Load the files: the new employees data and the hays data

data <- read.csv(file = "employees2.csv", fileEncoding="UTF-8-BOM")  
income\_data <- read.csv(file = "hays\_salary\_survey.csv")[-c(2,3,4)] #use post tax figures

map the industry income ranges to our data

lower <- income\_data[match(data$JobRole,income\_data$JobRole),][2]  
upper <- income\_data[match(data$JobRole,income\_data$JobRole),][3]  
  
#classify the monthly income  
data$IndustryRange <- ifelse(  
data$MonthlyIncome < lower,  
"Below Range",  
ifelse(  
data$MonthlyIncome >= lower & data$MonthlyIncome< upper,  
"Within Range",  
"Above Range")  
)

Enrich the data

data$IsAttrition <- ifelse(data$Attrition == "Yes",1,0)   
  
data$IsManagement <- ifelse(  
 data$JobRole == c("Manager", "Manufacturing Director",   
 "Research Director", "Sales Executive"),  
 "Yes",  
 "No"  
)

Linear Model lmIncome

numRows <- nrow(data)  
set.seed(57) #changed from 57  
train <- sample(numRows, 0.8\*numRows)  
dataTrain <- data[train, ]  
dataTest <- data[-train,]

Linear modelling Linear.Model.01

Linear.Model.01 <- lm(  
formula = YearsAtCompany ~ .,  
data = dataTrain  
)  
summary(Linear.Model.01)

##   
## Call:  
## lm(formula = YearsAtCompany ~ ., data = dataTrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.6771 -1.5229 -0.0429 1.2417 16.7111   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.0024381 2.1725757 2.303 0.021573 \*   
## Age -0.0229391 0.0173764 -1.320 0.187187   
## AttritionYes 0.5628472 0.3368803 1.671 0.095179 .   
## BusinessTravelTravel\_Frequently -0.2278177 0.4552909 -0.500 0.616953   
## BusinessTravelTravel\_Rarely -0.2239075 0.3888446 -0.576 0.564901   
## DepartmentResearch & Development -4.2839268 1.2325737 -3.476 0.000539 \*\*\*  
## DepartmentSales -2.5996062 1.2650675 -2.055 0.040227 \*   
## DistanceFromHome -0.0185455 0.0130748 -1.418 0.156479   
## Education 0.0147073 0.1134081 0.130 0.896850   
## EnvironmentSatisfaction 0.2049397 0.1008362 2.032 0.042459 \*   
## GenderMale 0.4052025 0.2277822 1.779 0.075654 .   
## JobInvolvement -0.2922933 0.1561496 -1.872 0.061606 .   
## JobRoleHuman Resources -3.7131162 1.4518302 -2.558 0.010734 \*   
## JobRoleLaboratory Technician 0.0079029 0.5270715 0.015 0.988041   
## JobRoleManager -1.3688200 0.9888525 -1.384 0.166686   
## JobRoleManufacturing Director -0.5912016 0.5358825 -1.103 0.270275   
## JobRoleResearch Director -1.3760249 0.8855216 -1.554 0.120619   
## JobRoleResearch Scientist 0.0860432 0.5326797 0.162 0.871720   
## JobRoleSales Executive -2.0777516 0.9604990 -2.163 0.030836 \*   
## JobRoleSales Representative -1.6568656 1.0808111 -1.533 0.125694   
## MaritalStatusMarried -0.1813662 0.2973339 -0.610 0.542060   
## MaritalStatusSingle -0.2097128 0.4111588 -0.510 0.610162   
## MonthlyIncome 0.0001507 0.0001020 1.478 0.139942   
## NumCompaniesWorked -0.3185540 0.0480975 -6.623 6.64e-11 \*\*\*  
## OverTimeYes -0.0297434 0.2608303 -0.114 0.909241   
## PerformanceRating -0.3504092 0.3114099 -1.125 0.260844   
## RelationshipSatisfaction 0.1835890 0.1040185 1.765 0.077970 .   
## StockOptionLevel -0.2290101 0.1819860 -1.258 0.208634   
## TotalWorkingYears 0.2653019 0.0295893 8.966 < 2e-16 \*\*\*  
## TrainingTimesLastYear 0.1260354 0.0877913 1.436 0.151519   
## WorkLifeBalance -0.1972665 0.1586038 -1.244 0.213966   
## YearsInCurrentRole 0.5321339 0.0458357 11.610 < 2e-16 \*\*\*  
## YearsSinceLastPromotion 0.3374339 0.0434305 7.770 2.54e-14 \*\*\*  
## YearsWithCurrManager 0.5555614 0.0456426 12.172 < 2e-16 \*\*\*  
## IndustryRangeBelow Range 0.7736786 0.5669328 1.365 0.172758   
## IndustryRangeWithin Range -0.4177114 0.5324305 -0.785 0.432969   
## IsAttrition NA NA NA NA   
## IsManagementYes -0.1156855 0.3906468 -0.296 0.767205   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.051 on 763 degrees of freedom  
## Multiple R-squared: 0.7926, Adjusted R-squared: 0.7829   
## F-statistic: 81.02 on 36 and 763 DF, p-value: < 2.2e-16

Use leaps package to work out which combinations best explain target

library(leaps)  
bestSubset <- regsubsets(  
YearsAtCompany ~ .,  
data = dataTrain,  
nvmax = 24 #number of dependant variables  
)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

summary(bestSubset)

