 2.000
##  2:132            Median :10.0      Median : 3.000
##  3: 68            Mean   :11.4      Mean   : 4.205
##                   3rd Qu.:15.0      3rd Qu.: 7.000
##                   Max.   :40.0      Max.   :18.000
```

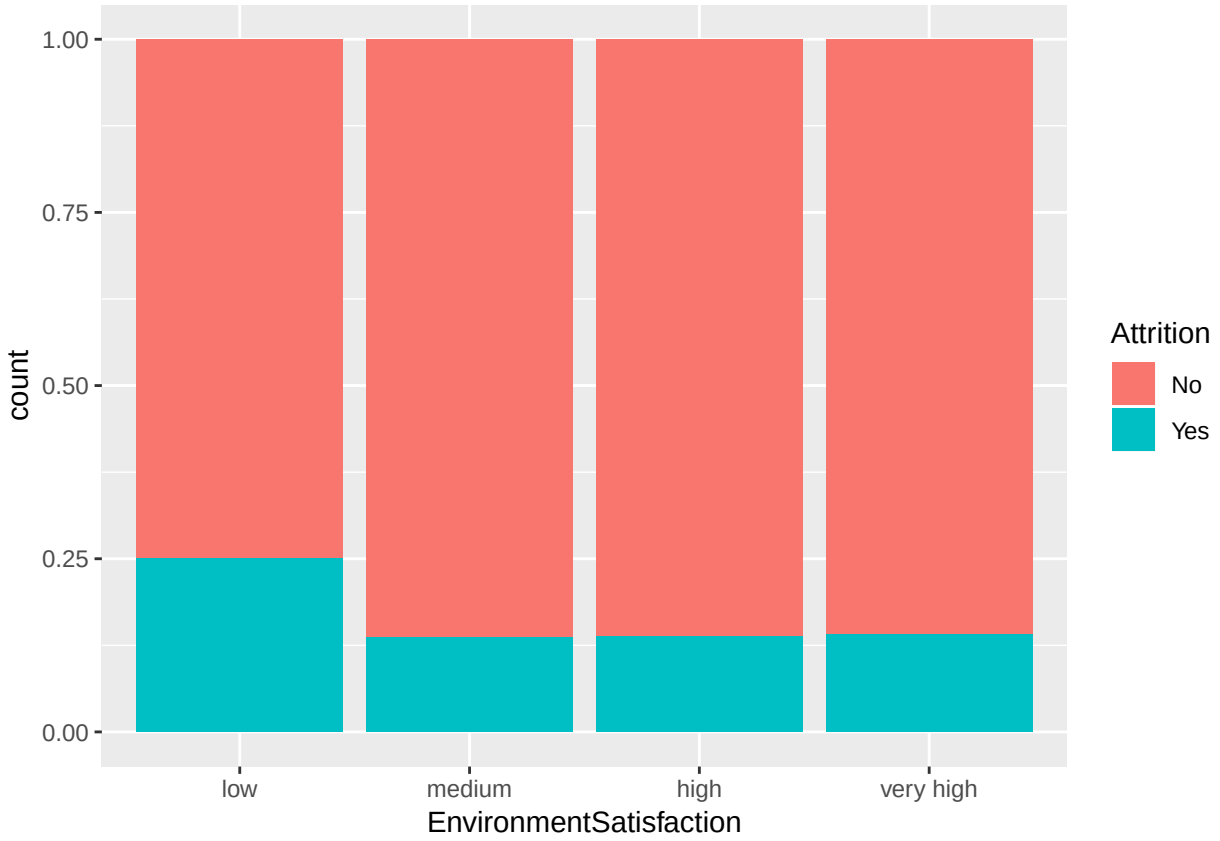
(4 points) Explore the relationship between the variables OverTime and Attrition. First calculate what is the probability that an emloyee leaves the company given that he/she worked overtime. Second, using ggplot2 construct a barplot to show the relationship between these variables (based on probabilities of attrition) Third, construct a barplot illustrating the relationship between Environment Satisfaction and Attrition Comment on all the results

