**Final Project Report**

**HR Employee Attrition**

**Spring 2019**

**Team Members**

Ann-Sophie Kouadio IV

Karan Dassi

Tushar Jethani

**Professor:**  J. Alberto Espinosa, Kogod School of Business, American University

Table of Contents

[Analytics question: 3](#_Toc8137534)

[Business Rationale: 3](#_Toc8137535)

[Dataset Identification: 3](#_Toc8137536)

[Descriptive Statistics: 4](#_Toc8137537)

[Best subset predictor selection: 4](#_Toc8137538)

[Data Preprocessing: 5](#_Toc8137539)

[Test for multicollinearity and autocorrelation: 5](#_Toc8137540)

[Assumptions to be met: 5](#_Toc8137541)

[Predictive models used: 5](#_Toc8137542)

[Logistic regression with cross validation: 6](#_Toc8137543)

[Classification Trees with cross validation: 6](#_Toc8137544)

[Conclusion: 7](#_Toc8137545)

[Results: 7](#_Toc8137546)

[Appendices 8](#_Toc8137547)

## Analytics question:

*What will be the future employee turnover for IBM?*

Here we are trying to predict a classification outcome primarily attrition (which is binomial in nature) – the two categories are: Terminated (Yes/1) Not Terminated (No/0) Our motive here is to build a strong logistic predictive model for IBM so that we can re-use the model in the future and use it as a framework to predict employee turnover. We are trying to find out the proportion of employees that will be terminated for IBM for our test dataset. Here our test data is a portion of all the employees for IBM.  
Also, note that we are building this predicitve model for every employee in the company irrespective of their department, age, gender or any other factor.

### Articulation of Goals:

1. Interpretation: What are the most significant factors in determining employee turnover?
2. Inference/Prediction: Building a strong predictive model for future employee churn (terminations) which can be referred again for predictions.

## Business Rationale:

While in the past decades, most companies have improved their analytics capabilities in areas such as Supply Chain, Accounting and Marketing, fewer are using analytics in their workforce to make predictions and take action on future HR issues.  
The aim of this project is to encourage companies such as IBM to invest more in HR analytics to improve their interactions with their employees before issues arise. We believe that the predictive models that will be built throughout this project will provide the company with insights on future employee turnover irrespective of department and roles. To do so, our predictive analysis will look at many factors such as distance from home, environment satisfaction, monthly income, years of experience, department, etc.  
We believe that IBM could benefit from implementing our models in their company to take actions on the identified areas of concerns and reduce costs of turnover related to employee resignations and terminations by improving upon the most significant factors such as Job satisfactio, distance from home, travel for the company, overtime, job satisfaction, etc.

## Dataset Identification:

The data set used for this analysis is based on a Kaggle dataset created by IBM data scientists. This data set contains information on 1470 employees of IBM with 35 variables representing specific characteristics of each employee including their Attrition status (if they left IBM or not). This IBM data consists of both numerical and categorical variables. An extensive list of these data and a brief description of each variable is available [see appendix A, table 1 ].  
The dependent variable here is Attrition which is a binomial categorical variable with an outcome of “Yes” meaning an employee left the company or “No” meaning the employee is still with the company.

## Descriptive Statistics:

The aim of our project is to build a model that will be able to predict if an employee will leave IBM or not. Contrary to what the majority believe the factors that contribute to an employee’s attrition are not all related to the financial benefits a company offers to its employees. In fact, it is impossible to conclude from the scatter plot [see appendix A, figure 4] the salary level of an employee lead to attrition.  
The points of this plot for both employees leaving or staying at the company do not vary much based on the salary range. Looking at employees that are receiving the lowest salary, we see that most of them decide to stay at IBM regardless of their income. This implies that other factors are more significant to determine attrition. Moreover, the bar chart in *appendix A, figure 1* showed that most of the employee at IBM tends to stay at the company. In fact, about only 16% of the employees have left the company. This rate is somewhat high knowing that a good attrition rate for a company should not exceed 10%.  
Additionally to monthly income, age seemed to be an interesting factor to take into consideration when analyzing attrition rate of a company. Looking at the density chart, we are able to determine the median age of the employee leaving the company which is approximately 31 years old; this indicates that the portion of employees leaving IBM are relatively young and suggest that some the reasons why they are leaving the company might be related to the level of satisfaction as well as other qualitative factors.

Another common and essential factor to consider when understanding the reasons why employees leave is the level of satisfaction. Indeed, job descriptions tend to differ a lot from what the company presents to the employee at the earlier phase of employment to the actual assigned duty when the employee begin his/her function.

Employees that feel betrayed by their job descriptions might tend to leave the company to explore other opportunities. A bar chart displaying employee attrition in terms of job satisfaction confirm this statement. In fact, the figure shows the proportion of employees that are not satisfied with their job (job satisfaction level 1) and leave the company is higher compared to the proportion of employees are satisfied with their job but still leave the IBM. We have also created various other plots to check for relationships between factors [see appendix A].

We also went ahead and checked for mean, median and other characteristics and found out that they very well support our graphs above. [See Appendix B]

## Best subset predictor selection:

1. Firstly, we went on created the best subset of predictors according to business problem at hand which gave us quite a good logistic predictive model.  
2. Then, we checked for correlation between predictors and were able to eliminate one variable [see appendix D].  
3. Then, we moved on to selecting the most significant variables using stepwise method:  
Here we went down to 14 variables. [see Appendix F]

## Data Preprocessing:

As we had a mixture of variables i.e. categorical as well as quantitative variables, so it was pretty hard to check for correlation between all the variables at once. So, to process the data we used the following steps:  
1. Reduced the full set of variables from 35 to 19 predictors based on business sense.  
2. Then, we ran a correlation matrix on all the quantitative variables [see appendix D], we were easily able to eliminate Total working years at the company because it has very high serial correlation and is not that significant.  
3. Then, we checked for correlation between categorical variables using chi square test and found out that Overtime was highly correlated but is very significant, thus, cannot remove it [see appendix D].  
4. We also checked for relation between categorical and quantitative variables by using box plots and could not find any good variables to eliminate [see appendix E].  
4. Finally, we ran a stepwise regression to find the best set of variables and could get to a final count of 14 predictors eliminating 5 variables [see appendix F].

## Test for multicollinearity and autocorrelation:

We checked for autocorrelation for both business model as well as final reduced model with Durbin Watson test where we found out that the DW statistic is 1.9045 and 1.9051 respectively, which are well between the limits, thus we can declare that there is acceptable autocorrelation between residuals for both models [see appendix I].  
We also checked for Variance Inflation Factors for all the predictors in bothh our models and they were well under 10 for both the models, while the VIFs for the final reduced model has all the VIF values for various variables close to one but for the business model had a couple of variables with VIF’s close to 9 which were eventually eliminated by stepwise regression. [see appendix I]

## Assumptions to be met:

Observation independence: Eliminate serial correlation [see section above]  
Appropriate outcome variable: Binomial outcome variable  
Multicollinearity: Checked it after running the models. [see section above]  
Linear relation between independent variables and log odds   
Large Sample Size: We have quite a big dataset  
All the above assumptions were met during our project with good results.

## Predictive models used:

We basically used two models with three different predictor sets. The models we used are logistic regression and classification trees as our outcome (dependent) variable is binomial in nature.  
As we can see that all the assumptions for these have been met, we could go ahead and implement these models.

## Logistic regression with cross validation:

Firstly, we applied the logistic regression to the full model, then to the business reduced model and finally to the model after perfoming the variable selection method (stepwise variable selection). When, we checked for AIC for the models the AIC for final reduced model dropped down to 998 from 1007 [see appendix F].  
Also, we cross validated the business model and the variable reduced final model using confusion matrix and after checking the predictive accuracy, we could see that at lambda = 0.275, we could get the most balanced model. As our aim was to check for people leaving the company we had to reduce the lambda value to get a strong predictive model. So our final logistic model had an accuracy rate of 82 percent over business models accuracy rate of 81 percent, the sensitivity rate of 66 percent over 63 percent and finally the specificity rate of 87 percent over 85 percent. We also checked the models with lambda values at 0.5, and 0.3, while these models gave us a good accuracy and specificity rate but the sensitivity rates were quite low and hence lamba = 0.275 [see appendix K].  
We also checked for area under ROC curve which originally came as 0.832 which was already pretty good and then, the area under ROC curve for final reduced model turned out to be 0.836 which is not significantly large but the final predictors gave us a pretty balanced predictive model [see appendix J & K].

## Classification Trees with cross validation:

Firstly, we tried to apply classification tree to the full business model which did not give quite significant results and there was bad accuracy as well. So, we went ahead and applied classification decision tree to the final reduced variables (created after eliminating serially correlated variables and selecting variables using stepwise regression) and got some pretty good results to decide whether if an employee will be terminated or not but as a whole, the accuracy of logistic model is pretty good when compared to classification tree.  
After cross validating the tree at mindev = 0.005, we found that the best number of nodes was 13 with a misclass of 238 which streamlined after 13 nodes for further number of nodes and misclass was very high before 13.  
We also found out that the 2LL for this model was pretty high at 1014 when compared to the 2LL of our logistic model of 962 [see appendix L].  
Then, we went ahead and cross validated the accuracy of the classification tree model using confusion matrix and Area under ROC curve. Accuracy found by using confusion matrix was 86 percent which is higher than logistic model at lambda = 0.275 but this did not give us a balanced model with a pretty low sensitivity of 57 percent compared to 666 percent for our logistics model [see appendix L].  
The above was also prooved by checking the area under ROC curve which was 0.739 for the classification tree which is significantly lower than 0.836 for our final logistic model [see appendix L].  
Looking at the above results, we concluded that our final results will be based upon reduced logistic mode.