## Subset selection object  
## Call: regsubsets.formula(YearsAtCompany ~ ., data = dataTrain, nvmax = 24)  
## 37 Variables (and intercept)  
## Forced in Forced out  
## Age FALSE FALSE  
## AttritionYes FALSE FALSE  
## BusinessTravelTravel\_Frequently FALSE FALSE  
## BusinessTravelTravel\_Rarely FALSE FALSE  
## DepartmentResearch & Development FALSE FALSE  
## DepartmentSales FALSE FALSE  
## DistanceFromHome FALSE FALSE  
## Education FALSE FALSE  
## EnvironmentSatisfaction FALSE FALSE  
## GenderMale FALSE FALSE  
## JobInvolvement FALSE FALSE  
## JobRoleHuman Resources FALSE FALSE  
## JobRoleLaboratory Technician FALSE FALSE  
## JobRoleManager FALSE FALSE  
## JobRoleManufacturing Director FALSE FALSE  
## JobRoleResearch Director FALSE FALSE  
## JobRoleResearch Scientist FALSE FALSE  
## JobRoleSales Executive FALSE FALSE  
## JobRoleSales Representative FALSE FALSE  
## MaritalStatusMarried FALSE FALSE  
## MaritalStatusSingle FALSE FALSE  
## MonthlyIncome FALSE FALSE  
## NumCompaniesWorked FALSE FALSE  
## OverTimeYes FALSE FALSE  
## PerformanceRating FALSE FALSE  
## RelationshipSatisfaction FALSE FALSE  
## StockOptionLevel FALSE FALSE  
## TotalWorkingYears FALSE FALSE  
## TrainingTimesLastYear FALSE FALSE  
## WorkLifeBalance FALSE FALSE  
## YearsInCurrentRole FALSE FALSE  
## YearsSinceLastPromotion FALSE FALSE  
## YearsWithCurrManager FALSE FALSE  
## IndustryRangeBelow Range FALSE FALSE  
## IndustryRangeWithin Range FALSE FALSE  
## IsManagementYes FALSE FALSE  
## IsAttrition FALSE FALSE  
## 1 subsets of each size up to 25  
## Selection Algorithm: exhaustive  
## Age AttritionYes BusinessTravelTravel\_Frequently  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " " "   
## 13 ( 1 ) " " " " " "   
## 14 ( 1 ) " " " " " "   
## 15 ( 1 ) " " " " " "   
## 16 ( 1 ) " " " " " "   
## 17 ( 1 ) " " " " " "   
## 18 ( 1 ) " " " " " "   
## 19 ( 1 ) " " "\*" " "   
## 20 ( 1 ) "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " "   
## 22 ( 1 ) "\*" " " " "   
## 23 ( 1 ) "\*" " " " "   
## 24 ( 1 ) "\*" " " " "   
## 25 ( 1 ) "\*" " " " "   
## BusinessTravelTravel\_Rarely DepartmentResearch & Development  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) " " " "   
## 6 ( 1 ) " " " "   
## 7 ( 1 ) " " " "   
## 8 ( 1 ) " " " "   
## 9 ( 1 ) " " "\*"   
## 10 ( 1 ) " " "\*"   
## 11 ( 1 ) " " "\*"   
## 12 ( 1 ) " " "\*"   
## 13 ( 1 ) " " "\*"   
## 14 ( 1 ) " " "\*"   
## 15 ( 1 ) " " "\*"   
## 16 ( 1 ) " " "\*"   
## 17 ( 1 ) " " "\*"   
## 18 ( 1 ) " " "\*"   
## 19 ( 1 ) " " "\*"   
## 20 ( 1 ) " " "\*"   
## 21 ( 1 ) " " "\*"   
## 22 ( 1 ) " " "\*"   
## 23 ( 1 ) " " "\*"   
## 24 ( 1 ) " " "\*"   
## 25 ( 1 ) " " "\*"   
## DepartmentSales DistanceFromHome Education EnvironmentSatisfaction  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) "\*" " " " " " "   
## 10 ( 1 ) "\*" " " " " " "   
## 11 ( 1 ) "\*" " " " " "\*"   
## 12 ( 1 ) "\*" " " " " "\*"   
## 13 ( 1 ) "\*" " " " " "\*"   
## 14 ( 1 ) "\*" " " " " "\*"   
## 15 ( 1 ) "\*" " " " " "\*"   
## 16 ( 1 ) "\*" " " " " "\*"   
## 17 ( 1 ) "\*" "\*" " " "\*"   
## 18 ( 1 ) "\*" " " " " "\*"   
## 19 ( 1 ) "\*" "\*" " " "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*"   
## 21 ( 1 ) "\*" "\*" " " "\*"   
## 22 ( 1 ) "\*" "\*" " " "\*"   
## 23 ( 1 ) "\*" "\*" " " "\*"   
## 24 ( 1 ) "\*" "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*"   
## GenderMale JobInvolvement JobRoleHuman Resources  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " "\*" " "   
## 9 ( 1 ) " " " " "\*"   
## 10 ( 1 ) " " "\*" "\*"   
## 11 ( 1 ) " " " " "\*"   
## 12 ( 1 ) " " "\*" "\*"   
## 13 ( 1 ) "\*" "\*" "\*"   
## 14 ( 1 ) "\*" "\*" "\*"   
## 15 ( 1 ) "\*" "\*" "\*"   
## 16 ( 1 ) "\*" "\*" "\*"   
## 17 ( 1 ) "\*" "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## JobRoleLaboratory Technician JobRoleManager  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) " " " "   
## 6 ( 1 ) " " " "   
## 7 ( 1 ) " " "\*"   
## 8 ( 1 ) " " "\*"   
## 9 ( 1 ) " " " "   
## 10 ( 1 ) " " " "   
## 11 ( 1 ) " " " "   
## 12 ( 1 ) " " " "   
## 13 ( 1 ) " " " "   
## 14 ( 1 ) " " " "   
## 15 ( 1 ) " " " "   
## 16 ( 1 ) " " " "   
## 17 ( 1 ) " " " "   
## 18 ( 1 ) " " " "   
## 19 ( 1 ) " " " "   
## 20 ( 1 ) " " " "   
## 21 ( 1 ) " " " "   
## 22 ( 1 ) " " " "   
## 23 ( 1 ) " " " "   
## 24 ( 1 ) " " " "   
## 25 ( 1 ) " " "\*"   
## JobRoleManufacturing Director JobRoleResearch Director  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) " " " "   
## 6 ( 1 ) " " " "   
## 7 ( 1 ) " " " "   
## 8 ( 1 ) " " " "   
## 9 ( 1 ) " " " "   
## 10 ( 1 ) " " " "   
## 11 ( 1 ) " " " "   
## 12 ( 1 ) " " " "   
## 13 ( 1 ) " " " "   
## 14 ( 1 ) " " " "   
## 15 ( 1 ) " " " "   
## 16 ( 1 ) " " " "   
## 17 ( 1 ) " " " "   
## 18 ( 1 ) " " " "   
## 19 ( 1 ) " " " "   
## 20 ( 1 ) " " " "   
## 21 ( 1 ) " " " "   
## 22 ( 1 ) " " " "   
## 23 ( 1 ) "\*" " "   
## 24 ( 1 ) "\*" "\*"   
## 25 ( 1 ) "\*" "\*"   
## JobRoleResearch Scientist JobRoleSales Executive  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) " " " "   
## 6 ( 1 ) " " " "   
## 7 ( 1 ) " " " "   
## 8 ( 1 ) " " " "   
## 9 ( 1 ) " " " "   
## 10 ( 1 ) " " " "   
## 11 ( 1 ) " " " "   
## 12 ( 1 ) " " " "   
## 13 ( 1 ) " " " "   
## 14 ( 1 ) " " " "   
## 15 ( 1 ) " " "\*"   
## 16 ( 1 ) " " "\*"   
## 17 ( 1 ) " " "\*"   
## 18 ( 1 ) " " "\*"   
## 19 ( 1 ) " " "\*"   
## 20 ( 1 ) " " "\*"   
## 21 ( 1 ) " " "\*"   
## 22 ( 1 ) " " "\*"   
## 23 ( 1 ) " " "\*"   
## 24 ( 1 ) " " "\*"   
## 25 ( 1 ) " " "\*"   
## JobRoleSales Representative MaritalStatusMarried MaritalStatusSingle  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " " "   
## 13 ( 1 ) " " " " " "   
## 14 ( 1 ) " " " " " "   
## 15 ( 1 ) "\*" " " " "   
## 16 ( 1 ) "\*" " " " "   
## 17 ( 1 ) "\*" " " " "   
## 18 ( 1 ) "\*" " " " "   
## 19 ( 1 ) "\*" " " " "   
## 20 ( 1 ) "\*" " " " "   
## 21 ( 1 ) "\*" " " " "   
## 22 ( 1 ) "\*" " " " "   
## 23 ( 1 ) "\*" " " " "   
## 24 ( 1 ) "\*" " " " "   
## 25 ( 1 ) "\*" " " " "   
## MonthlyIncome NumCompaniesWorked OverTimeYes PerformanceRating  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " "\*" " " " "   
## 6 ( 1 ) " " "\*" " " " "   
## 7 ( 1 ) " " "\*" " " " "   
## 8 ( 1 ) " " "\*" " " " "   
## 9 ( 1 ) " " "\*" " " " "   
## 10 ( 1 ) " " "\*" " " " "   
## 11 ( 1 ) " " "\*" " " " "   
## 12 ( 1 ) " " "\*" " " " "   
## 13 ( 1 ) " " "\*" " " " "   
## 14 ( 1 ) " " "\*" " " " "   
## 15 ( 1 ) " " "\*" " " " "   
## 16 ( 1 ) " " "\*" " " " "   
## 17 ( 1 ) " " "\*" " " " "   
## 18 ( 1 ) " " "\*" " " " "   
## 19 ( 1 ) " " "\*" " " " "   
## 20 ( 1 ) " " "\*" " " " "   
## 21 ( 1 ) " " "\*" " " " "   
## 22 ( 1 ) " " "\*" " " "\*"   
## 23 ( 1 ) " " "\*" " " "\*"   
## 24 ( 1 ) " " "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" " " " "   
## RelationshipSatisfaction StockOptionLevel TotalWorkingYears  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " "\*"   
## 3 ( 1 ) " " " " "\*"   
## 4 ( 1 ) " " " " "\*"   
## 5 ( 1 ) " " " " "\*"   
## 6 ( 1 ) " " " " "\*"   
## 7 ( 1 ) " " " " "\*"   
## 8 ( 1 ) " " " " "\*"   
## 9 ( 1 ) " " " " "\*"   
## 10 ( 1 ) " " " " "\*"   
## 11 ( 1 ) " " " " "\*"   
## 12 ( 1 ) " " " " "\*"   
## 13 ( 1 ) " " " " "\*"   
## 14 ( 1 ) "\*" " " "\*"   
## 15 ( 1 ) " " " " "\*"   
## 16 ( 1 ) "\*" " " "\*"   
## 17 ( 1 ) "\*" " " "\*"   
## 18 ( 1 ) "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## TrainingTimesLastYear WorkLifeBalance YearsInCurrentRole  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " "\*"   
## 4 ( 1 ) " " " " "\*"   
## 5 ( 1 ) " " " " "\*"   
## 6 ( 1 ) " " " " "\*"   
## 7 ( 1 ) " " " " "\*"   
## 8 ( 1 ) " " " " "\*"   
## 9 ( 1 ) " " " " "\*"   
## 10 ( 1 ) " " " " "\*"   
## 11 ( 1 ) " " " " "\*"   
## 12 ( 1 ) " " " " "\*"   
## 13 ( 1 ) " " " " "\*"   
## 14 ( 1 ) " " " " "\*"   
## 15 ( 1 ) " " " " "\*"   
## 16 ( 1 ) " " " " "\*"   
## 17 ( 1 ) " " " " "\*"   
## 18 ( 1 ) "\*" " " "\*"   
## 19 ( 1 ) "\*" " " "\*"   
## 20 ( 1 ) "\*" " " "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## YearsSinceLastPromotion YearsWithCurrManager IndustryRangeBelow Range  
## 1 ( 1 ) " " "\*" " "   
## 2 ( 1 ) " " "\*" " "   
## 3 ( 1 ) " " "\*" " "   
## 4 ( 1 ) "\*" "\*" " "   
## 5 ( 1 ) "\*" "\*" " "   
## 6 ( 1 ) "\*" "\*" " "   
## 7 ( 1 ) "\*" "\*" " "   
## 8 ( 1 ) "\*" "\*" " "   
## 9 ( 1 ) "\*" "\*" " "   
## 10 ( 1 ) "\*" "\*" " "   
## 11 ( 1 ) "\*" "\*" " "   
## 12 ( 1 ) "\*" "\*" " "   
## 13 ( 1 ) "\*" "\*" " "   
## 14 ( 1 ) "\*" "\*" " "   
## 15 ( 1 ) "\*" "\*" " "   
## 16 ( 1 ) "\*" "\*" " "   
## 17 ( 1 ) "\*" "\*" " "   
## 18 ( 1 ) "\*" "\*" " "   
## 19 ( 1 ) "\*" "\*" " "   
## 20 ( 1 ) "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" " "   
## 22 ( 1 ) "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" " "   
## 25 ( 1 ) "\*" "\*" "\*"   
## IndustryRangeWithin Range IsAttrition IsManagementYes  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) "\*" " " " "   
## 7 ( 1 ) "\*" " " " "   
## 8 ( 1 ) "\*" " " " "   
## 9 ( 1 ) "\*" " " " "   
## 10 ( 1 ) "\*" " " " "   
## 11 ( 1 ) "\*" "\*" " "   
## 12 ( 1 ) "\*" "\*" " "   
## 13 ( 1 ) "\*" "\*" " "   
## 14 ( 1 ) "\*" "\*" " "   
## 15 ( 1 ) "\*" "\*" " "   
## 16 ( 1 ) "\*" "\*" " "   
## 17 ( 1 ) "\*" "\*" " "   
## 18 ( 1 ) "\*" "\*" " "   
## 19 ( 1 ) "\*" " " " "   
## 20 ( 1 ) "\*" " " " "   
## 21 ( 1 ) "\*" " " " "   
## 22 ( 1 ) "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" " "   
## 25 ( 1 ) " " "\*" " "

coef(bestSubset, 1:24)