```
counts <- addmargins(table(df$Attrition, df$OverTime))
prob <- 100/336

plt<-ggplot(df, aes(x=OverTime, fill=Attrition)) + geom_bar(position = "fill")
plt
```



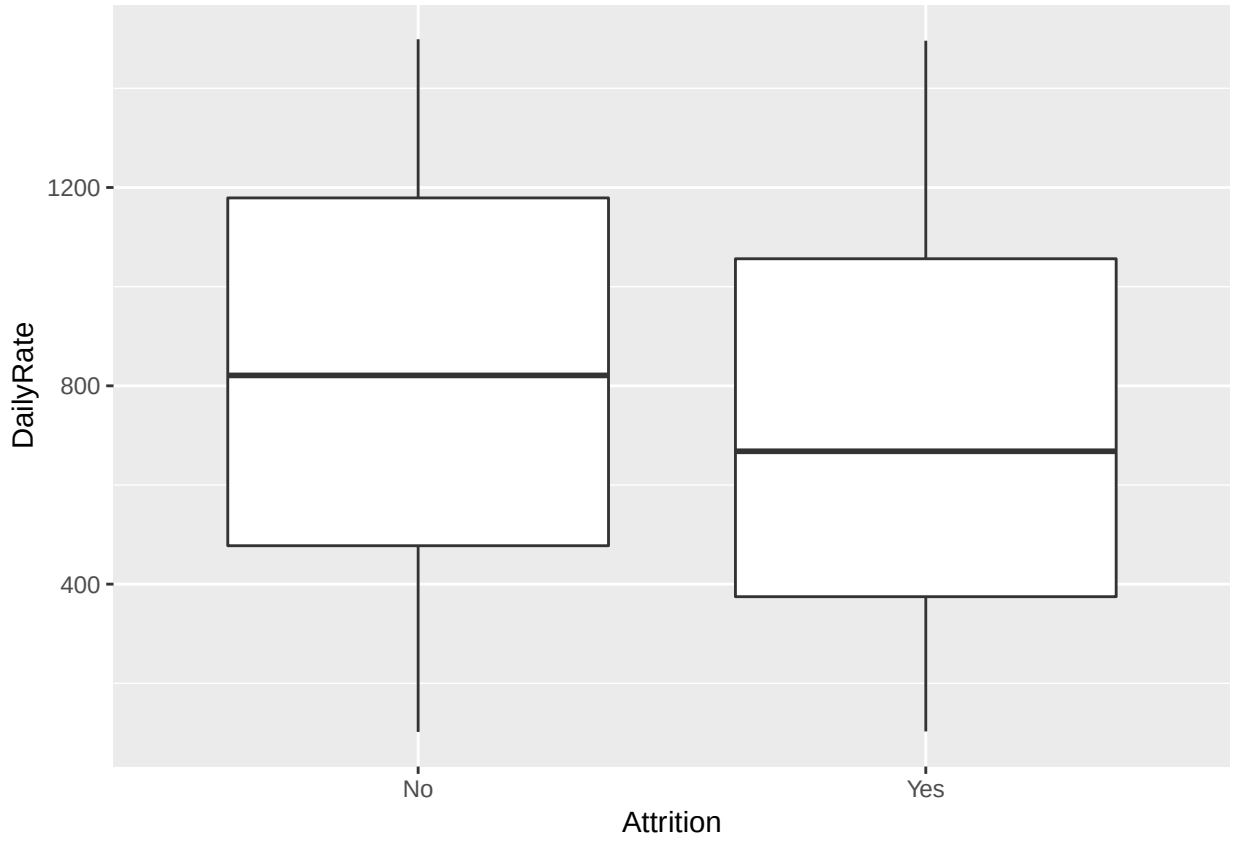

```
plt<-ggplot(df, aes(x=EnvironmentSatisfaction, fill=Attrition)) + geom_bar( position = "fill")
plt
```



The first plot shows that around 30 percent of the employees who worked overtime quitted, while around 12 percent of those who did not work overtime did. The second shows that the likelihood of Attrition is lower for the very high environment satisfaction level. For other levels attrition is almost equally likely.

(3 points) USing ggplot2 visualize the difference between the distribution of the DailyRate depending on the fact whether the employee left the company or stayed and comment on the differences.

