Core HR
Absence
Payroll
Recruitment
Attrition
Performance
Learning
Succession Planning
HR Analytics

A REPORT ON HUMAN RESOURCE ANALYTICS

BY

DEVIKA NAIK

TWISHA AJWANI

MUKUL ANAPINDI

RUTUJA PATIL

Under the guidance of

MR.DOUGLAS JONES

FOR ACADEMIC YEAR 2016

RUTGERS

THE STATE UNIVERSITY OF NEW JERSEY

ABSTRACT

When employees walk out the door, they take substantial value with them. Not surprisingly then, the broader application of predictive modeling across the enterprise along with the emergence of HR Analytics is leading organizations to ask how HR can start using data to predict and ultimately reduce employee turnover.

The specific goal here is to predict whether an employee will stay or voluntary leave within the next year. In the present data, this means predicting the variable “vol_leave” (0 = stay, 1 = leave) using the other columns of data. You can think of this data as historical data which tells us who did and who did not leave within the last year.

Our initial step is to describe and visualize our data. Then, we will develop two different kinds of predictive models. The first of these is a logistic regression model. Logistic regression models predict the likelihood of a categorical outcome, here staying or leaving.

The second kind of model is known as a decision tree (or a classification tree). A decision tree is essentially a set of rules for splitting the data into buckets to help us predict whether the employees in those buckets will end up in one group (staying) or another group (leaving).

In both cases, we are classifying into just two possible groups. This is known as “binary classification”.

TABLE OF CONTENTS

INTRODUCTION

PROBLEM STATEMENT

Each and every organization needs to diminish costs, increment income, amplify operational proficiency, and concentrate on vital activities to remain productive. Whether in developed or developing markets, HR pioneers regularly battle to bolster the business with the gifted workforce it needs, because of budget and time requirements. One of the biggest challenges a company faces when it plans to launch a new line of services or products is recruiting the right people for the job in time for execution. So also, organizations bear unequivocal and implicit costs when talent exits the organization. It is more regrettable when representatives quit not long, but after in the wake of partaking in a costly preparing program supported by the association. Is there an approach to foresee such dangers and decrease the expenses connected with them?

SOLUTION

The attrition risk score of individual workers can be evaluated with predictive models of attrition. This can aid organizations to keep the potential attrition of high performing workers, factors contributing to attrition guarantee business continuation, and recognize loyal representatives. Employee’s salary, designation, gender, performance, business area, age can be used for data analysis. Supervisors can recognize the key purposes for attrition and along these lines decrease its occurrence.

GOAL AND OVERVIEW

The particular objective here is to anticipate whether an employee will stay or voluntarily leave the following year. In the present information, this implies anticipating the variable "vol_leave" (0 = stay, 1 = leave) using the other columns of data. You can think about this information as historical data which lets us know who did and who did not leave the organization in the most recent year.

Our underlying stride is to depict and visualize our data. Then we will develop two kinds of predictive models. The first of these is a logistic regression model. Logistic regression models predict the likelihood of a categorical outcome, here staying or leaving.

The second sort of model is known as a decision tree. A decision tree is basically an arrangement of standards to part the data into bins to help us foresee whether the workers in those bins will end up in some group(staying) or another group (leaving). In both cases, we are characterizing into only two conceivable groups. This is known as "binary classification".

After we establish the framework for model development, we then disclose how to assess model quality utilizing overall   model accuracy and the Receiver Operator Curve (ROC). The ROC lets us know what the best "limit" or "cutoff"   while figuring out if somebody will leave
While there are absolutely more intricate techniques for foreseeing turnover, logistic regression and decision trees both work exceptionally well. Besides, they are similarly simple to execute and, above all, easier to interpret and explain. This is critical when translating modeling insights into action.

DATASET DESCRIPTION AND R STUDIO

- ## DATA ANALYSIS TECHNOLOGIES USED

R is the main instrument for data analysis, statistics and machine learning. It is a programming dialect, so you can make your own particular objects and packages.