## [[1]]  
## (Intercept) YearsWithCurrManager   
## 1.602009 1.363701   
##   
## [[2]]  
## (Intercept) TotalWorkingYears YearsWithCurrManager   
## -0.6638536 0.3160117 1.0317725   
##   
## [[3]]  
## (Intercept) TotalWorkingYears YearsInCurrentRole   
## -1.0986752 0.2652621 0.6066922   
## YearsWithCurrManager   
## 0.6495117   
##   
## [[4]]  
## (Intercept) TotalWorkingYears YearsInCurrentRole   
## -0.9138100 0.2378306 0.5163450   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.3802538 0.5760198   
##   
## [[5]]  
## (Intercept) NumCompaniesWorked TotalWorkingYears   
## -0.1300279 -0.3446535 0.2783919   
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager   
## 0.4813875 0.3653776 0.5442383   
##   
## [[6]]  
## (Intercept) NumCompaniesWorked TotalWorkingYears   
## -0.0896039 -0.3408123 0.2751316   
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager   
## 0.4935233 0.3609046 0.5510988   
## IndustryRangeWithin Range   
## -0.9693013   
##   
## [[7]]  
## (Intercept) JobRoleManager NumCompaniesWorked   
## -0.03322845 1.07998724 -0.32341234   
## TotalWorkingYears YearsInCurrentRole YearsSinceLastPromotion   
## 0.25397099 0.50106878 0.36004101   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.55709439 -0.96422427   
##   
## [[8]]  
## (Intercept) JobInvolvement JobRoleManager   
## 0.8448499 -0.3258653 1.1274985   
## NumCompaniesWorked TotalWorkingYears YearsInCurrentRole   
## -0.3222766 0.2522952 0.5021531   
## YearsSinceLastPromotion YearsWithCurrManager IndustryRangeWithin Range   
## 0.3582597 0.5598453 -0.9440142   
##   
## [[9]]  
## (Intercept) DepartmentResearch & Development   
## 3.8166048 -3.9697942   
## DepartmentSales JobRoleHuman Resources   
## -3.8023709 -3.1353849   
## NumCompaniesWorked TotalWorkingYears   
## -0.3305016 0.2639141   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.5003097 0.3570137   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.5557933 -0.9371852   
##   
## [[10]]  
## (Intercept) DepartmentResearch & Development   
## 4.7566653 -4.0368784   
## DepartmentSales JobInvolvement   
## -3.8903180 -0.3219733   
## JobRoleHuman Resources NumCompaniesWorked   
## -3.2080210 -0.3300113   
## TotalWorkingYears YearsInCurrentRole   
## 0.2629362 0.5013597   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.3551829 0.5581827   
## IndustryRangeWithin Range   
## -0.9171605   
##   
## [[11]]  
## (Intercept) DepartmentResearch & Development   
## 3.23508135 -3.92472198   
## DepartmentSales EnvironmentSatisfaction   
## -3.72390620 0.19498457   
## JobRoleHuman Resources NumCompaniesWorked   
## -3.04358576 -0.33443430   
## TotalWorkingYears YearsInCurrentRole   
## 0.26425804 0.49598487   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.35781811 0.56223957   
## IndustryRangeWithin Range IsManagementYes   
## -0.90097072 -0.09045087   
##   
## [[12]]  
## (Intercept) DepartmentResearch & Development   
## 4.17365627 -3.98833081   
## DepartmentSales EnvironmentSatisfaction   
## -3.81326060 0.19838926   
## JobInvolvement JobRoleHuman Resources   
## -0.32588926 -3.11127081   
## NumCompaniesWorked TotalWorkingYears   
## -0.33401699 0.26313915   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.49698867 0.35607733   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.56455891 -0.87972363   
## IsManagementYes   
## -0.06235985   
##   
## [[13]]  
## (Intercept) DepartmentResearch & Development   
## 3.82594819 -3.87362576   
## DepartmentSales EnvironmentSatisfaction   
## -3.69350965 0.20003213   
## GenderMale JobInvolvement   
## 0.39078451 -0.33598012   
## JobRoleHuman Resources NumCompaniesWorked   
## -3.03433070 -0.33225882   
## TotalWorkingYears YearsInCurrentRole   
## 0.26317837 0.50244877   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.35520743 0.56419051   
## IndustryRangeWithin Range IsManagementYes   
## -0.91722172 -0.06926514   
##   
## [[14]]  
## (Intercept) DepartmentResearch & Development   
## 3.49797912 -3.94390777   
## DepartmentSales EnvironmentSatisfaction   
## -3.77043759 0.19508470   
## GenderMale JobInvolvement   
## 0.40153467 -0.34672175   
## JobRoleHuman Resources NumCompaniesWorked   
## -3.17406330 -0.33828104   
## RelationshipSatisfaction TotalWorkingYears   
## 0.16649974 0.26404419   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.50344976 0.35062372   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.56396902 -0.92731518   
## IsManagementYes   
## -0.07622941   
##   
## [[15]]  
## (Intercept) DepartmentResearch & Development   
## 4.1114282 -4.0761249   
## DepartmentSales EnvironmentSatisfaction   
## -2.4021922 0.1940888   
## GenderMale JobInvolvement   
## 0.4162526 -0.3378059   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.2844914 -1.6763905   
## JobRoleSales Representative NumCompaniesWorked   
## -1.4971648 -0.3208287   
## TotalWorkingYears YearsInCurrentRole   
## 0.2501428 0.5096782   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.3629573 0.5626018   
## IndustryRangeWithin Range IsManagementYes   
## -0.8891394 -0.1177252   
##   
## [[16]]  
## (Intercept) DepartmentResearch & Development   
## 3.7844742 -4.1459799   
## DepartmentSales EnvironmentSatisfaction   
## -2.4807468 0.1891630   
## GenderMale JobInvolvement   
## 0.4269055 -0.3485163   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.4235141 -1.6737773   
## JobRoleSales Representative NumCompaniesWorked   
## -1.4960510 -0.3268538   
## RelationshipSatisfaction TotalWorkingYears   
## 0.1659168 0.2510204   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.5106615 0.3583827   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.5623756 -0.8993008   
## IsManagementYes   
## -0.1247430   
##   
## [[17]]  
## (Intercept) DepartmentResearch & Development   
## 3.99089583 -4.16296754   
## DepartmentSales DistanceFromHome   
## -2.48890531 -0.01862239   
## EnvironmentSatisfaction GenderMale   
## 0.18825127 0.44180116   
## JobInvolvement JobRoleHuman Resources   
## -0.35247997 -3.48784632   
## JobRoleSales Executive JobRoleSales Representative   
## -1.65849654 -1.49961622   
## NumCompaniesWorked RelationshipSatisfaction   
## -0.33040614 0.16845611   
## TotalWorkingYears YearsInCurrentRole   
## 0.25116749 0.51421763   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.35853759 0.55829991   
## IndustryRangeWithin Range IsManagementYes   
## -0.92212983 -0.14327593   
##   
## [[18]]  
## (Intercept) DepartmentResearch & Development   
## 3.6990861 -4.2528055   
## DepartmentSales EnvironmentSatisfaction   
## -2.6404621 0.1922341   
## GenderMale JobInvolvement   
## 0.4544443 -0.3468863   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.5419226 -1.6407058   
## JobRoleSales Representative NumCompaniesWorked   
## -1.5136903 -0.3199207   
## RelationshipSatisfaction StockOptionLevel   
## 0.1648429 -0.2177581   
## TotalWorkingYears TrainingTimesLastYear   
## 0.2512991 0.1170714   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.5156886 0.3522762   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.5618898 -0.8835325   
## IsManagementYes   
## -0.1369599   
##   
## [[19]]  
## (Intercept) AttritionYes   
## 3.56531441 0.55150210   
## DepartmentResearch & Development DepartmentSales   
## -4.26803386 -2.71388978   
## DistanceFromHome EnvironmentSatisfaction   
## -0.01787067 0.20966226   
## GenderMale JobInvolvement   
## 0.44652460 -0.30874495   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.60800121 -1.61600230   
## JobRoleSales Representative NumCompaniesWorked   
## -1.56023478 -0.33053452   
## RelationshipSatisfaction StockOptionLevel   
## 0.17371456 -0.18190667   
## TotalWorkingYears TrainingTimesLastYear   
## 0.25404004 0.11995364   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.52462333 0.34566805   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.56188465 -0.87815901   
##   
## [[20]]  
## (Intercept) Age   
## 4.08043666 -0.02153526   
## AttritionYes DepartmentResearch & Development   
## 0.53156757 -4.20399362   
## DepartmentSales DistanceFromHome   
## -2.65922812 -0.01748617   
## EnvironmentSatisfaction GenderMale   
## 0.20787633 0.44107489   
## JobInvolvement JobRoleHuman Resources   
## -0.29738571 -3.52037241   
## JobRoleSales Executive JobRoleSales Representative   
## -1.60580125 -1.59301279   
## NumCompaniesWorked RelationshipSatisfaction   
## -0.32130077 0.17497596   
## StockOptionLevel TotalWorkingYears   
## -0.17356764 0.27170806   
## TrainingTimesLastYear YearsInCurrentRole   
## 0.11897698 0.52056299   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.34670292 0.55641517   
## IndustryRangeWithin Range   
## -0.84263575   
##   
## [[21]]  
## (Intercept) Age   
## 4.67321808 -0.02273708   
## AttritionYes DepartmentResearch & Development   
## 0.51085365 -4.23555373   
## DepartmentSales DistanceFromHome   
## -2.70086387 -0.01843034   
## EnvironmentSatisfaction GenderMale   
## 0.21311384 0.44925410   
## JobInvolvement JobRoleHuman Resources   
## -0.30343999 -3.54864657   
## JobRoleSales Executive JobRoleSales Representative   
## -1.57607775 -1.54216058   
## NumCompaniesWorked RelationshipSatisfaction   
## -0.31757811 0.17518370   
## StockOptionLevel TotalWorkingYears   
## -0.17385940 0.27251918   
## TrainingTimesLastYear WorkLifeBalance   
## 0.12407101 -0.20401864   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.52566214 0.34641003   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.55373618 -0.84677834   
##   
## [[22]]  
## (Intercept) Age   
## 6.30451010 -0.02399646   
## DepartmentResearch & Development DepartmentSales   
## -4.23163906 -2.62741798   
## DistanceFromHome EnvironmentSatisfaction   
## -0.01691465 0.19424162   
## GenderMale JobInvolvement   
## 0.47316034 -0.33977871   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.58961130 -1.60683956   
## JobRoleSales Representative NumCompaniesWorked   
## -1.52792921 -0.31011012   
## PerformanceRating RelationshipSatisfaction   
## -0.38472797 0.16394116   
## StockOptionLevel TotalWorkingYears   
## -0.20692022 0.27027356   
## TrainingTimesLastYear WorkLifeBalance   
## 0.11193216 -0.22112660   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.52016449 0.35289033   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.55027611 -0.85341219   
## IsManagementYes   
## -0.17352199   
##   
## [[23]]  
## (Intercept) Age   
## 6.26766583 -0.02413741   
## DepartmentResearch & Development DepartmentSales   
## -4.17376268 -2.63669797   
## DistanceFromHome EnvironmentSatisfaction   
## -0.01662898 0.19630812   
## GenderMale JobInvolvement   
## 0.45856942 -0.34085994   
## JobRoleHuman Resources JobRoleManufacturing Director   
## -3.56504495 -0.27355222   
## JobRoleSales Executive JobRoleSales Representative   
## -1.59169544 -1.49316852   
## NumCompaniesWorked PerformanceRating   
## -0.30978701 -0.38020940   
## RelationshipSatisfaction StockOptionLevel   
## 0.16248332 -0.20594123   
## TotalWorkingYears TrainingTimesLastYear   
## 0.27061777 0.10736221   
## WorkLifeBalance YearsInCurrentRole   
## -0.21500405 0.52007484   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.35240092 0.55184728   
## IndustryRangeWithin Range IsManagementYes   
## -0.84414531 -0.12234063   
##   
## [[24]]  
## (Intercept) Age   
## 6.16633697 -0.02475961   
## DepartmentResearch & Development DepartmentSales   
## -4.06439259 -2.62680033   
## DistanceFromHome EnvironmentSatisfaction   
## -0.01697216 0.19577668   
## GenderMale JobInvolvement   
## 0.46248319 -0.34281051   
## JobRoleHuman Resources JobRoleManufacturing Director   
## -3.48019342 -0.31994565   
## JobRoleResearch Director JobRoleSales Executive   
## -0.35077818 -1.53341603   
## JobRoleSales Representative NumCompaniesWorked   
## -1.40073447 -0.30770221   
## PerformanceRating RelationshipSatisfaction   
## -0.38638671 0.16488027   
## StockOptionLevel TotalWorkingYears   
## -0.20252836 0.27514280   
## TrainingTimesLastYear WorkLifeBalance   
## 0.10898168 -0.21278900   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.52109037 0.35070792   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.55254101 -0.84502256   
## IsManagementYes   
## -0.11321394

summary(bestSubset)$bic # Get all of the adjusted R^2 values

## [1] -651.5250 -901.7283 -1047.0527 -1111.6394 -1158.7993 -1158.1219  
## [7] -1156.6318 -1154.5599 -1152.4342 -1150.3021 -1147.6885 -1144.4173  
## [13] -1140.5715 -1136.7702 -1132.9473 -1129.1425 -1124.9032 -1120.1521  
## [19] -1115.4574 -1110.4511 -1105.5109 -1100.4274 -1094.2841 -1087.9868  
## [25] -1081.6439

Judging which model is best there are different metrics

which.max(summary(bestSubset)$adjr2)

## [1] 22

which.min(summary(bestSubset)$rss)

## [1] 25

which.min(summary(bestSubset)$cp)

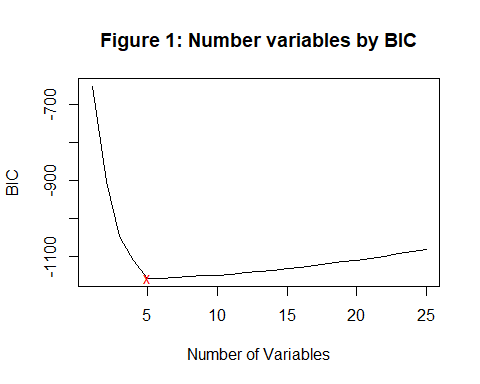
## [1] 17

which.min(summary(bestSubset)$bic)