```
ggplot(aes(Attrition, DailyRate), data = df) + geom_boxplot()
```



The median salary for the employees who did not quit is higher.

(1 point)Subset the data into 2 random samples, 1st one will serve as a training data set containing 80% of the cases from Attrition data set while the 2nd will be the test dataset containing the 20% of the cases, respectively.(Do not forget to set seed) This time, use caret package for data partitioning.

```
trainIndex <- createDataPartition(df$Attrition, p = 0.8, list = F)
Train <- df[trainIndex, ]
Test <- df[-trainIndex, ]
```

(5 points) Build a logistic regression model on the train data set having Attrition as a dependent variable and all the others as independent variables. Comment on which variables are significant for the model. Looking at the signs of coeffecients, comment whether the relationships are the same as while running the above analysis.

```
model <- glm(formula = Attrition~., family = "binomial", data = df)
summary(model)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6504  -0.5849  -0.3696  -0.1991   2.8065
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.4671274  0.2968764   1.573 0.115609
## DailyRate                    -0.0005004  0.0002188  -2.286 0.022226
```

```
## EnvironmentSatisfactionmedium    -1.0464047  0.2771713  -3.775 0.000160
## EnvironmentSatisfactionhigh      -0.9994466  0.2435764  -4.103 4.07e-05
## EnvironmentSatisfactionvery high -1.0119199  0.2457248  -4.118 3.82e-05
## JobLevel2                        -0.9915873  0.2341735  -4.234 2.29e-05
## JobLevel3                        -0.1734592  0.3171868  -0.547 0.584470
## JobLevel4                        -1.2713039  0.6302655  -2.017 0.043686
## JobLevel5                        -0.7244677  0.7496118  -0.966 0.333815
## OverTimeYes                       1.5397193  0.1854046   8.305  < 2e-16
## StockOptionLevel1                -1.1654695  0.2051293  -5.682 1.33e-08
## StockOptionLevel2                -1.3302965  0.3658287  -3.636 0.000276
## StockOptionLevel3                -0.7093284  0.3806028  -1.864 0.062364
## TotalWorkingYears                -0.0269147  0.0212725  -1.265 0.205789
## YearsInCurrentRole               -0.0849913  0.0328123  -2.590 0.009591
##
## (Intercept)
## DailyRate                        *
## EnvironmentSatisfactionmedium    ***
## EnvironmentSatisfactionhigh      ***
## EnvironmentSatisfactionvery high ***
## JobLevel2                        ***
## JobLevel3
## JobLevel4                        *
## JobLevel5
## OverTimeYes                      ***
## StockOptionLevel1                ***
## StockOptionLevel2                ***
## StockOptionLevel3                .
## TotalWorkingYears
## YearsInCurrentRole               **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1040.54  on 1176  degrees of freedom
## Residual deviance:  832.77  on 1162  degrees of freedom
## AIC: 862.77
##
## Number of Fisher Scoring iterations: 5
```

Categoric variables are considered significant if at least one level shows significance. Thus, all categoric variables are significant. Of the numeric variables, all are significant except TotalWorkingYears. In the analysis above, we were looking at the relationship between

1.Attrition and DailyRate As DailyRate increases the logits of Attrition decrease (negative correlation both in the plot and by coefficients)

2.Attrition and OverTime If an employee worked overtime the logits of Attrition increases (the same result both in the plot and by coefficients)

3.Attrition and Environment Satisfaction For higher levels of satisfaction level(the base level is low) the logits of Attrition decrease (the same result both in the plot and by coefficients)

(7 points) Print the coeficients of the model, also create the exponents of the coeficients. How will you interpret the coefficients in terms of their impact on odds and logit (log odds). Comment on 4 variables: 2 numeric and 2 categorical ones.