Like all projects, R programs unequivocally record the means of your analysis and make it simple to repeat as well as overhaul examination, which implies we can rapidly attempt numerous ideas or potentially revise issues. It is also platform independent and it's free, so you can utilize it at any business.

- ## DATASET AND R LIBRARIES

Our dataset is downloaded from Kaggle. This dataset has 8 unique attributes like those that you may have in your own HR information. These attributes and its description is mentioned below:

Role: This specifies the designation of the employee. We have 5 types of designation in our dataset namely CEO, VPs, Directors, Managers, and Individual Contributors.

Performance: The performance scale of an employee varies from 1-3. With 1 being lowest and 3 being highest.

Area: The area refers to the business department of an organization. We have areas such as Sales, Finance, Accounting, Marketing and others

Sex : Refers to the gender of the employee. It has two categories namely male female.

Id : Refers to the employee id

Age : Refers to the age of the employee

Salary : Refers to the salary of the employees

Vol_leave : This attribute is based on historical data. It depicts whether an employee has remained or voluntarily left an organization. ‘0’ means to stay and ‘1’ referring to leaving the organization.

Further more the R libraries used during the project implementation are given below:

- library(plyr) : The plyr package is a set of clean and consistent tools that implement the split-apply-combine pattern in R.

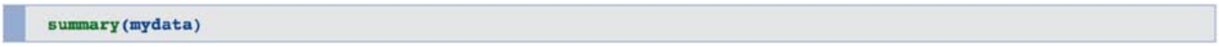
- library(ggplot2) : A system for 'declaratively' creating graphics, based on "The Grammar of Graphics”.

- library(caTools) : Tools: moving window statistics, GIF, Base64, ROC AUC, etc.
- library(RColorBrewer) : Sequential, diverging and qualitative colour scales from colorbrewer.org

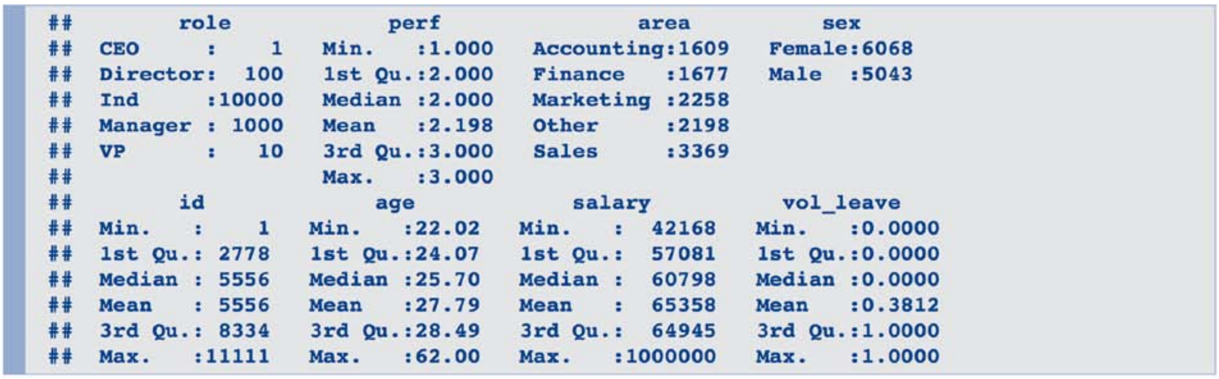
- library(rpart.plot) : Plot an rpart model, automatically tailoring the plot for the model's response type