## [1] 5

Plotting the best(bic, for example) measure

plot(  
summary(bestSubset)$bic,  
type = "l",  
main = "Figure 1: Number variables by BIC",  
xlab = "Number of Variables",  
ylab = "BIC",  
)  
  
bestBIC = which.min(summary(bestSubset)$bic)  
points( #showing the point with lowest bic for example  
x = bestBIC,  
y = summary(bestSubset)$bic[bestBIC],  
col = "red",  
pch = "x"  
)



coef(bestSubset, 5)

## (Intercept) NumCompaniesWorked TotalWorkingYears   
## -0.1300279 -0.3446535 0.2783919   
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager   
## 0.4813875 0.3653776 0.5442383

coef(bestSubset, 17)

## (Intercept) DepartmentResearch & Development   
## 3.99089583 -4.16296754   
## DepartmentSales DistanceFromHome   
## -2.48890531 -0.01862239   
## EnvironmentSatisfaction GenderMale   
## 0.18825127 0.44180116   
## JobInvolvement JobRoleHuman Resources   
## -0.35247997 -3.48784632   
## JobRoleSales Executive JobRoleSales Representative   
## -1.65849654 -1.49961622   
## NumCompaniesWorked RelationshipSatisfaction   
## -0.33040614 0.16845611   
## TotalWorkingYears YearsInCurrentRole   
## 0.25116749 0.51421763   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0.35853759 0.55829991   
## IndustryRangeWithin Range IsManagementYes   
## -0.92212983 -0.14327593

coef(bestSubset, 22)

## (Intercept) Age   
## 6.30451010 -0.02399646   
## DepartmentResearch & Development DepartmentSales   
## -4.23163906 -2.62741798   
## DistanceFromHome EnvironmentSatisfaction   
## -0.01691465 0.19424162   
## GenderMale JobInvolvement   
## 0.47316034 -0.33977871   
## JobRoleHuman Resources JobRoleSales Executive   
## -3.58961130 -1.60683956   
## JobRoleSales Representative NumCompaniesWorked   
## -1.52792921 -0.31011012   
## PerformanceRating RelationshipSatisfaction   
## -0.38472797 0.16394116   
## StockOptionLevel TotalWorkingYears   
## -0.20692022 0.27027356   
## TrainingTimesLastYear WorkLifeBalance   
## 0.11193216 -0.22112660   
## YearsInCurrentRole YearsSinceLastPromotion   
## 0.52016449 0.35289033   
## YearsWithCurrManager IndustryRangeWithin Range   
## 0.55027611 -0.85341219   
## IsManagementYes   
## -0.17352199

#linear models deemed by by regsubsets, bic and c scores  
#best 5  
Linear.Model.02 <- lm(  
 formula = YearsAtCompany ~ NumCompaniesWorked + TotalWorkingYears + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager ,  
 data = dataTrain,  
   
)  
  
#best 17  
Linear.Model.03 <- lm(  
 formula = YearsAtCompany ~ Department+DistanceFromHome+EnvironmentSatisfaction+Gender+  
 JobInvolvement+JobRole+NumCompaniesWorked+RelationshipSatisfaction+TotalWorkingYears+  
 YearsInCurrentRole+YearsSinceLastPromotion +YearsWithCurrManager+IndustryRange+IsManagement,  
 data = dataTrain,  
   
)  
  
#Best 22  
  
Linear.Model.04 <- lm(  
 formula = YearsAtCompany ~ Age+Department+DistanceFromHome+EnvironmentSatisfaction+  
 Gender+JobInvolvement+JobRole+NumCompaniesWorked+  
 PerformanceRating+RelationshipSatisfaction+StockOptionLevel+  
 TotalWorkingYears+TrainingTimesLastYear+WorkLifeBalance+  
 YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager+  
 IndustryRange+IsManagement,  
 data = dataTrain,  
   
)

#final test on linear models deemed by by regsubsets, bic and c scores  
dataTest$PredictionYears.lm02 <- predict(  
 Linear.Model.02,  
 newdata = dataTest,  
 type = "response"  
)  
dataTest$PredictionYears.lm03 <- predict(  
 Linear.Model.03,  
 newdata = dataTest,  
 type = "response"  
)  
  
dataTest$PredictionYears.lm02 <- predict(  
 Linear.Model.02,  
 newdata = dataTest,  
 type = "response"  
)  
dataTest$PredictionYears.lm04 <- predict(  
 Linear.Model.04,  
 newdata = dataTest,  
 type = "response"  
)  
  
mean((dataTest$YearsAtCompany - dataTest$PredictionYears.lm02)^2) #least so it is best

## [1] 6.16559

mean((dataTest$YearsAtCompany - dataTest$PredictionYears.lm03)^2)

## [1] 6.498167

mean((dataTest$YearsAtCompany - dataTest$PredictionYears.lm04)^2)

## [1] 6.628791

#what about on train data  
  
dataTrain$PredictionYears.lm02 <- predict(  
 Linear.Model.02,  
 newdata = dataTrain,  
 type = "response"  
)  
dataTrain$PredictionYears.lm03 <- predict(  
 Linear.Model.03,  
 newdata = dataTrain,  
 type = "response"  
)  
summary(Linear.Model.02)

##   
## Call:  
## lm(formula = YearsAtCompany ~ NumCompaniesWorked + TotalWorkingYears +   
## YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager,   
## data = dataTrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.3631 -1.5165 -0.0474 1.1963 17.0180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.13003 0.22839 -0.569 0.569   
## NumCompaniesWorked -0.34465 0.04636 -7.435 2.71e-13 \*\*\*  
## TotalWorkingYears 0.27839 0.01695 16.421 < 2e-16 \*\*\*  
## YearsInCurrentRole 0.48139 0.04494 10.711 < 2e-16 \*\*\*  
## YearsSinceLastPromotion 0.36538 0.04279 8.538 < 2e-16 \*\*\*  
## YearsWithCurrManager 0.54424 0.04500 12.094 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.104 on 794 degrees of freedom  
## Multiple R-squared: 0.7766, Adjusted R-squared: 0.7752   
## F-statistic: 551.9 on 5 and 794 DF, p-value: < 2.2e-16

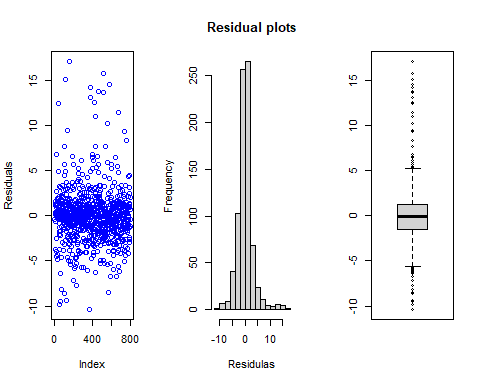
mean((dataTrain$YearsAtCompany - dataTrain$PredictionYears.lm02)^2) #least so it is best

## [1] 9.564905

mean((dataTrain$YearsAtCompany - dataTrain$PredictionYears.lm03)^2)

## [1] 9.050761

#checking assumptions of linearility  
par(mfrow = c(1,3))  
plot(Linear.Model.02$residuals, col = "blue", ylab = "Residuals")  
hist(Linear.Model.02$residuals, xlab = "Residulas", main = "Residual plots")  
boxplot(Linear.Model.02$residuals)



Logistic modelling

#gml model with all variables   
gLinear.Model.05 <- glm(  
 IsAttrition ~ Age +BusinessTravel +Department +   
 DistanceFromHome +Education + EnvironmentSatisfaction +   
 Gender + JobInvolvement + JobRole +   
 MaritalStatus + MonthlyIncome + NumCompaniesWorked +   
 OverTime + PerformanceRating +StockOptionLevel +  
 TotalWorkingYears +TrainingTimesLastYear +WorkLifeBalance +  
 YearsAtCompany +YearsInCurrentRole + YearsSinceLastPromotion +  
 YearsWithCurrManager +IndustryRange ,  
 data = dataTrain,  
 family = binomial,  
 control = list(maxit = 27)  
)  
  
summary(gLinear.Model.05)

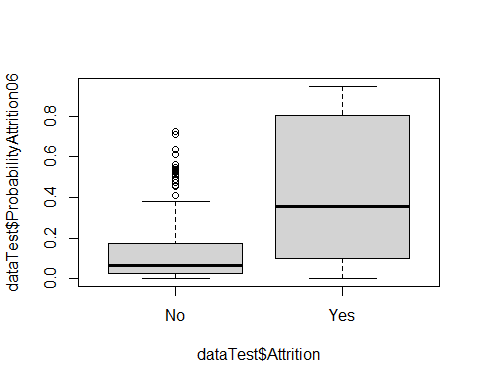
##   
## Call:  
## glm(formula = IsAttrition ~ Age + BusinessTravel + Department +   
## DistanceFromHome + Education + EnvironmentSatisfaction +   
## Gender + JobInvolvement + JobRole + MaritalStatus + MonthlyIncome +   
## NumCompaniesWorked + OverTime + PerformanceRating + StockOptionLevel +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager + IndustryRange, family = binomial,   
## data = dataTrain, control = list(maxit = 27))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.67019 -0.49567 -0.22711 -0.05249 3.07204   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.572e+01 7.207e+02 -0.022 0.982596   
## Age -4.400e-02 1.845e-02 -2.385 0.017082 \*   
## BusinessTravelTravel\_Frequently 2.886e+00 6.509e-01 4.434 9.26e-06 \*\*\*  
## BusinessTravelTravel\_Rarely 1.715e+00 6.079e-01 2.821 0.004792 \*\*   
## DepartmentResearch & Development 1.542e+01 7.207e+02 0.021 0.982925   
## DepartmentSales 1.524e+01 7.207e+02 0.021 0.983133   
## DistanceFromHome 3.096e-02 1.442e-02 2.147 0.031781 \*   
## Education 4.730e-02 1.211e-01 0.391 0.696129   
## EnvironmentSatisfaction -4.062e-01 1.126e-01 -3.609 0.000308 \*\*\*  
## GenderMale 3.665e-01 2.509e-01 1.461 0.144143   
## JobInvolvement -7.429e-01 1.727e-01 -4.303 1.69e-05 \*\*\*  
## JobRoleHuman Resources 1.749e+01 7.207e+02 0.024 0.980642   
## JobRoleLaboratory Technician 2.010e+00 7.906e-01 2.542 0.011026 \*   
## JobRoleManager 2.133e+00 1.342e+00 1.589 0.112093   
## JobRoleManufacturing Director 1.165e+00 8.180e-01 1.424 0.154467   
## JobRoleResearch Director -2.160e+00 1.853e+00 -1.165 0.243878   
## JobRoleResearch Scientist 1.212e+00 8.111e-01 1.495 0.135036   
## JobRoleSales Executive 2.299e+00 1.331e+00 1.727 0.084128 .   
## JobRoleSales Representative 2.974e+00 1.413e+00 2.105 0.035272 \*   
## MaritalStatusMarried 3.151e-01 3.613e-01 0.872 0.383168   
## MaritalStatusSingle 9.866e-01 4.777e-01 2.065 0.038899 \*   
## MonthlyIncome -8.080e-05 1.316e-04 -0.614 0.539092   
## NumCompaniesWorked 1.931e-01 5.206e-02 3.709 0.000208 \*\*\*  
## OverTimeYes 2.271e+00 2.736e-01 8.299 < 2e-16 \*\*\*  
## PerformanceRating -1.869e-01 3.463e-01 -0.540 0.589440   
## StockOptionLevel -1.221e-01 2.151e-01 -0.567 0.570380   
## TotalWorkingYears -2.197e-02 3.675e-02 -0.598 0.549922   
## TrainingTimesLastYear -1.424e-01 9.825e-02 -1.450 0.147190   
## WorkLifeBalance -1.508e-01 1.776e-01 -0.849 0.395773   
## YearsAtCompany 1.129e-01 5.015e-02 2.250 0.024428 \*   
## YearsInCurrentRole -2.546e-01 6.627e-02 -3.842 0.000122 \*\*\*  
## YearsSinceLastPromotion 2.063e-01 5.979e-02 3.451 0.000559 \*\*\*  
## YearsWithCurrManager -1.773e-01 6.340e-02 -2.797 0.005162 \*\*   
## IndustryRangeBelow Range -4.934e-01 6.900e-01 -0.715 0.474578   
## IndustryRangeWithin Range -4.444e-01 6.853e-01 -0.649 0.516635   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 738.85 on 799 degrees of freedom  
## Residual deviance: 469.52 on 765 degrees of freedom  
## AIC: 539.52  
##   
## Number of Fisher Scoring iterations: 15

gLinear.Model.06 <- glm(  
 IsAttrition ~ Age +DistanceFromHome + EnvironmentSatisfaction +   
 JobInvolvement + JobRole + MaritalStatus +   
 NumCompaniesWorked + OverTime +YearsInCurrentRole +   
 YearsSinceLastPromotion +YearsWithCurrManager +RelationshipSatisfaction,  
 data = dataTrain,  
 family = binomial,  
 control = list(maxit = 27)  
)  
summary(gLinear.Model.06)