## Results:

Using logistic regression, we found out that for our test dataset will have approximately 20 percent of the employees will be leaving the company (or the company will terminate them) [see appendix N]. The total number of employees in our test data set was almost 300 out of which 20 percent of them will be left from the company. We believe though at lambda 0.5 only 7 percent will be leaving but as can be seen above lambda at 0.275 gives the best true positives and that is what we would go for!

## Conclusion:

Employee attrition represent a threat for organizations, especially when there is no tools or strategy in place to address it. Predictive modeling could help HR departments to have a better understanding of the factors that motivates employee to leave. More importantly, predicting model generate insights for organizations which categories of employees might be at higher risk of attrition. While some level of attrition is expected in all businesses, it is important to minimize its level to avoid significant costs related to hiring and training. In our study we’ve built and tested several predictive methods on data set from IBM. The outcome of the analytic question here is binary therefore we applied Classification Tree and Logistic Regression models. We then compared the two model performances and chose the Logistic Regression model as the best model. While the classification model provided inconclusive results, the logistic regression provided consistent results with an outcome of 20.06 % employees that are subject to leave.

Having such information about significant predictors such as business travel, job satisfaction, age, gender, environment satisfaction, overtime, etc. could be beneficial for IBM to take preventive actions to retain its talent. We believe the final model is quite robust and will help IBM’s HR department overcome this business problem.

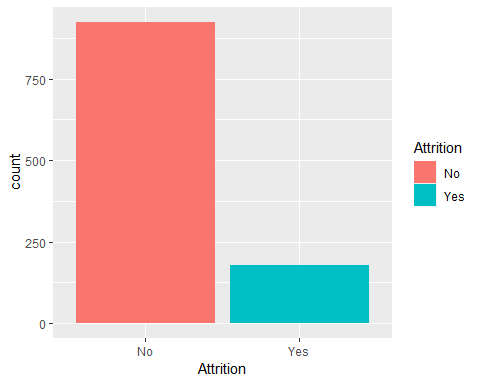
## Appendices

**Appendix A**

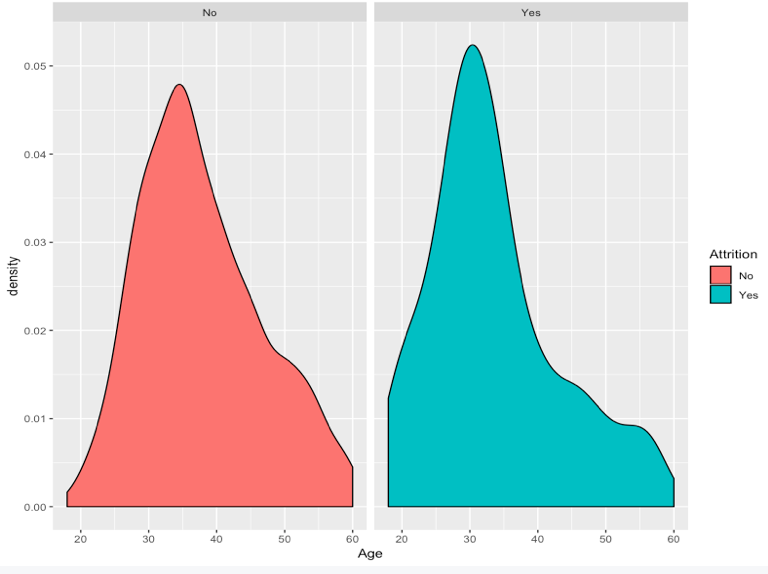
***Table 1: Few Variable Description***

|  |  |
| --- | --- |
| **Predictors** | **Descriptions** |
| **Age** | Integer with various values |
| **Business Travel** | Factor with 3 level indicating the level of travel included in an employee duty |
| **Department** | Factor with 3 levels indicating the Department where an employee works |
| **Distance From Home** | The distance in kilometers from work to the employee’s home |
| **Environment Satisfaction** | Factor with 4 level showing how satisfy is the employee with the work environment. 4 being the highest. |
| **Gender** | Factor with 2 levels "Female", "Male". Indicating the employee’s gender. |
| **Job Level** | Factor with 5 levels “1, 2, 3, 4, 5” with 5 being highest job level. Indicating the employee position within the organization. |
| **Job Satisfaction** | Factor with 4 levels “1, 2, 3, 4” with 4 being highly satisfied. Indicating how satisfy the employee is with his/her duty. |
| **Marital Status** | Factor with 3 levels “Single”, "Divorced", "Married". Indicating the matrimonial status of an employee. |
| **OverTime** | Binary variable with 2 levels “No”,”Yes” indicating if an employee is allow to work over time. |
| **Relationship Satisfaction** | Factor with 4 levels “1, 2, 3, 4” with 4 being the highest. Indicating how satisfy an employee is with his/her manager. |
| **WorkLifeBalance** | Factor with 4 levels “1, 2, 3, 4” with 4 being the highest. Indicating how satisfy an employee is with his/her job in relation to his life. |
| **Years Since Last Promotion** | Integer with various values indicating the number of years since an employee had a promotion. |
| **Years with current manager** | Integer with various values. Number of years an employee has been working under his/her current manager. |
| **Monthly Income** | Integer with various values. Indicating the monthly salary of an individual. |
| **Performance Rating** | Factor with 2 levels indicating an individual job performance. |
| **Total working years** | Integer with various values. Indicating an employee’s years of experience. |

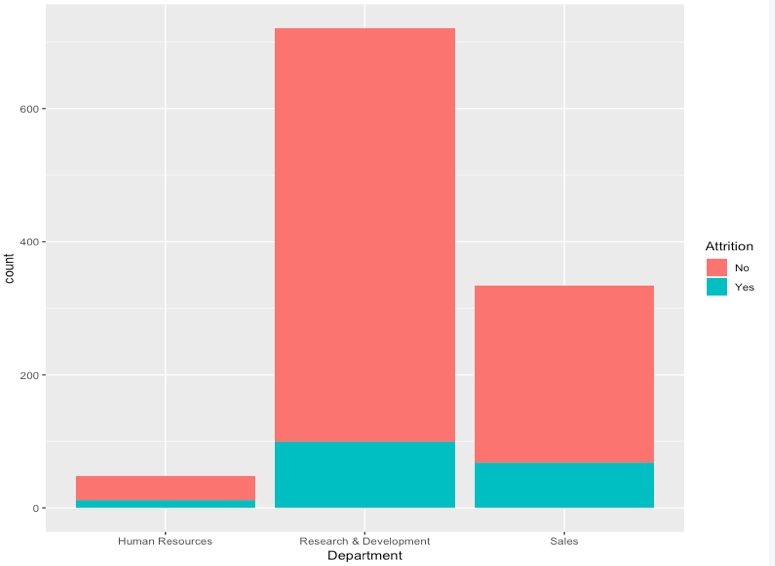
### Various graphs for our dataset for visualization:



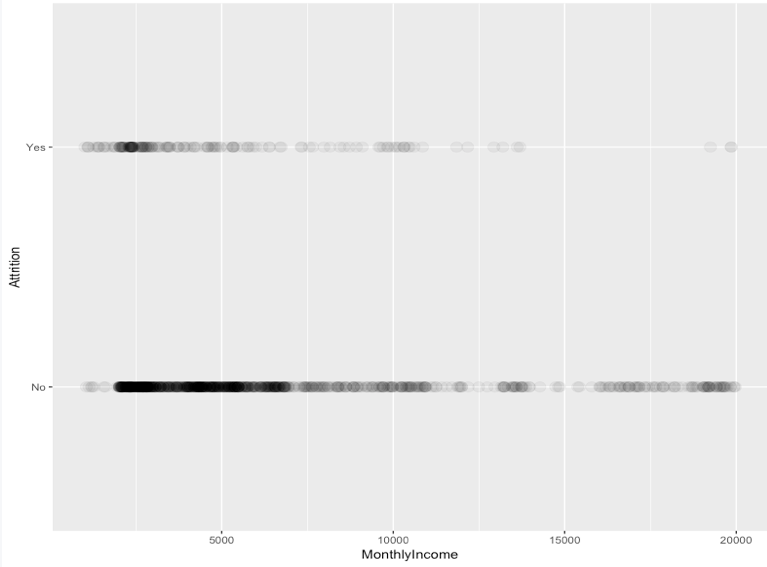
***Figure 1:*** *Overall attrition at IBM*: Approximately 236 employees leave IBM each year compared to 1234 that stayed.