DATA EXPLORATION AND VISUALIZATION

Let’s start with the summary command.



```
summary(mydata)
```

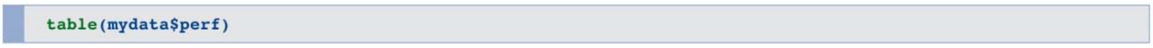


```
##         role            perf                area            sex      
##  CEO     :    1   Min.   :1.000   Accounting:1609   Female:6068  
##  Director:  100   1st Qu.:2.000   Finance   :1677   Male  :5043  
##  Ind     :10000   Median :2.000   Marketing :2258                
##  Manager : 1000   Mean   :2.198   Other     :2198                
##  VP      :   10   3rd Qu.:3.000   Sales     :3369                
##                   Max.   :3.000                                  
##        id              age             salary           vol_leave     
##  Min.   :    1   Min.   :22.02   Min.   :  42168   Min.   :0.0000  
##  1st Qu.: 2778   1st Qu.:24.07   1st Qu.:  57081   1st Qu.:0.0000  
##  Median : 5556   Median :25.70   Median :  60798   Median :0.0000  
##  Mean   : 5556   Mean   :27.79   Mean   :  65358   Mean   :0.3812  
##  3rd Qu.: 8334   3rd Qu.:28.49   3rd Qu.:  64945   3rd Qu.:1.0000  
##  Max.   :11111   Max.   :62.00   Max.   :1000000   Max.   :1.0000  
```

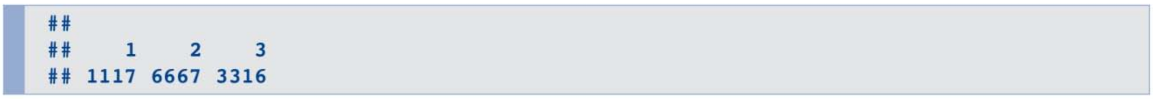
The summary information lets us know that we have 5 fundamental roles: CEO, VPs, Directors, Managers, and Individual Contributors. Since CEOs and VPs encounter an altogether different labormarket than Directors, Managers, and Individuals, hence incorporating them in our modelingeffort doesn't bode well.

□ PERFORMANCE

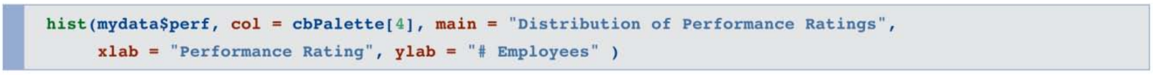
We’ll start with performance, here graded on a simple 1 (low) to 3 (high) scale.



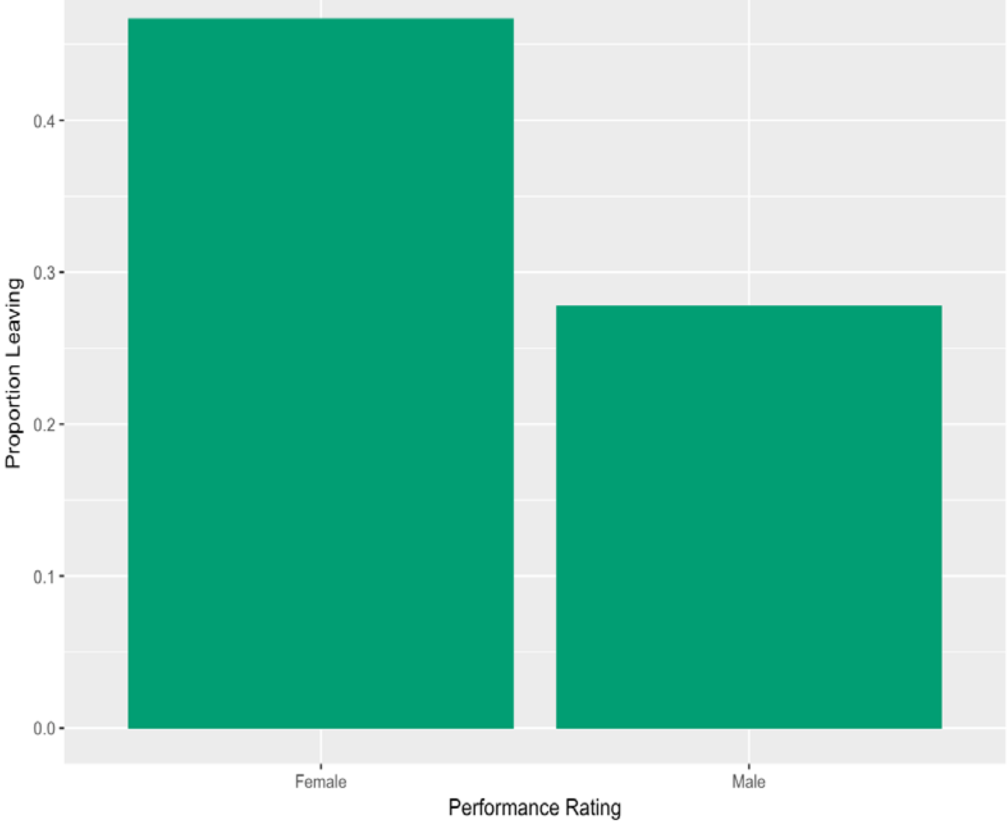
```
table(mydata$perf)
```



```
## 
##    1    2    3 
## 1117 6667 3316 
```