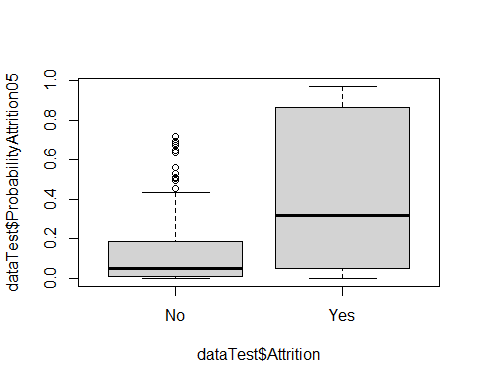
##   
## Call:  
## glm(formula = IsAttrition ~ Age + DistanceFromHome + EnvironmentSatisfaction +   
## JobInvolvement + JobRole + MaritalStatus + NumCompaniesWorked +   
## OverTime + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager + RelationshipSatisfaction, family = binomial,   
## data = dataTrain, control = list(maxit = 27))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7154 -0.5430 -0.2723 -0.1024 3.3158   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.34044 1.01231 -0.336 0.736643   
## Age -0.03430 0.01482 -2.314 0.020667 \*   
## DistanceFromHome 0.02713 0.01353 2.004 0.045056 \*   
## EnvironmentSatisfaction -0.31983 0.10618 -3.012 0.002595 \*\*   
## JobInvolvement -0.64902 0.15859 -4.092 4.27e-05 \*\*\*  
## JobRoleHuman Resources 2.60754 0.81953 3.182 0.001464 \*\*   
## JobRoleLaboratory Technician 2.03132 0.66875 3.037 0.002385 \*\*   
## JobRoleManager 0.93541 0.83488 1.120 0.262539   
## JobRoleManufacturing Director 1.00947 0.77119 1.309 0.190544   
## JobRoleResearch Director -0.98528 1.28073 -0.769 0.441705   
## JobRoleResearch Scientist 1.18966 0.68023 1.749 0.080305 .   
## JobRoleSales Executive 1.73418 0.65830 2.634 0.008431 \*\*   
## JobRoleSales Representative 2.83531 0.73333 3.866 0.000110 \*\*\*  
## MaritalStatusMarried 0.31135 0.33107 0.940 0.346985   
## MaritalStatusSingle 1.09065 0.33800 3.227 0.001252 \*\*   
## NumCompaniesWorked 0.15708 0.04716 3.331 0.000865 \*\*\*  
## OverTimeYes 2.33722 0.26259 8.901 < 2e-16 \*\*\*  
## YearsInCurrentRole -0.18014 0.05511 -3.269 0.001080 \*\*   
## YearsSinceLastPromotion 0.20658 0.05176 3.991 6.58e-05 \*\*\*  
## YearsWithCurrManager -0.09824 0.05653 -1.738 0.082239 .   
## RelationshipSatisfaction -0.25694 0.10793 -2.381 0.017280 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 738.85 on 799 degrees of freedom  
## Residual deviance: 509.57 on 779 degrees of freedom  
## AIC: 551.57  
##   
## Number of Fisher Scoring iterations: 6

gLm preditions

dataTest$ProbabilityAttrition05 <- predict(  
 gLinear.Model.05,  
 newdata = dataTest,  
 type = "response"  
)  
  
dataTest$ProbabilityAttrition06 <- predict(  
 gLinear.Model.06,  
 newdata = dataTest,  
 type = "response"  
)  
  
dataTest$PredictionAttrition05 <- ifelse(dataTest$ProbabilityAttrition05 > 0.5, 1, 0)  
dataTest$PredictionAttrition06 <- ifelse(dataTest$ProbabilityAttrition06 > 0.5, 1, 0)  
  
boxplot(  
 dataTest$ProbabilityAttrition06 ~ dataTest$Attrition  
   
)



boxplot(  
 dataTest$ProbabilityAttrition05 ~ dataTest$Attrition  
   
)



#Confusion matrix   
  
conf\_matrixAll <- table(dataTest$PredictionAttrition05, dataTest$IsAttrition)  
accuracy.gLinear.Model.05.test <- sum(diag(conf\_matrixAll))/sum(conf\_matrixAll)  
accuracy.gLinear.Model.05.test

## [1] 0.895

conf\_matrixSig <- table(dataTest$PredictionAttrition06, dataTest$IsAttrition)  
accuracy.gLinear.Model.06.test <- sum(diag(conf\_matrixSig))/sum(conf\_matrixSig)  
accuracy.gLinear.Model.06.test

## [1] 0.885

Accuracy on Train data

dataTrain$ProbabilityAttrition05 <- predict(  
 gLinear.Model.05,  
 newdata = dataTrain,  
 type = "response"  
)  
  
dataTrain$ProbabilityAttrition06 <- predict(  
 gLinear.Model.06,  
 newdata = dataTrain,  
 type = "response"  
)  
  
dataTrain$PredictionAttrition05 <- ifelse(dataTrain$ProbabilityAttrition05 > 0.5, 1, 0)  
dataTrain$PredictionAttrition06 <- ifelse(dataTrain$ProbabilityAttrition06 > 0.5, 1, 0)  
  
#Confusion matrix   
  
(conf\_matrixAll.train <- table(dataTrain$PredictionAttrition05, dataTrain$IsAttrition))

##   
## 0 1  
## 0 638 66  
## 1 23 73

accuracy.gLinear.Model.05.train <- sum(diag(conf\_matrixAll.train))/sum(conf\_matrixAll.train)  
accuracy.gLinear.Model.05.train

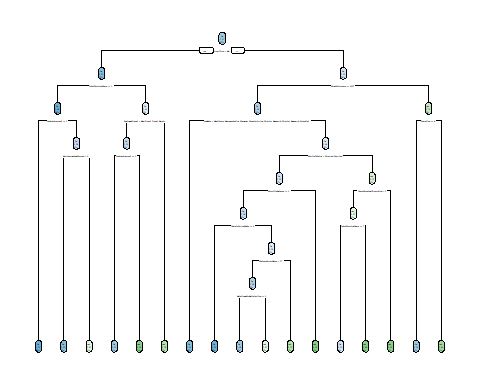
## [1] 0.88875

conf\_matrixSig.train <- table(dataTrain$PredictionAttrition06, dataTrain$IsAttrition)  
accuracy.gLinear.Model.06.train <- sum(diag(conf\_matrixSig.train))/sum(conf\_matrixSig.train)  
accuracy.gLinear.Model.06.train

## [1] 0.875

Modelling with a Decision Tree

library(rpart)  
library(rpart.plot)  
tree <- rpart(  
 Attrition ~ Age +BusinessTravel +Department +   
 DistanceFromHome + Education + EnvironmentSatisfaction +   
 Gender + JobInvolvement + JobRole + MaritalStatus +   
 MonthlyIncome + NumCompaniesWorked + OverTime +   
 PerformanceRating +StockOptionLevel +TotalWorkingYears +  
 TrainingTimesLastYear +WorkLifeBalance +YearsAtCompany +  
 YearsInCurrentRole + YearsSinceLastPromotion +YearsWithCurrManager +  
 IndustryRange + RelationshipSatisfaction + IsManagement,  
 data = dataTrain,  
 method = "class",  
 minsplit=20, #smallest number of observations in the parent node that could be split further.default is 20.  
 minbucket = 1 #smallest number of observations that are allowed in a terminal node  
)  
  
rpart.plot(tree)



summary(tree)