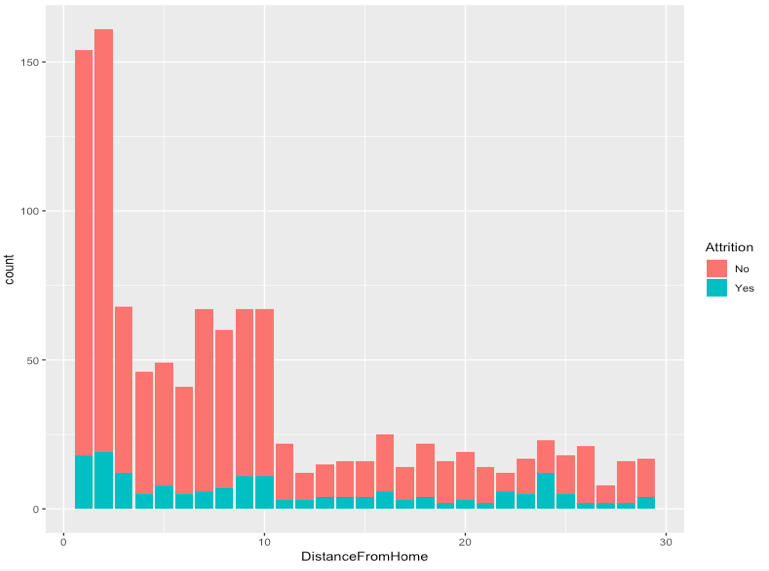
***Figure 2:*** Attrition based on Employee’s Age at IBM: *The median age of employee leaving the company is 31.*



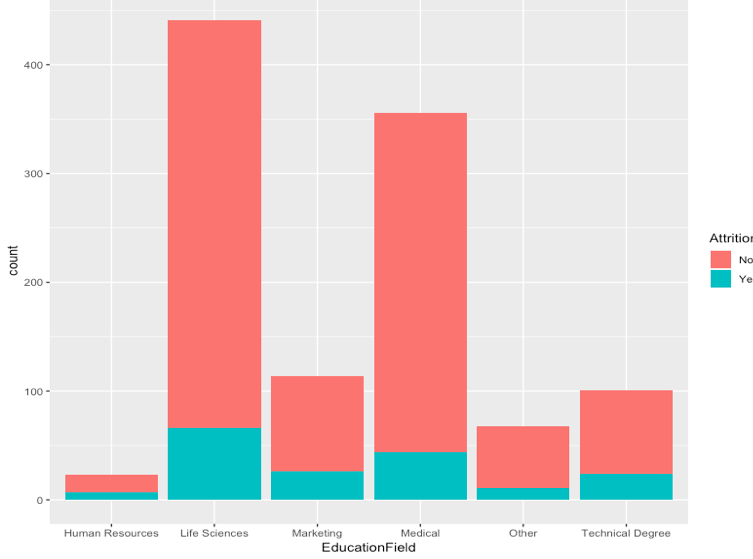
***Figure 3:*** Proportion of *Attrition within each Department:*  The proportion of employees leaving IBM vary from one department to another. The Research & Development department has the greatest number of employees leaving the company overall, closely followed by the sales department.



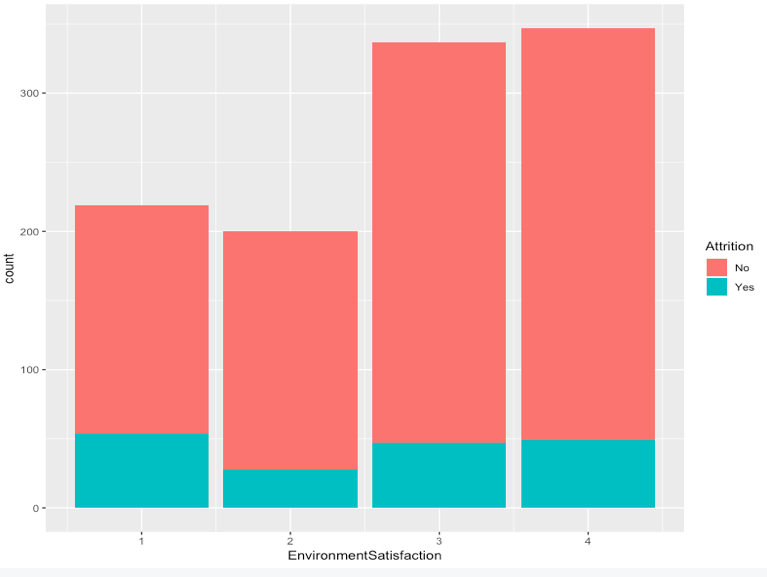
***Figure 4:*** *Attrition based on Employee’s Monthly Income*: It is impossible to conclude from this graph that Monthly Income play a significant role in an Employee’s decision to leave the company, as employee on the lower range of salary still decide to stay at IBM.



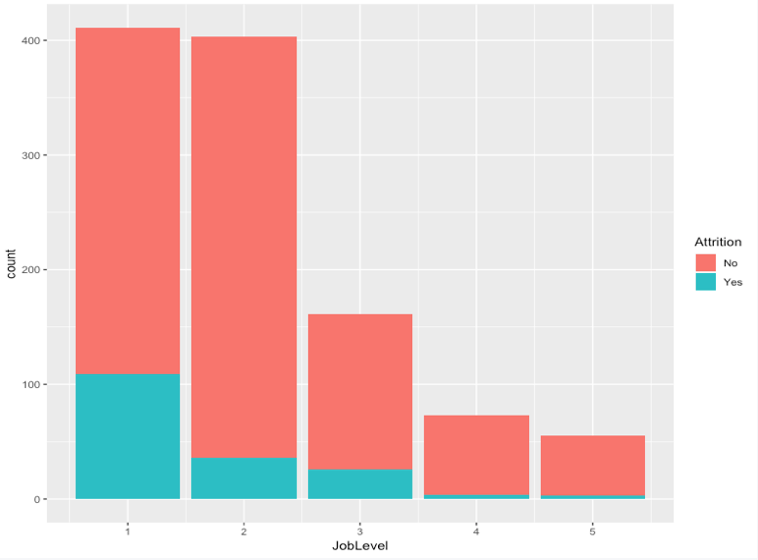
***Figure 5:*** *Attrition based on work’s distance from home*: Majority of employees of have left IBM leaved closed to the Office.



***Figure 4:*** *Attrition based on Employee’s Education Field*: The bar chart shows that few employees from the HR department have left the organization. This might be explained by the size of the division. As we can we can see on the graph the larger the department is the higher the number of employees leave.



***Figure 5:*** *Attrition based on Employee’s Environment Satisfaction*: The graph shows that majority of the employees are satisfy with their work environment. However, employees who gave a very low rate of environment satisfaction tend to be more motivated to leave.



***Figure 6:*** *Attrition based on Employee’s Job level*: The chart shows that as the job level increases the number of employees leaving the organization decreases.

### Appendix B

### Descriptive Statistics textual for business model

## Attrition Age BusinessTravel  
## No :1233 Min. :18.00 Non-Travel : 150   
## Yes: 237 1st Qu.:30.00 Travel\_Frequently: 277   
## Median :36.00 Travel\_Rarely :1043   
## Mean :36.92   
## 3rd Qu.:43.00   
## Max. :60.00   
## Department DistanceFromHome Education   
## Human Resources : 63 Min. : 1.000 Min. :1.000   
## Research & Development:961 1st Qu.: 2.000 1st Qu.:2.000   
## Sales :446 Median : 7.000 Median :3.000   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
## EnvironmentSatisfaction Gender HourlyRate JobLevel   
## Min. :1.000 Female:588 Min. : 30.00 Min. :1.000   
## 1st Qu.:2.000 Male :882 1st Qu.: 48.00 1st Qu.:1.000   
## Median :3.000 Median : 66.00 Median :2.000   
## Mean :2.722 Mean : 65.89 Mean :2.064   
## 3rd Qu.:4.000 3rd Qu.: 83.75 3rd Qu.:3.000   
## Max. :4.000 Max. :100.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome OverTime   
## Min. :1.000 Divorced:327 Min. : 1009 No :1054   
## 1st Qu.:2.000 Married :673 1st Qu.: 2911 Yes: 416   
## Median :3.000 Single :470 Median : 4919   
## Mean :2.729 Mean : 6503   
## 3rd Qu.:4.000 3rd Qu.: 8379   
## Max. :4.000 Max. :19999   
## PerformanceRating RelationshipSatisfaction TotalWorkingYears  
## Min. :3.000 Min. :1.000 Min. : 0.00   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.: 6.00   
## Median :3.000 Median :3.000 Median :10.00   
## Mean :3.154 Mean :2.712 Mean :11.28   
## 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:15.00   
## Max. :4.000 Max. :4.000 Max. :40.00   
## WorkLifeBalance YearsSinceLastPromotion YearsWithCurrManager  
## Min. :1.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.: 0.000 1st Qu.: 2.000   
## Median :3.000 Median : 1.000 Median : 3.000   
## Mean :2.761 Mean : 2.188 Mean : 4.123   
## 3rd Qu.:3.000 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :4.000 Max. :15.000 Max. :17.000

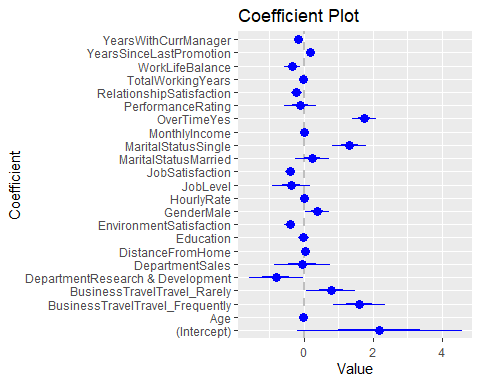
We can see that there are only a few employees who have left the company, also can see the data supports our graphs.