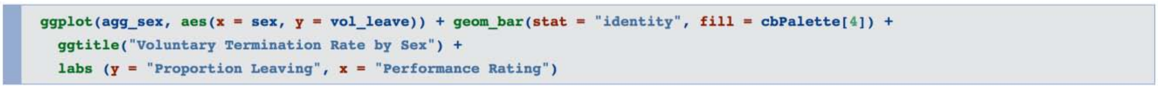
```
hist(mydata$perf, col = cbPalette[4], main = "Distribution of Performance Ratings",
     xlab = "Performance Rating", ylab = "# Employees" )
```



The table and histograms give us a glance at the distributions. With justthree conceivable qualities, it doesn't bode well to discuss it, as being normal or skewed, however we can see that there are reasonable number of individuals in each group.

When we utilize the aggregate function to separate the probabilities by performance groupings, we can see turnover is much higher for the elite group. That is a major issue and certainly something we decided to keep for the model.

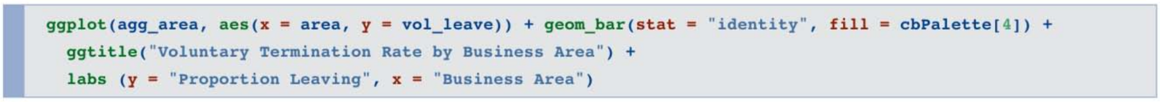
□ ANALYZING THE DISTRIBUTION OF MALES AND FEMALES.