## Call:  
## rpart(formula = Attrition ~ Age + BusinessTravel + Department +   
## DistanceFromHome + Education + EnvironmentSatisfaction +   
## Gender + JobInvolvement + JobRole + MaritalStatus + MonthlyIncome +   
## NumCompaniesWorked + OverTime + PerformanceRating + StockOptionLevel +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager + IndustryRange + RelationshipSatisfaction +   
## IsManagement, data = dataTrain, method = "class", minsplit = 20,   
## minbucket = 1)  
## n= 800   
##   
## CP nsplit rel error xerror xstd  
## 1 0.06834532 0 1.0000000 1.0000000 0.07709893  
## 2 0.04316547 2 0.8633094 0.9424460 0.07529895  
## 3 0.02877698 4 0.7769784 0.9208633 0.07459845  
## 4 0.02158273 5 0.7482014 0.9280576 0.07483355  
## 5 0.01798561 6 0.7266187 0.9352518 0.07506704  
## 6 0.01438849 9 0.6690647 0.9208633 0.07459845  
## 7 0.01079137 14 0.5971223 0.9496403 0.07552929  
## 8 0.01000000 16 0.5755396 0.9352518 0.07506704  
##   
## Variable importance  
## MonthlyIncome OverTime TotalWorkingYears   
## 15 13 13   
## JobRole Age Department   
## 10 6 4   
## JobInvolvement MaritalStatus StockOptionLevel   
## 4 4 3   
## Education DistanceFromHome YearsInCurrentRole   
## 3 3 3   
## YearsAtCompany NumCompaniesWorked WorkLifeBalance   
## 3 3 2   
## BusinessTravel RelationshipSatisfaction YearsSinceLastPromotion   
## 2 2 2   
## IndustryRange YearsWithCurrManager IsManagement   
## 1 1 1   
## EnvironmentSatisfaction TrainingTimesLastYear   
## 1 1   
##   
## Node number 1: 800 observations, complexity param=0.06834532  
## predicted class=No expected loss=0.17375 P(node) =1  
## class counts: 661 139  
## probabilities: 0.826 0.174   
## left son=2 (573 obs) right son=3 (227 obs)  
## Primary splits:  
## OverTime splits as LR, improve=20.23525, (0 missing)  
## TotalWorkingYears < 2.5 to the right, improve=17.26673, (0 missing)  
## YearsAtCompany < 2.5 to the right, improve=15.23627, (0 missing)  
## MonthlyIncome < 2802 to the right, improve=12.86746, (0 missing)  
## YearsWithCurrManager < 0.5 to the right, improve=11.24626, (0 missing)  
##   
## Node number 2: 573 observations, complexity param=0.01798561  
## predicted class=No expected loss=0.1029668 P(node) =0.71625  
## class counts: 514 59  
## probabilities: 0.897 0.103   
## left son=4 (529 obs) right son=5 (44 obs)  
## Primary splits:  
## TotalWorkingYears < 2.5 to the right, improve=8.932573, (0 missing)  
## YearsAtCompany < 1.5 to the right, improve=6.095809, (0 missing)  
## MonthlyIncome < 2089.5 to the right, improve=5.581737, (0 missing)  
## JobInvolvement < 1.5 to the right, improve=5.016271, (0 missing)  
## YearsWithCurrManager < 0.5 to the right, improve=4.202507, (0 missing)  
## Surrogate splits:  
## MonthlyIncome < 2009.5 to the right, agree=0.949, adj=0.341, (0 split)  
## Age < 20.5 to the right, agree=0.941, adj=0.227, (0 split)  
##   
## Node number 3: 227 observations, complexity param=0.06834532  
## predicted class=No expected loss=0.3524229 P(node) =0.28375  
## class counts: 147 80  
## probabilities: 0.648 0.352   
## left son=6 (182 obs) right son=7 (45 obs)  
## Primary splits:  
## MonthlyIncome < 2537 to the right, improve=14.442130, (0 missing)  
## JobRole splits as LRRLLLRRR, improve=12.084960, (0 missing)  
## TotalWorkingYears < 8.5 to the right, improve=10.179660, (0 missing)  
## YearsAtCompany < 2.5 to the right, improve= 9.818211, (0 missing)  
## Age < 35.5 to the right, improve= 8.360302, (0 missing)  
## Surrogate splits:  
## TotalWorkingYears < 1.5 to the right, agree=0.850, adj=0.244, (0 split)  
## Age < 23.5 to the right, agree=0.841, adj=0.200, (0 split)  
## JobRole splits as LLLLLLLLR, agree=0.815, adj=0.067, (0 split)  
##   
## Node number 4: 529 observations, complexity param=0.01079137  
## predicted class=No expected loss=0.07750473 P(node) =0.66125  
## class counts: 488 41  
## probabilities: 0.922 0.078   
## left son=8 (501 obs) right son=9 (28 obs)  
## Primary splits:  
## JobInvolvement < 1.5 to the right, improve=2.563346, (0 missing)  
## BusinessTravel splits as LRL, improve=2.039958, (0 missing)  
## NumCompaniesWorked < 4.5 to the left, improve=1.816652, (0 missing)  
## YearsAtCompany < 30.5 to the left, improve=1.748608, (0 missing)  
## TotalWorkingYears < 39 to the left, improve=1.705219, (0 missing)  
##   
## Node number 5: 44 observations, complexity param=0.01798561  
## predicted class=No expected loss=0.4090909 P(node) =0.055  
## class counts: 26 18  
## probabilities: 0.591 0.409   
## left son=10 (35 obs) right son=11 (9 obs)  
## Primary splits:  
## BusinessTravel splits as LRL, improve=3.075902, (0 missing)  
## JobInvolvement < 1.5 to the right, improve=3.072727, (0 missing)  
## TrainingTimesLastYear < 3.5 to the right, improve=2.969697, (0 missing)  
## JobRole splits as -RR---L-R, improve=2.009235, (0 missing)  
## EnvironmentSatisfaction < 2.5 to the right, improve=1.929569, (0 missing)  
## Surrogate splits:  
## WorkLifeBalance < 3.5 to the left, agree=0.818, adj=0.111, (0 split)  
##   
## Node number 6: 182 observations, complexity param=0.04316547  
## predicted class=No expected loss=0.2637363 P(node) =0.2275  
## class counts: 134 48  
## probabilities: 0.736 0.264   
## left son=12 (99 obs) right son=13 (83 obs)  
## Primary splits:  
## JobRole splits as LRRLLLLRR, improve=11.496700, (0 missing)  
## Department splits as LLR, improve= 4.554437, (0 missing)  
## MaritalStatus splits as LLR, improve= 4.039536, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve= 3.500220, (0 missing)  
## EnvironmentSatisfaction < 2.5 to the right, improve= 3.087226, (0 missing)  
## Surrogate splits:  
## Department splits as LLR, agree=0.808, adj=0.578, (0 split)  
## MonthlyIncome < 10580 to the right, agree=0.637, adj=0.205, (0 split)  
## TotalWorkingYears < 15.5 to the right, agree=0.632, adj=0.193, (0 split)  
## IndustryRange splits as LRR, agree=0.615, adj=0.157, (0 split)  
## Age < 34.5 to the right, agree=0.599, adj=0.120, (0 split)  
##   
## Node number 7: 45 observations, complexity param=0.02877698  
## predicted class=Yes expected loss=0.2888889 P(node) =0.05625  
## class counts: 13 32  
## probabilities: 0.289 0.711   
## left son=14 (8 obs) right son=15 (37 obs)  
## Primary splits:  
## Education < 3.5 to the right, improve=4.137538, (0 missing)  
## YearsInCurrentRole < 0.5 to the right, improve=3.266667, (0 missing)  
## YearsAtCompany < 3.5 to the right, improve=3.208889, (0 missing)  
## YearsWithCurrManager < 1.5 to the right, improve=2.953175, (0 missing)  
## Age < 33.5 to the right, improve=2.277350, (0 missing)  
## Surrogate splits:  
## TotalWorkingYears < 14 to the right, agree=0.867, adj=0.250, (0 split)  
## YearsAtCompany < 5.5 to the right, agree=0.844, adj=0.125, (0 split)  
##   
## Node number 8: 501 observations  
## predicted class=No expected loss=0.06586826 P(node) =0.62625  
## class counts: 468 33  
## probabilities: 0.934 0.066   
##   
## Node number 9: 28 observations, complexity param=0.01079137  
## predicted class=No expected loss=0.2857143 P(node) =0.035  
## class counts: 20 8  
## probabilities: 0.714 0.286   
## left son=18 (19 obs) right son=19 (9 obs)  
## Primary splits:  
## NumCompaniesWorked < 4.5 to the left, improve=3.849624, (0 missing)  
## EnvironmentSatisfaction < 1.5 to the right, improve=2.578571, (0 missing)  
## YearsAtCompany < 21 to the left, improve=2.197802, (0 missing)  
## JobRole splits as R-LRLLLRR, improve=2.115751, (0 missing)  
## YearsWithCurrManager < 0.5 to the right, improve=2.011905, (0 missing)  
## Surrogate splits:  
## JobRole splits as R-LLLLLRL, agree=0.821, adj=0.444, (0 split)  
## IsManagement splits as LR, agree=0.821, adj=0.444, (0 split)  
## EnvironmentSatisfaction < 1.5 to the right, agree=0.750, adj=0.222, (0 split)  
## StockOptionLevel < 0.5 to the right, agree=0.714, adj=0.111, (0 split)  
## WorkLifeBalance < 1.5 to the right, agree=0.714, adj=0.111, (0 split)  
##   
## Node number 10: 35 observations, complexity param=0.01798561  
## predicted class=No expected loss=0.3142857 P(node) =0.04375  
## class counts: 24 11  
## probabilities: 0.686 0.314   
## left son=20 (32 obs) right son=21 (3 obs)  
## Primary splits:  
## JobInvolvement < 1.5 to the right, improve=3.085714, (0 missing)  
## TrainingTimesLastYear < 3.5 to the right, improve=2.393407, (0 missing)  
## MaritalStatus splits as LLR, improve=1.371429, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=1.285714, (0 missing)  
## Department splits as RLR, improve=1.259627, (0 missing)  
## Surrogate splits:  
## DistanceFromHome < 21.5 to the left, agree=0.943, adj=0.333, (0 split)  
##   
## Node number 11: 9 observations  
## predicted class=Yes expected loss=0.2222222 P(node) =0.01125  
## class counts: 2 7  
## probabilities: 0.222 0.778   
##   
## Node number 12: 99 observations  
## predicted class=No expected loss=0.1010101 P(node) =0.12375  
## class counts: 89 10  
## probabilities: 0.899 0.101   
##   
## Node number 13: 83 observations, complexity param=0.04316547  
## predicted class=No expected loss=0.4578313 P(node) =0.10375  
## class counts: 45 38  
## probabilities: 0.542 0.458   
## left son=26 (53 obs) right son=27 (30 obs)  
## Primary splits:  
## MaritalStatus splits as LLR, improve=5.510480, (0 missing)  
## StockOptionLevel < 0.5 to the right, improve=4.168176, (0 missing)  
## Age < 36.5 to the right, improve=2.724587, (0 missing)  
## YearsAtCompany < 8.5 to the right, improve=2.693348, (0 missing)  
## DistanceFromHome < 23.5 to the left, improve=2.662354, (0 missing)  
## Surrogate splits:  
## StockOptionLevel < 0.5 to the right, agree=0.928, adj=0.800, (0 split)  
## RelationshipSatisfaction < 3.5 to the left, agree=0.699, adj=0.167, (0 split)  
## EnvironmentSatisfaction < 1.5 to the right, agree=0.687, adj=0.133, (0 split)  
## Age < 24.5 to the right, agree=0.663, adj=0.067, (0 split)  
## JobRole splits as -LL----LR, agree=0.663, adj=0.067, (0 split)  
##   
## Node number 14: 8 observations  
## predicted class=No expected loss=0.25 P(node) =0.01  
## class counts: 6 2  
## probabilities: 0.750 0.250   
##   
## Node number 15: 37 observations  
## predicted class=Yes expected loss=0.1891892 P(node) =0.04625  
## class counts: 7 30  
## probabilities: 0.189 0.811   
##   
## Node number 18: 19 observations  
## predicted class=No expected loss=0.1052632 P(node) =0.02375  
## class counts: 17 2  
## probabilities: 0.895 0.105   
##   
## Node number 19: 9 observations  
## predicted class=Yes expected loss=0.3333333 P(node) =0.01125  
## class counts: 3 6  
## probabilities: 0.333 0.667   
##   
## Node number 20: 32 observations  
## predicted class=No expected loss=0.25 P(node) =0.04  
## class counts: 24 8  
## probabilities: 0.750 0.250   
##   
## Node number 21: 3 observations  
## predicted class=Yes expected loss=0 P(node) =0.00375  
## class counts: 0 3  
## probabilities: 0.000 1.000   
##   
## Node number 26: 53 observations, complexity param=0.02158273  
## predicted class=No expected loss=0.3207547 P(node) =0.06625  
## class counts: 36 17  
## probabilities: 0.679 0.321   
## left son=52 (50 obs) right son=53 (3 obs)  
## Primary splits:  
## WorkLifeBalance < 1.5 to the right, improve=2.934340, (0 missing)  
## YearsInCurrentRole < 7.5 to the right, improve=2.856244, (0 missing)  
## YearsAtCompany < 7.5 to the right, improve=2.395927, (0 missing)  
## IndustryRange splits as LRL, improve=2.395927, (0 missing)  
## MonthlyIncome < 3805 to the right, improve=2.391043, (0 missing)  
##   
## Node number 27: 30 observations, complexity param=0.01438849  
## predicted class=Yes expected loss=0.3 P(node) =0.0375  
## class counts: 9 21  
## probabilities: 0.300 0.700   
## left son=54 (21 obs) right son=55 (9 obs)  
## Primary splits:  
## YearsSinceLastPromotion < 1.5 to the left, improve=2.314286, (0 missing)  
## Education < 2.5 to the left, improve=2.304545, (0 missing)  
## RelationshipSatisfaction < 3.5 to the right, improve=2.100000, (0 missing)  
## YearsAtCompany < 2.5 to the right, improve=1.963636, (0 missing)  
## DistanceFromHome < 4.5 to the left, improve=1.600000, (0 missing)  
## Surrogate splits:  
## MonthlyIncome < 9430 to the left, agree=0.833, adj=0.444, (0 split)  
## YearsAtCompany < 10.5 to the left, agree=0.800, adj=0.333, (0 split)  
## YearsInCurrentRole < 6 to the left, agree=0.800, adj=0.333, (0 split)  
## YearsWithCurrManager < 6.5 to the left, agree=0.800, adj=0.333, (0 split)  
## Age < 51.5 to the left, agree=0.767, adj=0.222, (0 split)  
##   
## Node number 52: 50 observations, complexity param=0.01438849  
## predicted class=No expected loss=0.28 P(node) =0.0625  
## class counts: 36 14  
## probabilities: 0.720 0.280   
## left son=104 (11 obs) right son=105 (39 obs)  
## Primary splits:  
## YearsInCurrentRole < 7.5 to the right, improve=2.211282, (0 missing)  
## JobInvolvement < 1.5 to the right, improve=2.160000, (0 missing)  
## TrainingTimesLastYear < 5.5 to the left, improve=2.160000, (0 missing)  
## DistanceFromHome < 16.5 to the left, improve=1.987506, (0 missing)  
## RelationshipSatisfaction < 2.5 to the right, improve=1.926667, (0 missing)  
## Surrogate splits:  
## YearsAtCompany < 8.5 to the right, agree=0.90, adj=0.545, (0 split)  
## YearsSinceLastPromotion < 7.5 to the right, agree=0.88, adj=0.455, (0 split)  
## YearsWithCurrManager < 7.5 to the right, agree=0.84, adj=0.273, (0 split)  
## BusinessTravel splits as LRR, agree=0.82, adj=0.182, (0 split)  
## IndustryRange splits as RRL, agree=0.82, adj=0.182, (0 split)  
##   
## Node number 53: 3 observations  
## predicted class=Yes expected loss=0 P(node) =0.00375  
## class counts: 0 3  
## probabilities: 0.000 1.000   
##   
## Node number 54: 21 observations, complexity param=0.01438849  
## predicted class=Yes expected loss=0.4285714 P(node) =0.02625  
## class counts: 9 12  
## probabilities: 0.429 0.571   
## left son=108 (12 obs) right son=109 (9 obs)  
## Primary splits:  
## TotalWorkingYears < 6.5 to the right, improve=3.174603, (0 missing)  
## YearsAtCompany < 7.5 to the right, improve=2.752381, (0 missing)  
## YearsWithCurrManager < 4.5 to the right, improve=2.752381, (0 missing)  
## DistanceFromHome < 1.5 to the right, improve=1.815126, (0 missing)  
## YearsInCurrentRole < 4.5 to the right, improve=1.810714, (0 missing)  
## Surrogate splits:  
## MonthlyIncome < 3946.5 to the right, agree=0.762, adj=0.444, (0 split)  
## Age < 27 to the right, agree=0.714, adj=0.333, (0 split)  
## TrainingTimesLastYear < 3.5 to the left, agree=0.714, adj=0.333, (0 split)  
## YearsAtCompany < 4.5 to the right, agree=0.714, adj=0.333, (0 split)  
## YearsInCurrentRole < 3.5 to the right, agree=0.714, adj=0.333, (0 split)  
##   
## Node number 55: 9 observations  
## predicted class=Yes expected loss=0 P(node) =0.01125  
## class counts: 0 9  
## probabilities: 0.000 1.000   
##   
## Node number 104: 11 observations  
## predicted class=No expected loss=0 P(node) =0.01375  
## class counts: 11 0  
## probabilities: 1.000 0.000   
##   
## Node number 105: 39 observations, complexity param=0.01438849  
## predicted class=No expected loss=0.3589744 P(node) =0.04875  
## class counts: 25 14  
## probabilities: 0.641 0.359   
## left son=210 (31 obs) right son=211 (8 obs)  
## Primary splits:  
## DistanceFromHome < 16.5 to the left, improve=3.077750, (0 missing)  
## Age < 41.5 to the right, improve=2.970457, (0 missing)  
## YearsSinceLastPromotion < 1.5 to the left, improve=2.357809, (0 missing)  
## RelationshipSatisfaction < 1.5 to the right, improve=2.154075, (0 missing)  
## JobInvolvement < 1.5 to the right, improve=1.732502, (0 missing)  
## Surrogate splits:  
## YearsWithCurrManager < 7.5 to the left, agree=0.846, adj=0.25, (0 split)  
##   
## Node number 108: 12 observations  
## predicted class=No expected loss=0.3333333 P(node) =0.015  
## class counts: 8 4  
## probabilities: 0.667 0.333   
##   
## Node number 109: 9 observations  
## predicted class=Yes expected loss=0.1111111 P(node) =0.01125  
## class counts: 1 8  
## probabilities: 0.111 0.889   
##   
## Node number 210: 31 observations, complexity param=0.01438849  
## predicted class=No expected loss=0.2580645 P(node) =0.03875  
## class counts: 23 8  
## probabilities: 0.742 0.258   
## left son=420 (25 obs) right son=421 (6 obs)  
## Primary splits:  
## RelationshipSatisfaction < 1.5 to the right, improve=2.484301, (0 missing)  
## YearsSinceLastPromotion < 4.5 to the left, improve=2.353726, (0 missing)  
## Age < 41.5 to the right, improve=1.778531, (0 missing)  
## YearsInCurrentRole < 6.5 to the left, improve=1.262272, (0 missing)  
## JobInvolvement < 1.5 to the right, improve=1.137634, (0 missing)  
## Surrogate splits:  
## Age < 49.5 to the left, agree=0.839, adj=0.167, (0 split)  
## YearsAtCompany < 1.5 to the right, agree=0.839, adj=0.167, (0 split)  
##   
## Node number 211: 8 observations  
## predicted class=Yes expected loss=0.25 P(node) =0.01  
## class counts: 2 6  
## probabilities: 0.250 0.750   
##   
## Node number 420: 25 observations  
## predicted class=No expected loss=0.16 P(node) =0.03125  
## class counts: 21 4  
## probabilities: 0.840 0.160   
##   
## Node number 421: 6 observations  
## predicted class=Yes expected loss=0.3333333 P(node) =0.0075  
## class counts: 2 4  
## probabilities: 0.333 0.667