```
coefficients(model)
```

```
##                       (Intercept)                         DailyRate
##                      0.4671273806                     -0.0005003621
##    EnvironmentSatisfactionmedium      EnvironmentSatisfactionhigh
##                     -1.0464046897                     -0.9994465651
## EnvironmentSatisfactionvery high                         JobLevel2
##                     -1.0119199474                     -0.9915872518
##                         JobLevel3                         JobLevel4
##                     -0.1734591594                     -1.2713039343
##                         JobLevel5                       OverTimeYes
##                     -0.7244677248                      1.5397193377
##                 StockOptionLevel1                 StockOptionLevel2
##                     -1.1654695143                     -1.3302965045
##                 StockOptionLevel3                 TotalWorkingYears
##                     -0.7093284407                     -0.0269146527
##                YearsInCurrentRole
##                     -0.0849912996
```

```
exp(coefficients(model))
```

```
##                       (Intercept)                         DailyRate
##                         1.5954046                         0.9994998
##    EnvironmentSatisfactionmedium      EnvironmentSatisfactionhigh
##                         0.3511981                         0.3680831
## EnvironmentSatisfactionvery high                         JobLevel2
##                         0.3635204                         0.3709874
##                         JobLevel3                         JobLevel4
##                         0.8407515                         0.2804657
##                         JobLevel5                       OverTimeYes
##                         0.4845824                         4.6632813
##                 StockOptionLevel1                 StockOptionLevel2
##                         0.3117762                         0.2643989
##                 StockOptionLevel3                 TotalWorkingYears
##                         0.4919745                         0.9734443
##                YearsInCurrentRole
##                         0.9185203
```

For each unit increase in Years in current role, the logits of Attrition decrease by 0.0849, and the odds decrease by 8.14 percent. For each unit increase in Daily Rate, the logits of Attrition decrease by 0.0005, and the odds decrease by 0.05 percent.

If an employee worked overtime, the logits of Attrition increase by 1.53, and the odds increase by 466 percent. Odds of Attrition for StockOptionLevel1 is 31 percent of the odds of StockOption0, base level.

(4 points) Using the coeficients, calculate the probability of an employee to leave the company if the given employee's daily rate is 356, environment satisfaction is high, joblevel is 2, has worked overtime, uses 2nd stock option, has worked for 3 years and 2 years of this were in the same position.

```
exp <- exp( 0.4671273806 + 356 * -0.0005003621 -0.9994465651 -0.9915872518
            + 1.5397193377 -1.3302965045 + 3* -0.0269146527 + 2* -0.0849912996)
prob <- exp/(1+exp)
prob
```

```
## [1] 0.1488893
```

(3 points) Build and save another model eliminating the variables which were not significant predictors for

attrition.

```
model2 <- glm(formula = Attrition~.-TotalWorkingYears, family = "binomial", data = df)
summary(model2)
```