### Appendix C

### Full business Logistic model output, 2LL and AIC

## # A tibble: 2 x 3  
## Attrition n freq  
## <fct> <int> <dbl>  
## 1 No 1233 0.839  
## 2 Yes 237 0.161

##   
## Call:  
## glm(formula = hraReducedBusiness$Attrition ~ ., family = binomial(link = "logit"),   
## data = hraReducedBusiness)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8595 -0.5377 -0.3163 -0.1352 3.6683   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 2.191e+00 1.189e+00 1.843 0.065265  
## Age -2.574e-02 1.266e-02 -2.034 0.041941  
## BusinessTravelTravel\_Frequently 1.599e+00 3.789e-01 4.221 2.43e-05  
## BusinessTravelTravel\_Rarely 7.834e-01 3.525e-01 2.223 0.026249  
## DepartmentResearch & Development -8.109e-01 3.912e-01 -2.073 0.038157  
## DepartmentSales -4.486e-02 4.020e-01 -0.112 0.911149  
## DistanceFromHome 3.702e-02 9.944e-03 3.722 0.000197  
## Education -7.328e-03 8.144e-02 -0.090 0.928305  
## EnvironmentSatisfaction -4.026e-01 7.750e-02 -5.195 2.05e-07  
## GenderMale 3.888e-01 1.723e-01 2.257 0.024030  
## HourlyRate -9.372e-04 4.097e-03 -0.229 0.819059  
## JobLevel -3.689e-01 2.700e-01 -1.366 0.171855  
## JobSatisfaction -4.040e-01 7.546e-02 -5.354 8.59e-08  
## MaritalStatusMarried 2.397e-01 2.431e-01 0.986 0.324228  
## MaritalStatusSingle 1.315e+00 2.448e-01 5.374 7.69e-08  
## MonthlyIncome -2.602e-05 6.481e-05 -0.401 0.688053  
## OverTimeYes 1.747e+00 1.761e-01 9.922 < 2e-16  
## PerformanceRating -1.053e-01 2.330e-01 -0.452 0.651393  
## RelationshipSatisfaction -2.158e-01 7.605e-02 -2.838 0.004537  
## TotalWorkingYears -1.149e-02 2.461e-02 -0.467 0.640662  
## WorkLifeBalance -3.364e-01 1.155e-01 -2.912 0.003591  
## YearsSinceLastPromotion 1.697e-01 3.555e-02 4.774 1.81e-06  
## YearsWithCurrManager -1.731e-01 3.551e-02 -4.875 1.09e-06  
##   
## (Intercept) .   
## Age \*   
## BusinessTravelTravel\_Frequently \*\*\*  
## BusinessTravelTravel\_Rarely \*   
## DepartmentResearch & Development \*   
## DepartmentSales   
## DistanceFromHome \*\*\*  
## Education   
## EnvironmentSatisfaction \*\*\*  
## GenderMale \*   
## HourlyRate   
## JobLevel   
## JobSatisfaction \*\*\*  
## MaritalStatusMarried   
## MaritalStatusSingle \*\*\*  
## MonthlyIncome   
## OverTimeYes \*\*\*  
## PerformanceRating   
## RelationshipSatisfaction \*\*   
## TotalWorkingYears   
## WorkLifeBalance \*\*   
## YearsSinceLastPromotion \*\*\*  
## YearsWithCurrManager \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 961.51 on 1447 degrees of freedom  
## AIC: 1007.5  
##   
## Number of Fisher Scoring iterations: 6



## 'log Lik.' 961.5126 (df=23)

## [1] 961.5126

## [1] 1007.513

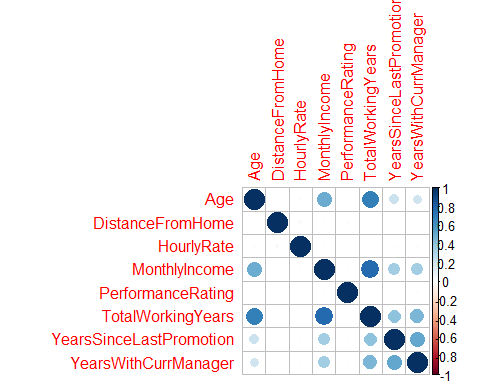
## log.odds odds prob  
## (Intercept) 2.191411e+00 8.9478295 0.8994756  
## Age -2.574470e-02 0.9745839 0.4935642  
## BusinessTravelTravel\_Frequently 1.599379e+00 4.9499596 0.8319316  
## BusinessTravelTravel\_Rarely 7.833953e-01 2.1888916 0.6864114  
## DepartmentResearch & Development -8.109360e-01 0.4444419 0.3076911  
## DepartmentSales -4.486049e-02 0.9561309 0.4887868  
## DistanceFromHome 3.701588e-02 1.0377095 0.5092529  
## Education -7.328053e-03 0.9926987 0.4981680  
## EnvironmentSatisfaction -4.025770e-01 0.6685948 0.4006933  
## GenderMale 3.887894e-01 1.4751938 0.5959912  
## HourlyRate -9.371729e-04 0.9990633 0.4997657  
## JobLevel -3.688814e-01 0.6915074 0.4088114  
## JobSatisfaction -4.040082e-01 0.6676386 0.4003497  
## MaritalStatusMarried 2.396903e-01 1.2708556 0.5596373  
## MaritalStatusSingle 1.315390e+00 3.7262056 0.7884138  
## MonthlyIncome -2.602077e-05 0.9999740 0.4999935  
## OverTimeYes 1.747048e+00 5.7376418 0.8515801  
## PerformanceRating -1.052954e-01 0.9000586 0.4737004  
## RelationshipSatisfaction -2.158391e-01 0.8058650 0.4462487  
## TotalWorkingYears -1.148561e-02 0.9885801 0.4971286  
## WorkLifeBalance -3.363955e-01 0.7143405 0.4166853  
## YearsSinceLastPromotion 1.697208e-01 1.1849740 0.5423287  
## YearsWithCurrManager -1.730845e-01 0.8410666 0.4568366

### Appendix D

### Correlation Matrix for quantitative variables

## Age DistanceFromHome HourlyRate  
## Age 1.000000000 -0.001686120 0.024286543  
## DistanceFromHome -0.001686120 1.000000000 0.031130586  
## HourlyRate 0.024286543 0.031130586 1.000000000  
## MonthlyIncome 0.497854567 -0.017014445 -0.015794304  
## PerformanceRating 0.001903896 0.027109618 -0.002171697  
## TotalWorkingYears 0.680380536 0.004628426 -0.002333682  
## YearsSinceLastPromotion 0.216513368 0.010028836 -0.026715586  
## YearsWithCurrManager 0.202088602 0.014406048 -0.020123200  
## MonthlyIncome PerformanceRating TotalWorkingYears  
## Age 0.49785457 0.001903896 0.680380536  
## DistanceFromHome -0.01701444 0.027109618 0.004628426  
## HourlyRate -0.01579430 -0.002171697 -0.002333682  
## MonthlyIncome 1.00000000 -0.017120138 0.772893246  
## PerformanceRating -0.01712014 1.000000000 0.006743668  
## TotalWorkingYears 0.77289325 0.006743668 1.000000000  
## YearsSinceLastPromotion 0.34497764 0.017896066 0.404857759  
## YearsWithCurrManager 0.34407888 0.022827169 0.459188397  
## YearsSinceLastPromotion YearsWithCurrManager  
## Age 0.21651337 0.20208860  
## DistanceFromHome 0.01002884 0.01440605  
## HourlyRate -0.02671559 -0.02012320  
## MonthlyIncome 0.34497764 0.34407888  
## PerformanceRating 0.01789607 0.02282717  
## TotalWorkingYears 0.40485776 0.45918840  
## YearsSinceLastPromotion 1.00000000 0.51022364  
## YearsWithCurrManager 0.51022364 1.00000000