dataTest$PredictionTree <-predict(  
 tree,  
 newdata = dataTest,  
 type = "class"  
)  
Probability.Tree <-predict(  
 tree,  
 newdata = dataTest  
   
)  
  
conf\_matrixTree <- table(dataTest$PredictionTree, dataTest$Attrition)  
accuracy.Tree.test <- sum(diag(conf\_matrixTree))/sum(conf\_matrixTree)  
accuracy.Tree.test

## [1] 0.875

dataTrain$PredictionTree <-predict(  
 tree,  
 newdata = dataTrain,  
 type = "class"  
)  
  
conf\_matrixTree.train <- table(dataTrain$PredictionTree, dataTrain$Attrition)  
accuracy.Tree.train <- sum(diag(conf\_matrixTree.train))/sum(conf\_matrixTree.train)  
accuracy.Tree.train

## [1] 0.9

Random forest

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

require(caTools)

## Loading required package: caTools

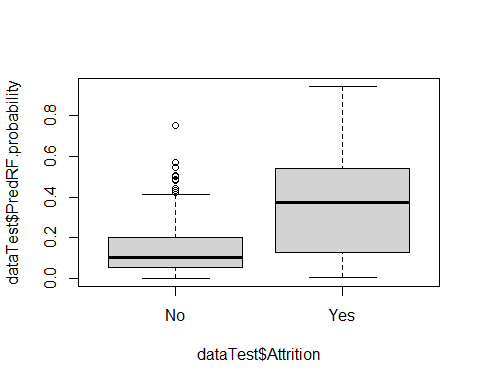
random.forest <- randomForest(  
 IsAttrition ~ Age +BusinessTravel +Department + DistanceFromHome +  
 Education + EnvironmentSatisfaction + Gender +   
 JobInvolvement + JobRole + MaritalStatus +   
 MonthlyIncome + NumCompaniesWorked + OverTime +   
 PerformanceRating +StockOptionLevel +TotalWorkingYears +  
 TrainingTimesLastYear +WorkLifeBalance +YearsAtCompany +  
 YearsInCurrentRole + YearsSinceLastPromotion +YearsWithCurrManager +  
 IndustryRange + IsManagement,  
 data = dataTrain,  
 parms=list(split="information")  
)

## Warning in randomForest.default(m, y, ...): The response has five or fewer  
## unique values. Are you sure you want to do regression?

dataTest$PredRF.probability <-predict(  
 random.forest,  
 newdata = dataTest  
   
)  
dataTest$PredRF <- ifelse(dataTest$PredRF.probability > 0.5, 1, 0)  
  
conf\_matrixRF.test <- table(dataTest$PredRF, dataTest$IsAttrition)  
accuracy.RF.test <- sum(diag(conf\_matrixRF.test))/sum(conf\_matrixRF.test)  
accuracy.RF.test

## [1] 0.9

boxplot(  
 dataTest$PredRF.probability ~ dataTest$Attrition  
   
)



random.forest$importance

## IncNodePurity  
## Age 8.6852081  
## BusinessTravel 2.7412721  
## Department 1.8301940  
## DistanceFromHome 6.8196638  
## Education 2.6985011  
## EnvironmentSatisfaction 3.9355085  
## Gender 1.2247368  
## JobInvolvement 4.2146931  
## JobRole 4.1108727  
## MaritalStatus 2.6875718  
## MonthlyIncome 11.7477706  
## NumCompaniesWorked 4.6401266  
## OverTime 9.1686821  
## PerformanceRating 0.6986922  
## StockOptionLevel 3.4099398  
## TotalWorkingYears 9.0298463  
## TrainingTimesLastYear 4.0266760  
## WorkLifeBalance 2.8704429  
## YearsAtCompany 7.0455829  
## YearsInCurrentRole 3.5287209  
## YearsSinceLastPromotion 3.2485703  
## YearsWithCurrManager 3.5624232  
## IndustryRange 0.7510072  
## IsManagement 0.4488820

What about on training set

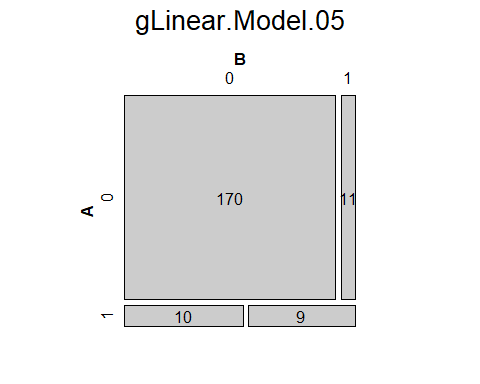
dataTrain$PredRF.probability <-predict(  
 random.forest,  
 newdata = dataTrain,  
   
)  
dataTrain$PredRF <- ifelse(dataTrain$PredRF.probability > 0.5, 1, 0)  
  
conf\_matrixRF.Train <- table(dataTrain$PredRF, dataTrain$IsAttrition)  
accuracy.RF.Train <- sum(diag(conf\_matrixRF.Train))/sum(conf\_matrixRF.Train)  
accuracy.RF.Train