```
##
## Call:
## glm(formula = Attrition ~ . - TotalWorkingYears, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7347  -0.5788  -0.3702  -0.1975   2.7914
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      0.3363204  0.2780390   1.210 0.226426
## DailyRate                       -0.0005076  0.0002186  -2.323 0.020199
## EnvironmentSatisfactionmedium   -1.0198407  0.2762232  -3.692 0.000222
## EnvironmentSatisfactionhigh     -0.9608977  0.2412876  -3.982 6.82e-05
## EnvironmentSatisfactionvery high -0.9962501  0.2453831  -4.060 4.91e-05
## JobLevel2                       -1.0979763  0.2187515  -5.019 5.19e-07
## JobLevel3                       -0.3788375  0.2740245  -1.382 0.166820
## JobLevel4                       -1.7540267  0.5041042  -3.479 0.000502
## JobLevel5                       -1.2463610  0.6280333  -1.985 0.047195
## OverTimeYes                      1.5283182  0.1847965   8.270  < 2e-16
## StockOptionLevel1               -1.1761947  0.2049162  -5.740 9.47e-09
## StockOptionLevel2               -1.3195862  0.3660513  -3.605 0.000312
## StockOptionLevel3               -0.7064126  0.3810152  -1.854 0.063735
## YearsInCurrentRole              -0.0971542  0.0310724  -3.127 0.001768
##
## (Intercept)
## DailyRate                       *
## EnvironmentSatisfactionmedium   ***
## EnvironmentSatisfactionhigh     ***
## EnvironmentSatisfactionvery high ***
## JobLevel2                       ***
## JobLevel3
## JobLevel4                       ***
## JobLevel5                       *
## OverTimeYes                     ***
## StockOptionLevel1               ***
## StockOptionLevel2               ***
## StockOptionLevel3               .
## YearsInCurrentRole              **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1040.54  on 1176  degrees of freedom
## Residual deviance:  834.41  on 1163  degrees of freedom
## AIC: 862.41
##
## Number of Fisher Scoring iterations: 5
```

(8 points) Use the 2 models you built to make predictions on the test data set Using the threshold of 0.5, convert probabilities into classes.Create a confusion matrixes for both. Calculate the accuracy of the models,sensitivity and specificity,Positive and negative predictive values(PPV and NPV). Make comments on the models' peformance based on these indicators. Are those performing better than comparing accuracy with no information rate?

```
pr <- predict(model, newdata = Test, type = "response")
pr1 <- factor(ifelse(pr > 0.5, "Yes", "No"))
confusionMatrix(data = pr1, Test$Attrition, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  194  27
##        Yes   3  11
##
##                Accuracy : 0.8723
##                  95% CI : (0.8228, 0.9122)
##     No Information Rate : 0.8383
##     P-Value [Acc > NIR] : 0.08907
##
##                   Kappa : 0.3681
##  Mcnemar's Test P-Value : 2.679e-05
##
##             Sensitivity : 0.28947
##             Specificity : 0.98477
##          Pos Pred Value : 0.78571
##          Neg Pred Value : 0.87783
##              Prevalence : 0.16170
##          Detection Rate : 0.04681
##    Detection Prevalence : 0.05957
##       Balanced Accuracy : 0.63712
##
##        'Positive' Class : Yes
##
```

```
Accuracy1 <- (195 + 11)/235
Sensitivity1 <- 11/38
Specificity1 <- 195/197
PPV1 <- 11/13
NPV1 <- 195/(195+27)

pr2 <- predict(model2, newdata = Test, type = "response")
pr2 <- factor(ifelse(pr2 > 0.5, "Yes", "No"))
confusionMatrix(data = pr2, Test$Attrition, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  193  27
##        Yes   4  11
##
##                Accuracy : 0.8681
```

```
##                  95% CI : (0.818, 0.9086)
##     No Information Rate : 0.8383
##     P-Value [Acc > NIR] : 0.1231
##
##                   Kappa : 0.3562
##  Mcnemar's Test P-Value : 7.772e-05
##
##             Sensitivity : 0.28947
##             Specificity : 0.97970
##          Pos Pred Value : 0.73333
##          Neg Pred Value : 0.87727
##              Prevalence : 0.16170
##          Detection Rate : 0.04681
##    Detection Prevalence : 0.06383
##       Balanced Accuracy : 0.63458
##
##        'Positive' Class : Yes
##
```

```
Accuracy2 <- (195 + 10)/235
Sensitivity2 <- 10/38
Specificity2 <- 195/197
PPV2 <- 10/12
NPV2 <- 195/(195+28)
```

Both models have high specificity, meaning they correctly identify the employees who are not going to quit. However, the sensitivity value is very low, meaning that many employees who are going to quit are not identified correctly. This might mean that the cutoff value is too high. Both models perform better than the No information Rate model.

(7 points) Use the threshhold of 0.7 while creating the confusion matrix for the second model. Elaborate on the changes in sensitivity, specificity, PPV, NPV and overall accuracy. which classification makes your model better?