## Age DistanceFromHome HourlyRate MonthlyIncome  
## Age 1.00 0.00 0.02 0.50  
## DistanceFromHome 0.00 1.00 0.03 -0.02  
## HourlyRate 0.02 0.03 1.00 -0.02  
## MonthlyIncome 0.50 -0.02 -0.02 1.00  
## PerformanceRating 0.00 0.03 0.00 -0.02  
## TotalWorkingYears 0.68 0.00 0.00 0.77  
## YearsSinceLastPromotion 0.22 0.01 -0.03 0.34  
## YearsWithCurrManager 0.20 0.01 -0.02 0.34  
## PerformanceRating TotalWorkingYears  
## Age 0.00 0.68  
## DistanceFromHome 0.03 0.00  
## HourlyRate 0.00 0.00  
## MonthlyIncome -0.02 0.77  
## PerformanceRating 1.00 0.01  
## TotalWorkingYears 0.01 1.00  
## YearsSinceLastPromotion 0.02 0.40  
## YearsWithCurrManager 0.02 0.46  
## YearsSinceLastPromotion YearsWithCurrManager  
## Age 0.22 0.20  
## DistanceFromHome 0.01 0.01  
## HourlyRate -0.03 -0.02  
## MonthlyIncome 0.34 0.34  
## PerformanceRating 0.02 0.02  
## TotalWorkingYears 0.40 0.46  
## YearsSinceLastPromotion 1.00 0.51  
## YearsWithCurrManager 0.51 1.00  
##   
## n= 1470   
##   
##   
## P  
## Age DistanceFromHome HourlyRate MonthlyIncome  
## Age 0.9485 0.3521 0.0000   
## DistanceFromHome 0.9485 0.2329 0.5145   
## HourlyRate 0.3521 0.2329 0.5451   
## MonthlyIncome 0.0000 0.5145 0.5451   
## PerformanceRating 0.9419 0.2989 0.9337 0.5119   
## TotalWorkingYears 0.0000 0.8593 0.9288 0.0000   
## YearsSinceLastPromotion 0.0000 0.7008 0.3060 0.0000   
## YearsWithCurrManager 0.0000 0.5810 0.4407 0.0000   
## PerformanceRating TotalWorkingYears  
## Age 0.9419 0.0000   
## DistanceFromHome 0.2989 0.8593   
## HourlyRate 0.9337 0.9288   
## MonthlyIncome 0.5119 0.0000   
## PerformanceRating 0.7961   
## TotalWorkingYears 0.7961   
## YearsSinceLastPromotion 0.4930 0.0000   
## YearsWithCurrManager 0.3818 0.0000   
## YearsSinceLastPromotion YearsWithCurrManager  
## Age 0.0000 0.0000   
## DistanceFromHome 0.7008 0.5810   
## HourlyRate 0.3060 0.4407   
## MonthlyIncome 0.0000 0.0000   
## PerformanceRating 0.4930 0.3818   
## TotalWorkingYears 0.0000 0.0000   
## YearsSinceLastPromotion 0.0000   
## YearsWithCurrManager 0.0000



### Chi Square test for categorical correlation:

##   
## Human Resources Research & Development Sales  
## Non-Travel 6 97 47  
## Travel\_Frequently 11 182 84  
## Travel\_Rarely 46 682 315

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 0.20189, df = 4, p-value = 0.9952

##   
## Female Male  
## Non-Travel 49 101  
## Travel\_Frequently 117 160  
## Travel\_Rarely 422 621

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 4.0314, df = 2, p-value = 0.1332

##   
## Human Resources Research & Development Sales  
## Female 20 379 189  
## Male 43 582 257

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 2.9645, df = 2, p-value = 0.2271

##   
## No Yes  
## Divorced 228 99  
## Married 487 186  
## Single 339 131

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 0.81672, df = 2, p-value = 0.6647

##   
## Human Resources Research & Development Sales  
## Divorced 16 224 87  
## Married 34 433 206  
## Single 13 304 153

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 6.6485, df = 4, p-value = 0.1557

##   
## Divorced Married Single  
## Non-Travel 44 59 47  
## Travel\_Frequently 63 118 96  
## Travel\_Rarely 220 496 327

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 7.5021, df = 4, p-value = 0.1116

##   
## Human Resources Research & Development Sales  
## No 46 690 318  
## Yes 17 271 128

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 0.093607, df = 2, p-value = 0.9543

##   
## No Yes  
## Non-Travel 115 35  
## Travel\_Frequently 191 86  
## Travel\_Rarely 748 295

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 2.8538, df = 2, p-value = 0.2401

##   
## No Yes  
## Female 408 180  
## Male 646 236

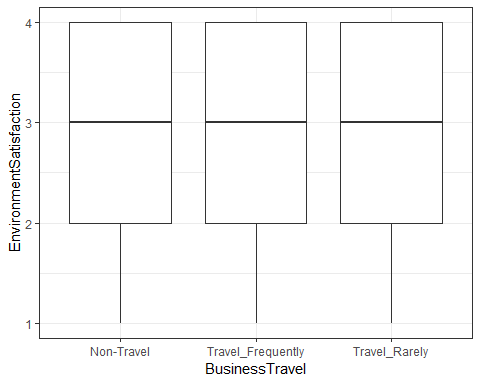
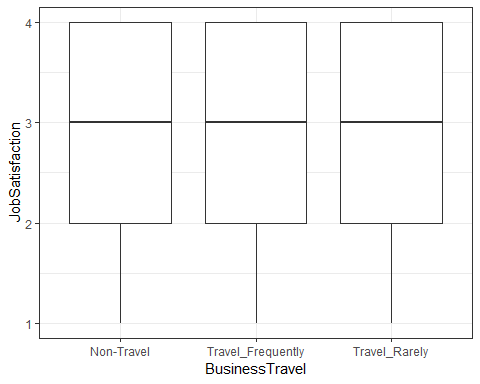
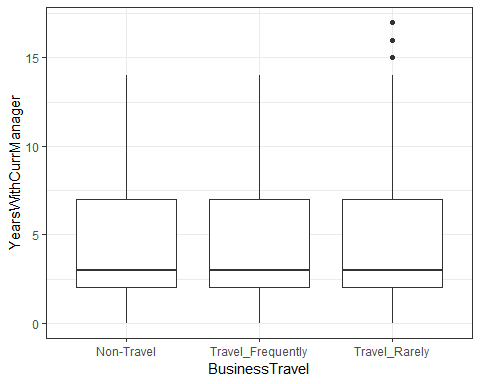
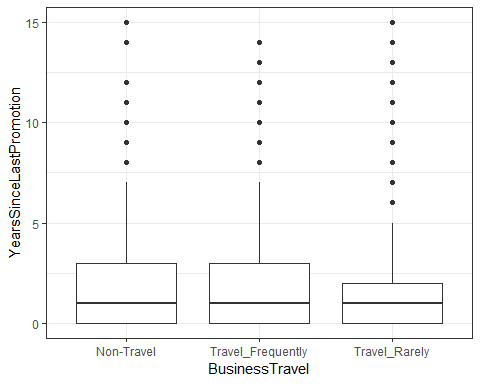
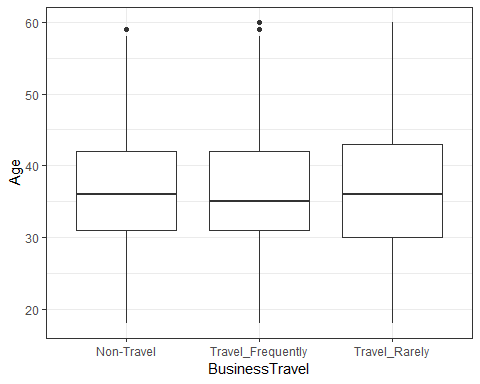
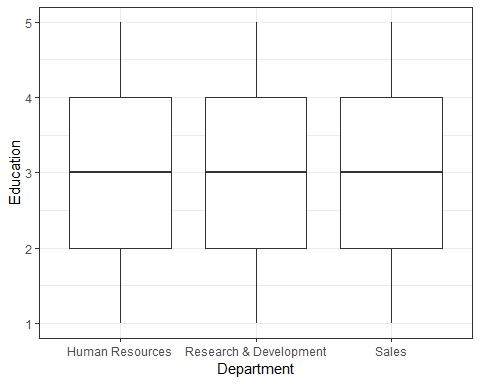
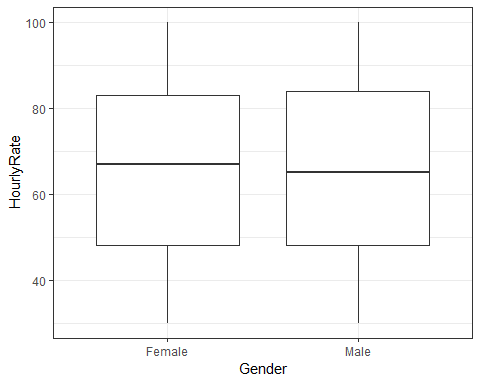
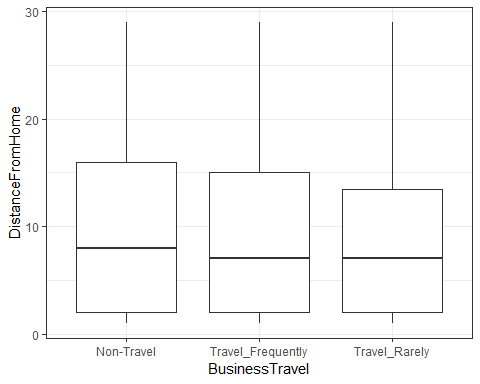
##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tbl  
## X-squared = 2.3973, df = 1, p-value = 0.1215

##   
## Divorced Married Single  
## Female 117 272 199  
## Male 210 401 271

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 3.5478, df = 2, p-value = 0.1697