## [1] 0.9925

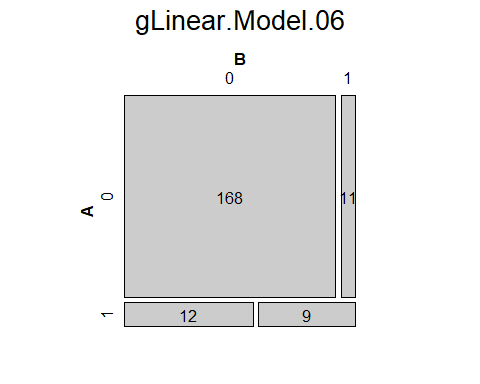
#plot of the models perfomance  
par(mfcol = c(2,2))  
library(vcd)

## Loading required package: grid

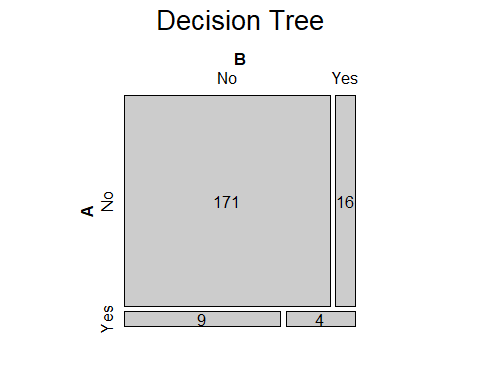
mosaic(  
 structable(conf\_matrixAll),  
 labeling = labeling\_values,   
 pop = FALSE,   
 main = "gLinear.Model.05")



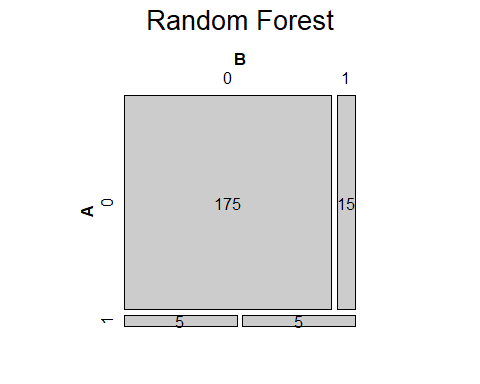
mosaic(  
 structable(conf\_matrixSig),  
 labeling = labeling\_values,   
 pop = FALSE,   
 main = "gLinear.Model.06")



mosaic(  
 structable(conf\_matrixTree),  
 labeling = labeling\_values,   
 pop = FALSE,   
 main = "Decision Tree")



mosaic(  
 structable(conf\_matrixRF.test),  
 labeling = labeling\_values,   
 pop = FALSE,   
 main = "Random Forest")



library(cutpointr)  
tpr <- tpr(c(9,4,5), c(11,16,15))\*100  
fpr <- fpr(c(12,9,5), c(168,171,175))\*100  
tnr <-tnr(c(12,9,5), c(168,171,175))\*100  
fnr <- fnr(c(9,4,5), c(11,16,15))\*100  
models.summary <- data.frame(Model = c("gLinear.Model.06","Decision Tree","Random Forest"), TPR = tpr,FPR = fpr,TNR = tnr,FNR = fnr)  
models.summary

## Model tpr fpr tnr fnr  
## 1 gLinear.Model.06 45 6.666667 93.33333 55  
## 2 Decision Tree 20 5.000000 95.00000 80  
## 3 Random Forest 25 2.777778 97.22222 75

Comparing LMs and GLMs

library("rcompanion")

##   
## Attaching package: 'rcompanion'

## The following object is masked from 'package:cutpointr':  
##   
## accuracy

compareLM(Linear.Model.01,Linear.Model.02,Linear.Model.03)

## $Models  
## Formula   
## 1 "YearsAtCompany ~ Age + Attrition + BusinessTravel + Department + DistanceFromHome + Education + EnvironmentSatisfaction + Gender + JobInvolvement + JobRole + MaritalStatus + MonthlyIncome + NumCompaniesWorked + OverTime + PerformanceRating + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager + IndustryRange + IsAttrition + IsManagement"  
## 2 "YearsAtCompany ~ NumCompaniesWorked + TotalWorkingYears + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager"   
## 3 "YearsAtCompany ~ Department + DistanceFromHome + EnvironmentSatisfaction + Gender + JobInvolvement + JobRole + NumCompaniesWorked + RelationshipSatisfaction + TotalWorkingYears + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager + IndustryRange + IsManagement"   
##   
## $Fit.criteria  
## Rank Df.res AIC AICc BIC R.squared Adj.R.sq p.value Shapiro.W Shapiro.p  
## 1 37 763 4095 4099 4278 0.7926 0.7829 1.336e-233 0.9051 7.057e-22  
## 2 6 794 4091 4091 4124 0.7766 0.7752 1.681e-255 0.8952 6.521e-23  
## 3 24 776 4083 4084 4200 0.7886 0.7823 1.798e-243 0.8996 1.856e-22

compareGLM(gLinear.Model.05,gLinear.Model.06)

## $Models  
## Formula   
## 1 "IsAttrition ~ Age + BusinessTravel + Department + DistanceFromHome + Education + EnvironmentSatisfaction + Gender + JobInvolvement + JobRole + MaritalStatus + MonthlyIncome + NumCompaniesWorked + OverTime + PerformanceRating + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager + IndustryRange"  
## 2 "IsAttrition ~ Age + DistanceFromHome + EnvironmentSatisfaction + JobInvolvement + JobRole + MaritalStatus + NumCompaniesWorked + OverTime + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager + RelationshipSatisfaction"   
##   
## $Fit.criteria  
## Rank Df.res AIC AICc BIC McFadden Cox.and.Snell Nagelkerke p.value  
## 1 35 765 541.5 545.0 710.2 0.3645 0.2859 0.4741 9.143e-39  
## 2 21 779 553.6 554.9 656.6 0.3103 0.2492 0.4133 7.663e-38

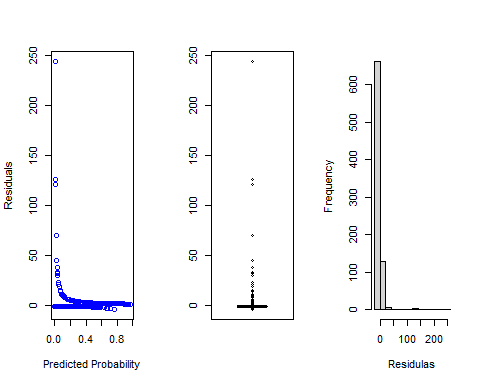
atRisk <- subset(  
 dataTest,  
 (dataTest$YearsAtCompany+1) > dataTest$PredictionYears.lm02  
 )  
table(dataTest$Attrition)

##   
## No Yes   
## 180 20

table(atRisk$Attrition)

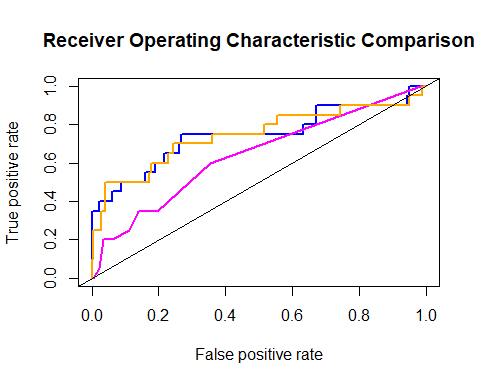
##   
## No Yes   
## 114 12

#residuals plot  
par(mfrow = c(1,3))  
plot(gLinear.Model.06$residuals ~ dataTrain$ProbabilityAttrition06, col = "blue", ylab = "Residuals", xlab = "Predicted Probability")  
boxplot(gLinear.Model.06$residuals)  
hist(gLinear.Model.06$residuals, xlab = "Residulas", main ="")

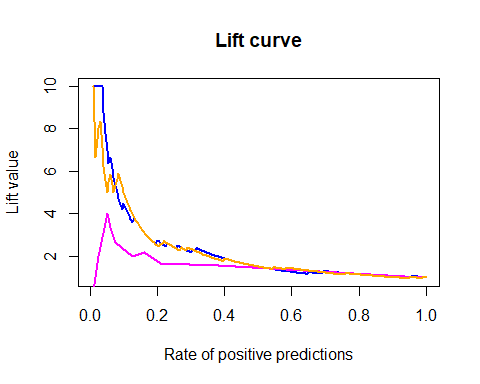


Performance

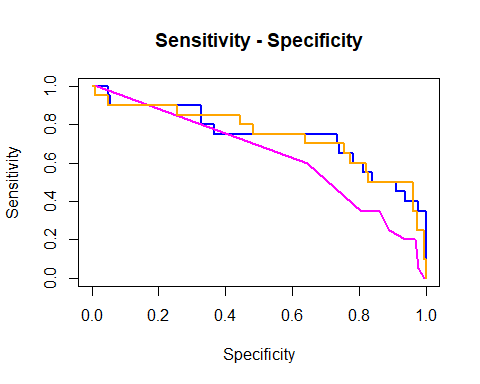
library(ROCR)  
  
pred6 <- prediction(  
 dataTest$ProbabilityAttrition06, dataTest$Attrition  
 )  
pred7 <- prediction(  
 data.frame(Probability.Tree)$Yes, dataTest$Attrition  
 )  
pred8 <- prediction(  
 dataTest$PredRF.probability, dataTest$Attrition  
 )  
  
  
#ROC curves  
  
roc6 <- performance(pred6,"tpr","fpr")  
roc7 <- performance(pred7,"tpr","fpr")  
roc8 <- performance(pred8,"tpr","fpr")  
plot(roc6, col = "blue", lwd = 2,main="Receiver Operating Characteristic Comparison")  
plot(roc7, col = "magenta", lwd = 2, add = TRUE)  
plot(roc8, col="orange", lwd = 2, add = TRUE)  
abline(a = 0, b = 1)



#Lift  
lift6 <- performance(pred6,"lift","rpp")  
lift7 <- performance(pred7,"lift","rpp")  
lift8 <- performance(pred8,"lift","rpp")  
plot(lift6, col = "blue", lwd = 2, main = "Lift curve")  
plot(lift7, col = "magenta", lwd = 2, add = TRUE)  
plot(lift8, col="orange", lwd = 2, add = TRUE)



#Sensitivity - specificity  
  
sens6 <- performance(pred6,"sens","spec")  
sens7 <- performance(pred7,"sens","spec")  
sens8 <- performance(pred8,"sens","spec")  
plot(sens6, col = "blue", lwd = 2,main ="Sensitivity - Specificity")  
plot(sens7, col = "magenta", lwd = 2, add = TRUE)  
plot(sens8, col="orange", lwd = 2, add = TRUE)

 Compare performance

library(performance)

##   
## Attaching package: 'performance'

## The following object is masked from 'package:ROCR':  
##   
## performance

compare\_performance(gLinear.Model.05,gLinear.Model.06, rank = TRUE)

## # Comparison of Model Performance Indices  
##   
## Name | Model | Tjur's R2 | RMSE | Sigma | Log\_loss | Score\_log | Score\_spherical | PCP | AIC\_wt | BIC\_wt | Performance-Score  
## --------------------------------------------------------------------------------------------------------------------------------------------  
## gLinear.Model.05 | glm | 0.384 | 0.293 | 0.783 | 0.293 | -36.865 | 0.016 | 0.823 | 0.998 | < 0.001 | 66.67%  
## gLinear.Model.06 | glm | 0.331 | 0.306 | 0.809 | 0.318 | -34.024 | 0.014 | 0.808 | 0.002 | 1.000 | 33.33%