```
pr2 <- predict(model2, newdata = Test, type = "response")
pr2 <- factor(ifelse(pr2 > 0.7, "Yes", "No"))
confusionMatrix(data = pr2, Test$Attrition, positive = "Yes")
```

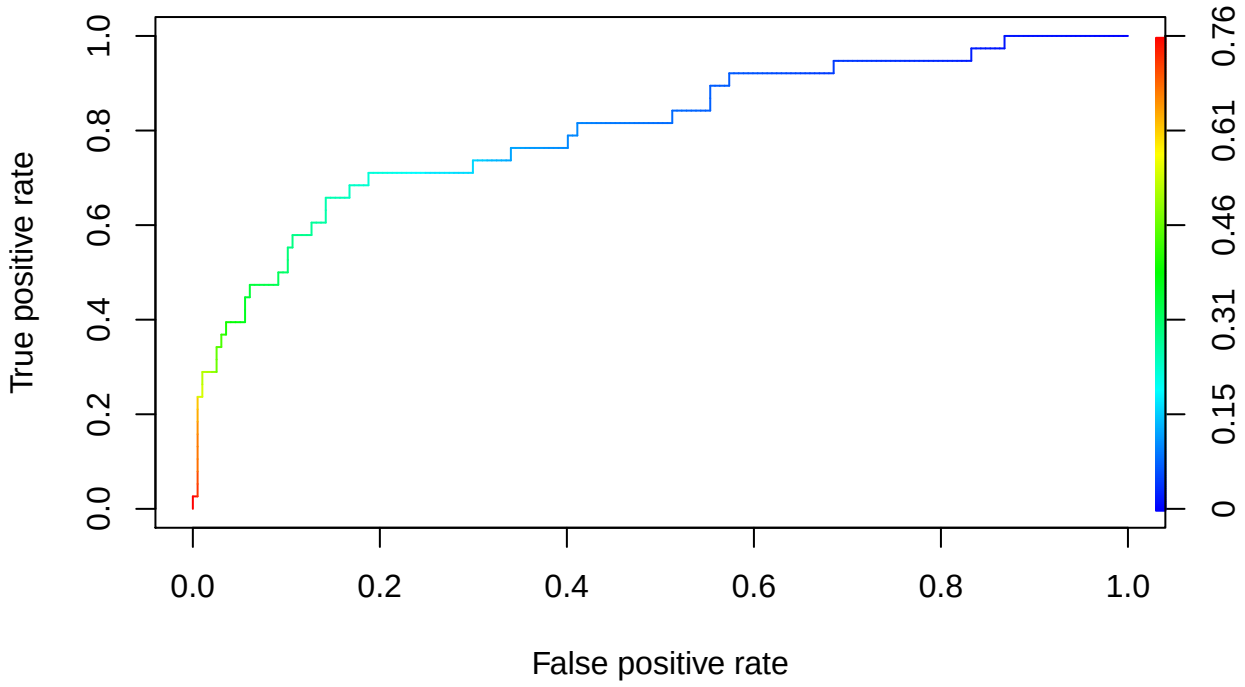
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  196  37
##        Yes   1   1
##
##                Accuracy : 0.8383
##                  95% CI : (0.7849, 0.883)
##     No Information Rate : 0.8383
##     P-Value [Acc > NIR] : 0.5432
##
##                   Kappa : 0.0344
##  Mcnemar's Test P-Value : 1.365e-08
##
##             Sensitivity : 0.026316
##             Specificity : 0.994924
##          Pos Pred Value : 0.500000
```

```
##          Neg Pred Value : 0.841202
##              Prevalence : 0.161702
##          Detection Rate : 0.004255
##    Detection Prevalence : 0.008511
##       Balanced Accuracy : 0.510620
##
##        'Positive' Class : Yes
##
```