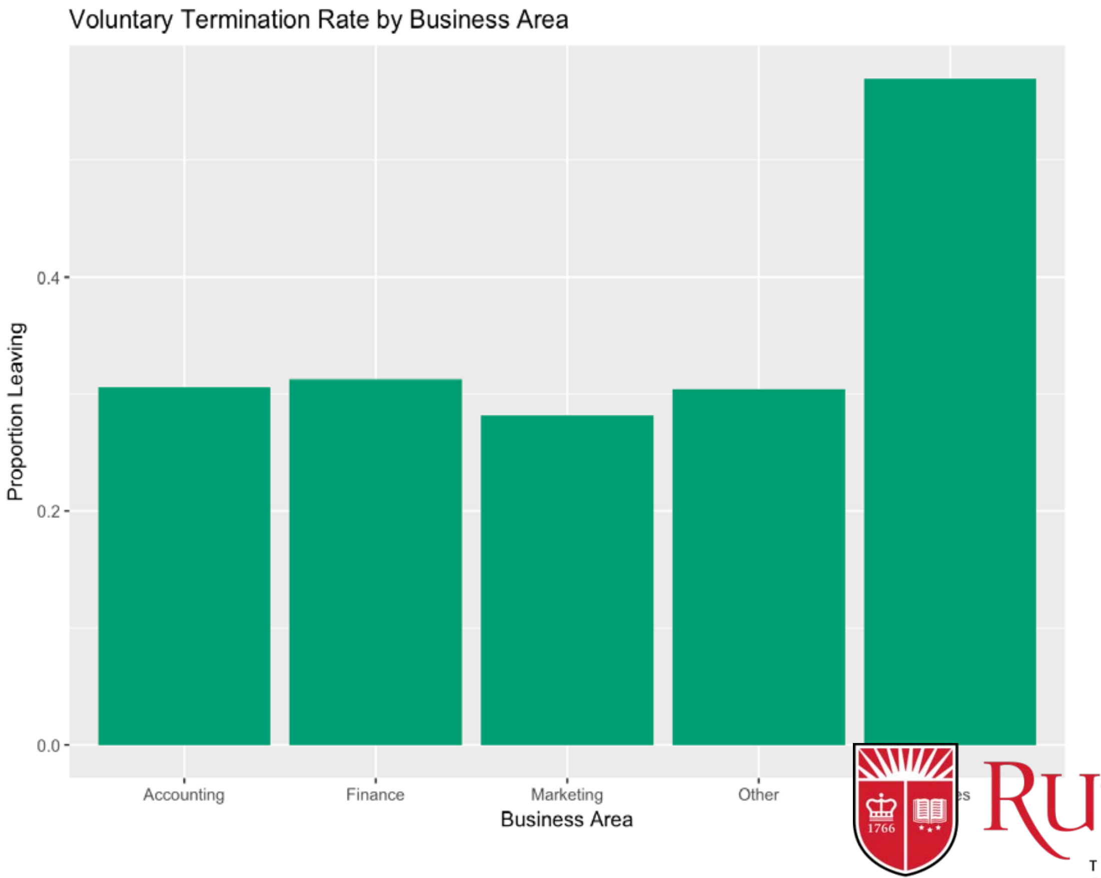
```
ggplot(agg_sex, aes(x = sex, y = vol_leave)) + geom_bar(stat = "identity", fill = cbPalette[4]) +
  ggtitle("Voluntary Termination Rate by Sex") +
  labs (y = "Proportion Leaving", x = "Performance Rating")
```

Unquestionably significant issues for female workers v/s. the male workers are observed. In such an event, alerts ought to go off and the investigative needs would begin with more distinct analyses and some extra focused modeling.