### Appendix E

### Box Plots for quant vs Categorical variables



### 

### Appendix F

### Stepwise Results

##   
## Call:  
## glm(formula = hrareducedcor$Attrition ~ Age + BusinessTravel +   
## Department + DistanceFromHome + EnvironmentSatisfaction +   
## Gender + JobLevel + JobSatisfaction + MaritalStatus + OverTime +   
## RelationshipSatisfaction + WorkLifeBalance + YearsSinceLastPromotion +   
## YearsWithCurrManager, family = binomial(link = "logit"),   
## data = hraReducedBusiness)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8490 -0.5423 -0.3176 -0.1372 3.6214   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.890526 0.823744 2.295 0.021731 \*   
## Age -0.029413 0.010772 -2.731 0.006322 \*\*   
## BusinessTravelTravel\_Frequently 1.597987 0.378674 4.220 2.44e-05 \*\*\*  
## BusinessTravelTravel\_Rarely 0.778156 0.352708 2.206 0.027368 \*   
## DepartmentResearch & Development -0.818595 0.389634 -2.101 0.035647 \*   
## DepartmentSales -0.024956 0.399141 -0.063 0.950145   
## DistanceFromHome 0.037113 0.009899 3.749 0.000177 \*\*\*  
## EnvironmentSatisfaction -0.398428 0.077260 -5.157 2.51e-07 \*\*\*  
## GenderMale 0.390383 0.172063 2.269 0.023278 \*   
## JobLevel -0.503615 0.118335 -4.256 2.08e-05 \*\*\*  
## JobSatisfaction -0.399869 0.075089 -5.325 1.01e-07 \*\*\*  
## MaritalStatusMarried 0.241072 0.242294 0.995 0.319757   
## MaritalStatusSingle 1.315500 0.244577 5.379 7.50e-08 \*\*\*  
## OverTimeYes 1.736649 0.175298 9.907 < 2e-16 \*\*\*  
## RelationshipSatisfaction -0.213294 0.075912 -2.810 0.004958 \*\*   
## WorkLifeBalance -0.334912 0.115504 -2.900 0.003737 \*\*   
## YearsSinceLastPromotion 0.165925 0.034793 4.769 1.85e-06 \*\*\*  
## YearsWithCurrManager -0.176818 0.034238 -5.164 2.41e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 962.21 on 1452 degrees of freedom  
## AIC: 998.21  
##   
## Number of Fisher Scoring iterations: 6

**The model with the lowest AIC shows that totalworkinghours, monthlyincome, performancerating, hourlyrate, and education are not significant and AIC has been reduced to 998.21 from 1007.5**

### Appendix G

### Stepwise reduced model

##   
## Call:  
## glm(formula = hrastepreduced$Attrition ~ ., family = binomial(link = "logit"),   
## data = hrastepreduced)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8490 -0.5423 -0.3176 -0.1372 3.6214   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.890526 0.823744 2.295 0.021731 \*   
## Age -0.029413 0.010772 -2.731 0.006322 \*\*   
## BusinessTravelTravel\_Frequently 1.597987 0.378674 4.220 2.44e-05 \*\*\*  
## BusinessTravelTravel\_Rarely 0.778156 0.352708 2.206 0.027368 \*   
## DepartmentResearch & Development -0.818595 0.389634 -2.101 0.035647 \*   
## DepartmentSales -0.024956 0.399141 -0.063 0.950145   
## DistanceFromHome 0.037113 0.009899 3.749 0.000177 \*\*\*  
## EnvironmentSatisfaction -0.398428 0.077260 -5.157 2.51e-07 \*\*\*  
## GenderMale 0.390383 0.172063 2.269 0.023278 \*   
## JobLevel -0.503615 0.118335 -4.256 2.08e-05 \*\*\*  
## JobSatisfaction -0.399869 0.075089 -5.325 1.01e-07 \*\*\*  
## MaritalStatusMarried 0.241072 0.242294 0.995 0.319757   
## MaritalStatusSingle 1.315500 0.244577 5.379 7.50e-08 \*\*\*  
## OverTimeYes 1.736649 0.175298 9.907 < 2e-16 \*\*\*  
## RelationshipSatisfaction -0.213294 0.075912 -2.810 0.004958 \*\*   
## WorkLifeBalance -0.334912 0.115504 -2.900 0.003737 \*\*   
## YearsSinceLastPromotion 0.165925 0.034793 4.769 1.85e-06 \*\*\*  
## YearsWithCurrManager -0.176818 0.034238 -5.164 2.41e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 962.21 on 1452 degrees of freedom  
## AIC: 998.21  
##   
## Number of Fisher Scoring iterations: 6

### Appendix H

### Comparison of 2LL and AIC:

## 'log Lik.' 961.5126 (df=23)

## [1] 961.5126

## [1] 1007.513

## 'log Lik.' 962.2132 (df=18)

## [1] 962.2132

## [1] 998.2132

### Appendix I

### DW Test:

##   
## Durbin-Watson test  
##   
## data: hra.logit  
## DW = 1.9045, p-value = 0.03356  
## alternative hypothesis: true autocorrelation is greater than 0

##   
## Durbin-Watson test  
##   
## data: hra.logit.final.red  
## DW = 1.9051, p-value = 0.03433  
## alternative hypothesis: true autocorrelation is greater than 0

### VIF for business and final reduced model:

## GVIF Df GVIF^(1/(2\*Df))  
## Age 1.798273 1 1.340997  
## BusinessTravel 1.068523 2 1.016707  
## Department 1.199726 2 1.046575  
## DistanceFromHome 1.049930 1 1.024661  
## Education 1.066579 1 1.032753  
## EnvironmentSatisfaction 1.069392 1 1.034114  
## Gender 1.025979 1 1.012906  
## HourlyRate 1.029275 1 1.014532  
## JobLevel 8.121260 1 2.849782  
## JobSatisfaction 1.060399 1 1.029757  
## MaritalStatus 1.099333 2 1.023958  
## MonthlyIncome 7.371458 1 2.715043  
## OverTime 1.132810 1 1.064335  
## PerformanceRating 1.016270 1 1.008102  
## RelationshipSatisfaction 1.038892 1 1.019260  
## TotalWorkingYears 3.595274 1 1.896121  
## WorkLifeBalance 1.033935 1 1.016826  
## YearsSinceLastPromotion 1.781397 1 1.334690  
## YearsWithCurrManager 1.707506 1 1.306716

## GVIF Df GVIF^(1/(2\*Df))  
## Age 1.305834 1 1.142731  
## BusinessTravel 1.059084 2 1.014454  
## Department 1.136996 2 1.032618  
## DistanceFromHome 1.042746 1 1.021149  
## EnvironmentSatisfaction 1.063621 1 1.031320  
## Gender 1.023543 1 1.011703  
## JobLevel 1.588919 1 1.260523  
## JobSatisfaction 1.051758 1 1.025553  
## MaritalStatus 1.090662 2 1.021933  
## OverTime 1.123190 1 1.059807  
## RelationshipSatisfaction 1.032986 1 1.016359  
## WorkLifeBalance 1.032246 1 1.015995  
## YearsSinceLastPromotion 1.722801 1 1.312555  
## YearsWithCurrManager 1.593121 1 1.262189

### Appendix J

### Cross validation for Business modeL:

## Actual  
## Predicted No Yes  
## No 347 55  
## Yes 15 24

## Accuracy Rate Error Rate Sensitivity Specificity   
## 0.841 0.159 0.304 0.959   
## False Positives   
## 0.041

## 5 9 14 15 16 17 27 29 30 31   
## 0 0 0 1 0 1 1 0 0 0

## Actual  
## Predicted No Yes  
## No 285 18  
## Yes 77 61

## 5 9 14 15 16 17 27 29 30 31   
## 0 0 0 1 0 0 1 0 0 0

## Actual  
## Predicted No Yes  
## No 317 32  
## Yes 45 47

## Accuracy Rate Error Rate Sensitivity Specificity   
## 0.83 0.17 0.59 0.88   
## False Positives   
## 0.12

## Accuracy Rate Error Rate Sensitivity Specificity  
## logit.rates.50 0.84 0.16 0.30 0.96  
## logit.rates.20 0.78 0.22 0.77 0.79  
## logit.rates.30 0.83 0.17 0.59 0.88  
## False Positives  
## logit.rates.50 0.041  
## logit.rates.20 0.213  
## logit.rates.30 0.124

## 5 9 14 15 16 17 27 29 30 31   
## 0 0 0 1 0 1 1 0 0 0

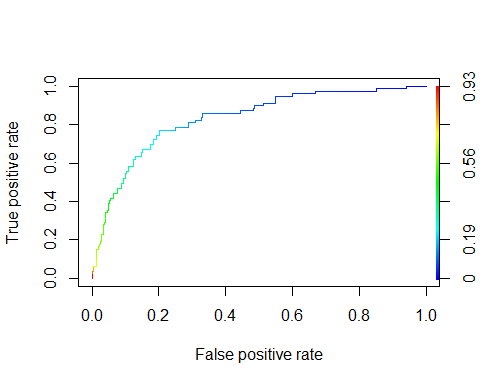
## Actual  
## Predicted No Yes  
## No 309 29  
## Yes 53 50