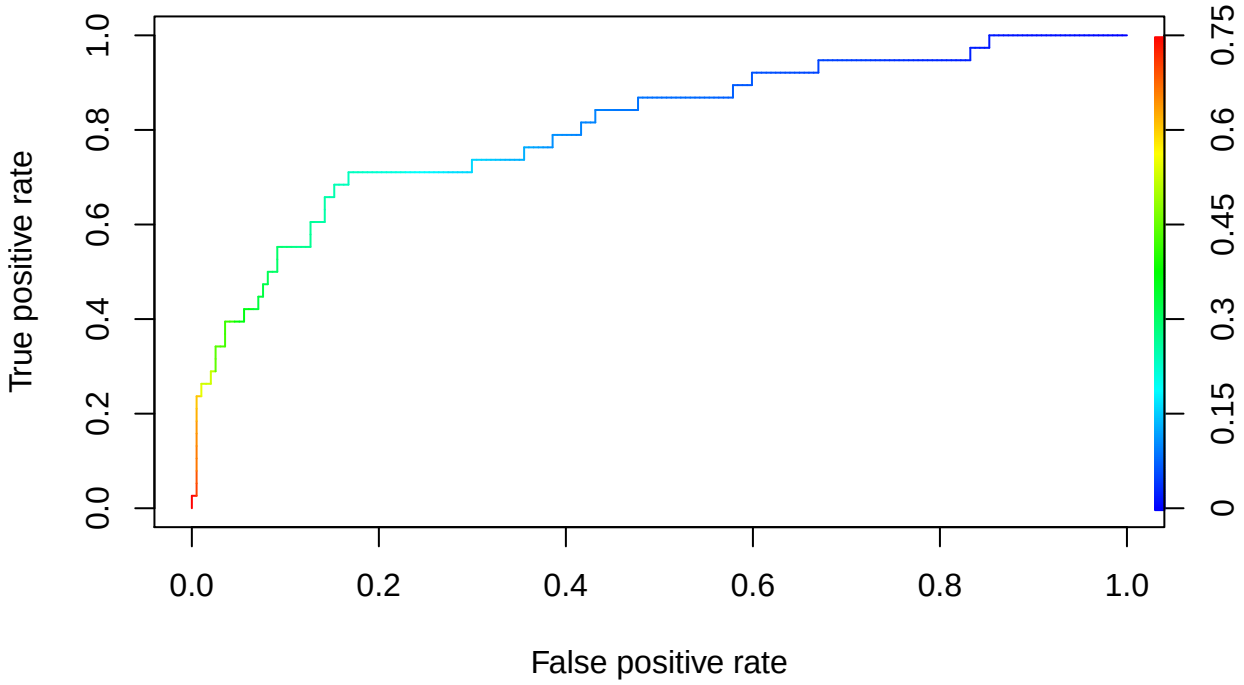
We have lower sensitivity, which means more people who are going to quit are not identified. This supports the claim that the cutoff value should be lower.

(7 points) Create ROC curves for each model and comment on it. Calculate AUC value for each model and make relevant comparisons.

```
pr <- predict(model, newdata = Test, type = "response")
pred1 <- prediction(pr, Test$Attrition)
perf = performance(pred1, "tpr", "fpr")
plot(perf, colorize = TRUE)
```




```
pr2 <- predict(model2, newdata = Test, type = "response")
pred2 <- prediction(pr2, Test$Attrition)
perf = performance(pred2, "tpr", "fpr")
plot(perf, colorize = TRUE)
```




```
auc <- performance(pred1, "auc")@y.values
auc
```

```
## [[1]]
## [1] 0.8013625
```

```
auc2 <- performance(pred2, "auc")@y.values
auc2
```

```
## [[1]]
## [1] 0.8048357
```

The second model has higher AUC value, because we removed a non-significant variable, although the values are not too different. The higher the ROC value, the worse is the model.