## Accuracy Rate Error Rate Sensitivity Specificity   
## 0.81 0.19 0.63 0.85   
## False Positives   
## 0.15

## [1] "Final Comparision of the four models:"

## Accuracy Rate Error Rate Sensitivity Specificity  
## logit.rates.50 0.84 0.16 0.30 0.96  
## logit.rates.20 0.78 0.22 0.77 0.79  
## logit.rates.30 0.83 0.17 0.59 0.88  
## logit.rates.275 0.81 0.19 0.63 0.85  
## False Positives  
## logit.rates.50 0.041  
## logit.rates.20 0.213  
## logit.rates.30 0.124  
## logit.rates.275 0.146

### ROC curve for business model:



## [1] "Area under the ROC curve" "0.832086159871317"

### Appendix K

### Cross Validation for final reduced model:

## Actual  
## Predicted No Yes  
## No 349 53  
## Yes 13 26



## Accuracy Rate Error Rate Sensitivity Specificity  
## logit.rates.50.red 0.85 0.15 0.33 0.96  
## logit.rates.50 0.84 0.16 0.30 0.96  
## logit.rates.20 0.78 0.22 0.77 0.79  
## logit.rates.30 0.83 0.17 0.59 0.88  
## logit.rates.275 0.81 0.19 0.63 0.85  
## False Positives  
## logit.rates.50.red 0.036  
## logit.rates.50 0.041  
## logit.rates.20 0.213  
## logit.rates.30 0.124  
## logit.rates.275 0.146

## Actual  
## Predicted No Yes  
## No 286 19  
## Yes 76 60

## Actual  
## Predicted No Yes  
## No 316 30  
## Yes 46 49

## [1] 0.8276644

## [1] 0.1270718

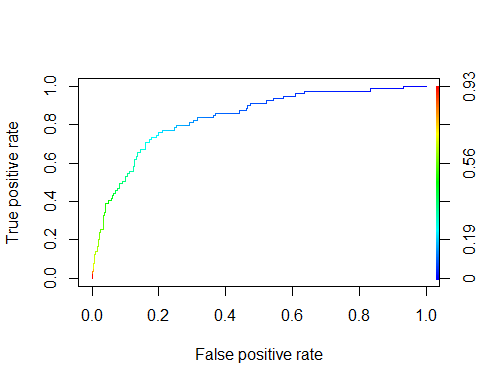
## Actual  
## Predicted No Yes  
## No 310 27  
## Yes 52 52

## [1] "Final Comparision of the four models:"

## Accuracy Rate Error Rate Sensitivity Specificity  
## logit.rates.50.red 0.85 0.15 0.33 0.96  
## logit.rates.20.red 0.78 0.22 0.76 0.79  
## logit.rates.30.red 0.83 0.17 0.62 0.87  
## logit.rates.275.red 0.82 0.18 0.66 0.86  
## False Positives  
## logit.rates.50.red 0.036  
## logit.rates.20.red 0.210  
## logit.rates.30.red 0.127  
## logit.rates.275.red 0.144

## Accuracy Rate Error Rate Sensitivity Specificity  
## logit.rates.50.red 0.85 0.15 0.33 0.96  
## logit.rates.20.red 0.78 0.22 0.76 0.79  
## logit.rates.30.red 0.83 0.17 0.62 0.87  
## logit.rates.275.red 0.82 0.18 0.66 0.86  
## logit.rates.50 0.84 0.16 0.30 0.96  
## logit.rates.20 0.78 0.22 0.77 0.79  
## logit.rates.30 0.83 0.17 0.59 0.88  
## logit.rates.275 0.81 0.19 0.63 0.85  
## False Positives  
## logit.rates.50.red 0.036  
## logit.rates.20.red 0.210  
## logit.rates.30.red 0.127  
## logit.rates.275.red 0.144  
## logit.rates.50 0.041  
## logit.rates.20 0.213  
## logit.rates.30 0.124  
## logit.rates.275 0.146

### ROC Curve for final reduced model:

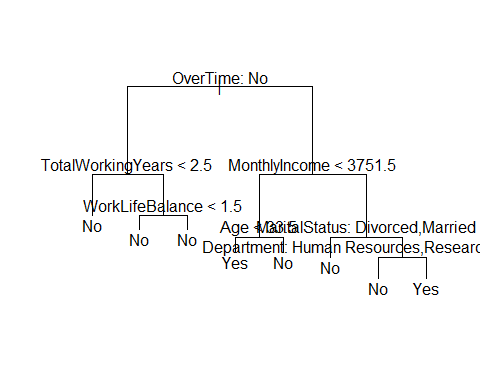


## [1] "Area under the ROC curve" "0.83635219246101"

### Appendix L

### Classification Tree on the business reduced model

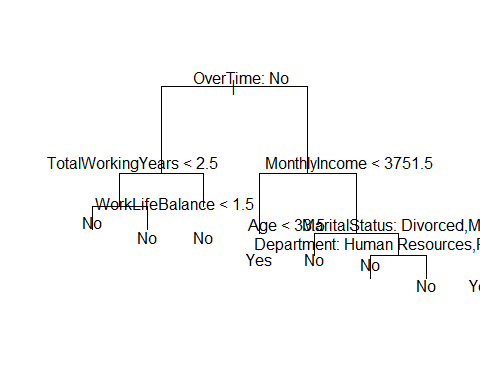
##   
## Classification tree:  
## tree(formula = Attrition ~ ., data = hraReducedBusiness)  
## Variables actually used in tree construction:  
## [1] "OverTime" "TotalWorkingYears" "WorkLifeBalance"   
## [4] "MonthlyIncome" "Age" "MaritalStatus"   
## [7] "Department"   
## Number of terminal nodes: 8   
## Residual mean deviance: 0.7234 = 1058 / 1462   
## Misclassification error rate: 0.1367 = 201 / 1470



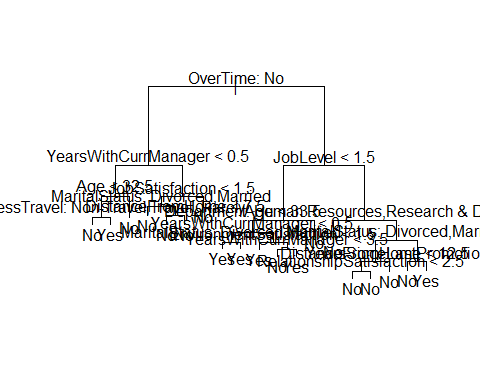
## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 1470 1299.00 No ( 0.83878 0.16122 )   
## 2) OverTime: No 1054 705.30 No ( 0.89564 0.10436 )   
## 4) TotalWorkingYears < 2.5 88 111.60 No ( 0.67045 0.32955 ) \*  
## 5) TotalWorkingYears > 2.5 966 556.60 No ( 0.91615 0.08385 )   
## 10) WorkLifeBalance < 1.5 53 59.05 No ( 0.75472 0.24528 ) \*  
## 11) WorkLifeBalance > 1.5 913 484.00 No ( 0.92552 0.07448 ) \*  
## 3) OverTime: Yes 416 511.90 No ( 0.69471 0.30529 )   
## 6) MonthlyIncome < 3751.5 143 197.10 Yes ( 0.45455 0.54545 )   
## 12) Age < 33.5 85 107.70 Yes ( 0.32941 0.67059 ) \*  
## 13) Age > 33.5 58 75.93 No ( 0.63793 0.36207 ) \*  
## 7) MonthlyIncome > 3751.5 273 257.00 No ( 0.82051 0.17949 )   
## 14) MaritalStatus: Divorced,Married 194 137.20 No ( 0.88660 0.11340 ) \*  
## 15) MaritalStatus: Single 79 101.50 No ( 0.65823 0.34177 )   
## 30) Department: Human Resources,Research & Development 44 35.05 No ( 0.86364 0.13636 ) \*  
## 31) Department: Sales 35 47.11 Yes ( 0.40000 0.60000 ) \*

### Classification Tree on the stepwise reduced model:

##   
## Classification tree:  
## tree(formula = Attrition ~ ., data = hrastepreduced)  
## Variables actually used in tree construction:  
## [1] "OverTime" "YearsWithCurrManager" "Age"   
## [4] "JobLevel" "Department" "MaritalStatus"   
## Number of terminal nodes: 7   
## Residual mean deviance: 0.7285 = 1066 / 1463   
## Misclassification error rate: 0.151 = 222 / 1470



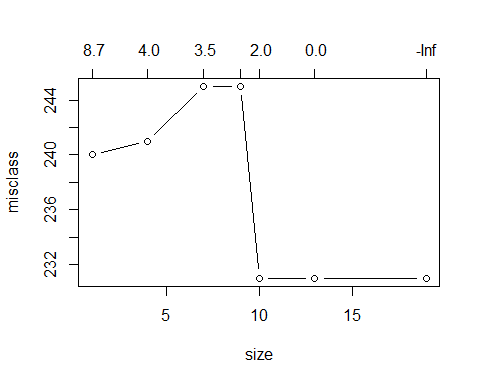
## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 1470 1299.00 No ( 0.83878 0.16122 )   
## 2) OverTime: No 1054 705.30 No ( 0.89564 0.10436 )   
## 4) YearsWithCurrManager < 0.5 181 196.10 No ( 0.76796 0.23204 )   
## 8) Age < 32.5 78 104.80 No ( 0.60256 0.39744 ) \*  
## 9) Age > 32.5 103 69.99 No ( 0.89320 0.10680 ) \*  
## 5) YearsWithCurrManager > 0.5 873 477.70 No ( 0.92211 0.07789 ) \*  
## 3) OverTime: Yes 416 511.90 No ( 0.69471 0.30529 )   
## 6) JobLevel < 1.5 156 215.90 Yes ( 0.47436 0.52564 ) \*  
## 7) JobLevel > 1.5 260 239.60 No ( 0.82692 0.17308 )   
## 14) Department: Human Resources,Research & Development 154 89.14 No ( 0.91558 0.08442 ) \*  
## 15) Department: Sales 106 129.80 No ( 0.69811 0.30189 )   
## 30) MaritalStatus: Divorced,Married 71 61.23 No ( 0.84507 0.15493 ) \*  
## 31) MaritalStatus: Single 35 47.11 Yes ( 0.40000 0.60000 ) \*



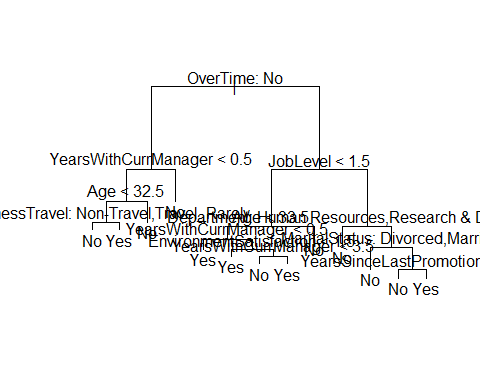
### Cross validate the classification tree:

## $size  
## [1] 19 13 10 9 7 4 1  
##   
## $dev  
## [1] 231 231 231 245 245 241 240  
##   
## $k  
## [1] -Inf 0.000000 2.000000 3.000000 3.500000 4.000000 8.666667  
##   
## $method  
## [1] "misclass"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

## Tree Size Misclass  
## [1,] 19 231  
## [2,] 13 231  
## [3,] 10 231  
## [4,] 9 245  
## [5,] 7 245  
## [6,] 4 241  
## [7,] 1 240



##   
## Classification tree:  
## snip.tree(tree = hra.tree.step.large, nodes = c(5L, 24L, 30L))  
## Variables actually used in tree construction:  
## [1] "OverTime" "YearsWithCurrManager"   
## [3] "Age" "BusinessTravel"   
## [5] "JobLevel" "EnvironmentSatisfaction"  
## [7] "Department" "MaritalStatus"   
## [9] "YearsSinceLastPromotion"  
## Number of terminal nodes: 13   
## Residual mean deviance: 0.6956 = 1014 / 1457   
## Misclassification error rate: 0.1245 = 183 / 1470

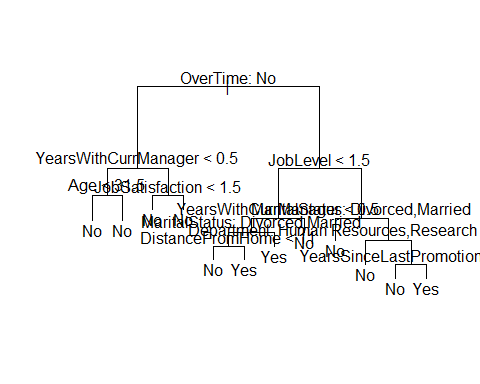


### Appendix M

### Cross validation using subset sampling:

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 1029 932.10 No ( 0.83188 0.16812 )   
## 2) OverTime: No 735 501.60 No ( 0.89252 0.10748 )   
## 4) YearsWithCurrManager < 0.5 124 134.90 No ( 0.76613 0.23387 )   
## 8) Age < 31.5 51 69.74 No ( 0.56863 0.43137 ) \*  
## 9) Age > 31.5 73 46.13 No ( 0.90411 0.09589 ) \*  
## 5) YearsWithCurrManager > 0.5 611 346.10 No ( 0.91817 0.08183 )   
## 10) JobSatisfaction < 1.5 117 103.80 No ( 0.83761 0.16239 ) \*  
## 11) JobSatisfaction > 1.5 494 231.70 No ( 0.93725 0.06275 ) \*  
## 3) OverTime: Yes 294 368.50 No ( 0.68027 0.31973 )   
## 6) JobLevel < 1.5 112 154.70 Yes ( 0.46429 0.53571 )   
## 12) YearsWithCurrManager < 0.5 36 40.49 Yes ( 0.25000 0.75000 )   
## 24) MaritalStatus: Divorced,Married 22 29.77 Yes ( 0.40909 0.59091 )   
## 48) DistanceFromHome < 11 15 20.19 No ( 0.60000 0.40000 ) \*  
## 49) DistanceFromHome > 11 7 0.00 Yes ( 0.00000 1.00000 ) \*  
## 25) MaritalStatus: Single 14 0.00 Yes ( 0.00000 1.00000 ) \*  
## 13) YearsWithCurrManager > 0.5 76 104.00 No ( 0.56579 0.43421 ) \*  
## 7) JobLevel > 1.5 182 175.30 No ( 0.81319 0.18681 )   
## 14) MaritalStatus: Divorced,Married 129 88.60 No ( 0.89147 0.10853 ) \*  
## 15) MaritalStatus: Single 53 70.25 No ( 0.62264 0.37736 )   
## 30) Department: Human Resources,Research & Development 30 23.56 No ( 0.86667 0.13333 ) \*  
## 31) Department: Sales 23 28.27 Yes ( 0.30435 0.69565 )   
## 62) YearsSinceLastPromotion < 1.5 13 17.94 No ( 0.53846 0.46154 ) \*  
## 63) YearsSinceLastPromotion > 1.5 10 0.00 Yes ( 0.00000 1.00000 ) \*

##   
## Classification tree:  
## tree(formula = Attrition ~ ., data = hrastepreduced[train, ])  
## Variables actually used in tree construction:  
## [1] "OverTime" "YearsWithCurrManager"   
## [3] "Age" "JobSatisfaction"   
## [5] "JobLevel" "MaritalStatus"   
## [7] "DistanceFromHome" "Department"   
## [9] "YearsSinceLastPromotion"  
## Number of terminal nodes: 12   
## Residual mean deviance: 0.6939 = 705.7 / 1017   
## Misclassification error rate: 0.138 = 142 / 1029



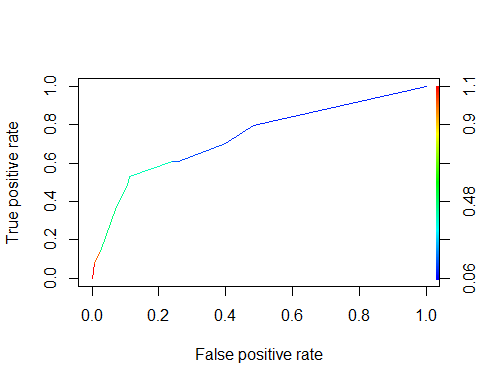
## Actual  
## Predicted No Yes  
## No 374 59  
## Yes 3 5

##   
## hra.tree.pred.class.275 No Yes  
## 0 335 30  
## 1 42 34

## Accuracy Rate Error Rate Sensitivity Specificity   
## 0.8367347 0.1632653 0.5312500 0.8885942   
## False Positives   
## 0.1114058

## Accuracy Rate Error Rate Sensitivity Specificity  
## tree.rates.50 0.8594104 0.1405896 0.0781250 0.9920424  
## tree.rates.275 0.8367347 0.1632653 0.5312500 0.8885942  
## logit.rates.50.red 0.8503401 0.1496599 0.3291139 0.9640884  
## logit.rates.20.red 0.7845805 0.2154195 0.7594937 0.7900552  
## logit.rates.30.red 0.8276644 0.1723356 0.6202532 0.8729282  
## logit.rates.275.red 0.8208617 0.1791383 0.6582278 0.8563536  
## logit.rates.50 0.8412698 0.1587302 0.3037975 0.9585635  
## logit.rates.20 0.7845805 0.2154195 0.7721519 0.7872928  
## logit.rates.30 0.8253968 0.1746032 0.5949367 0.8756906  
## logit.rates.275 0.8140590 0.1859410 0.6329114 0.8535912  
## False Positives  
## tree.rates.50 0.00795756  
## tree.rates.275 0.11140584  
## logit.rates.50.red 0.03591160  
## logit.rates.20.red 0.20994475  
## logit.rates.30.red 0.12707182  
## logit.rates.275.red 0.14364641  
## logit.rates.50 0.04143646  
## logit.rates.20 0.21270718  
## logit.rates.30 0.12430939  
## logit.rates.275 0.14640884

### ROC curve for our tree model:



## [1] "Area under the ROC curve" "0.735949933687003"

### Appendix N

### Results based on logistic model:

## [1] "Final porion of employees leaving: "

## [1] 20.06803