□ PLOTTING THE VOLUNTARY TERMINATION ON THE BASIS OF BUSINESS AREA



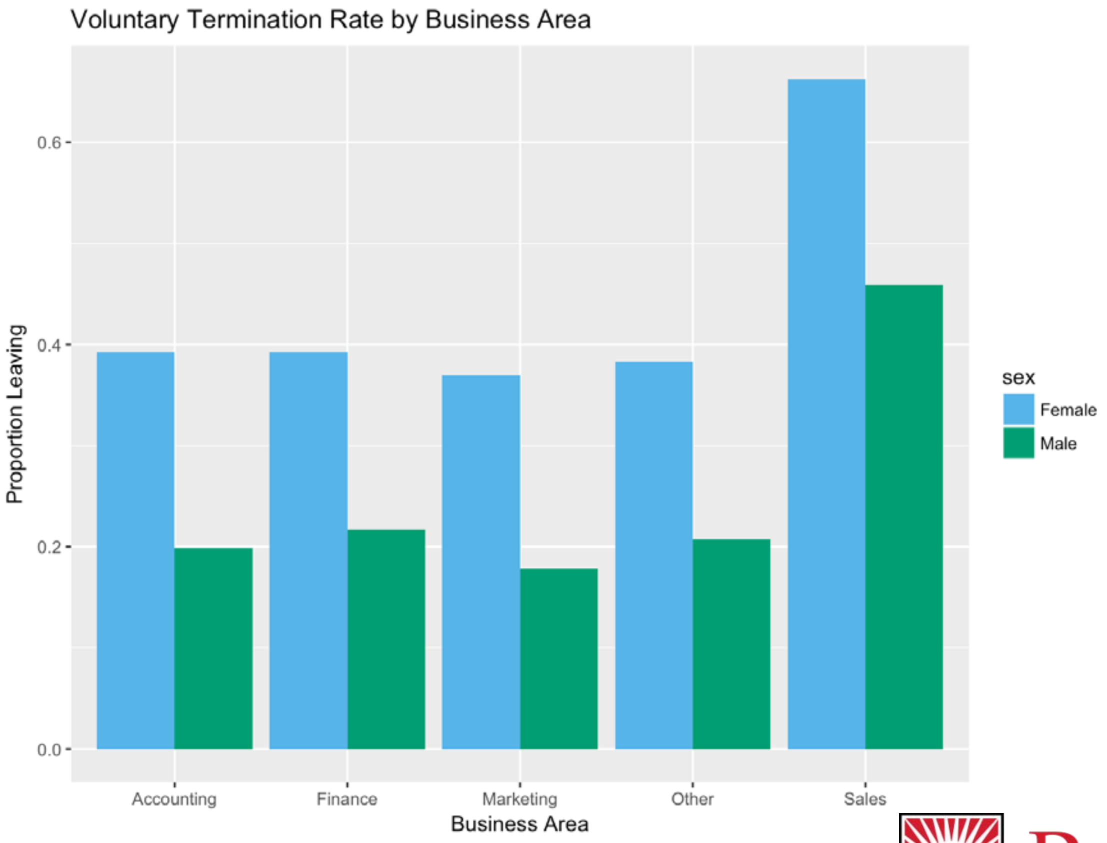
```
ggplot(agg_area, aes(x = area, y = vol_leave)) + geom_bar(stat = "identity", fill = cbPalette[4]) +
  ggtitle("Voluntary Termination Rate by Business Area") +
  labs (y = "Proportion Leaving", x = "Business Area")
```

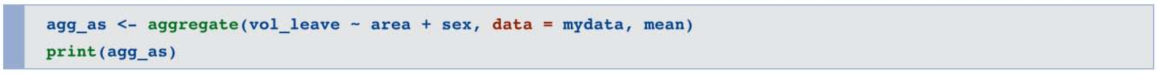


Those in Sales are much more likely to leave. Again, we’ll definitelywant to keep this for our models too.

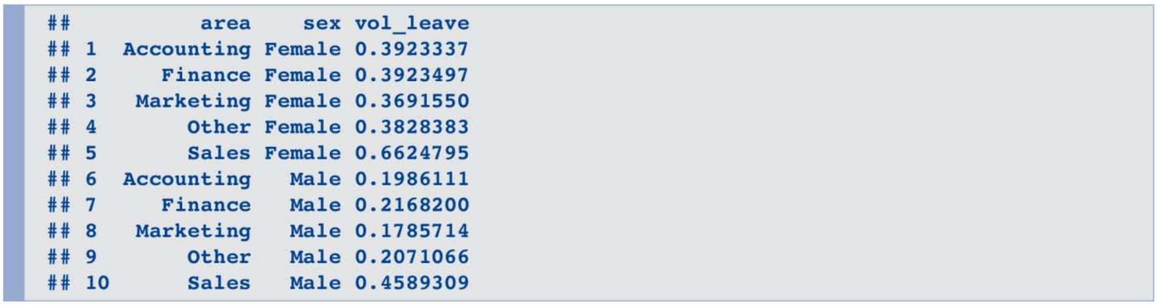
□ SEGMENTING BUSINESS AREA BASED ON GENDER AND ANALYZING THE BUSINESS SECTORS AGAIN

Plotting the findings





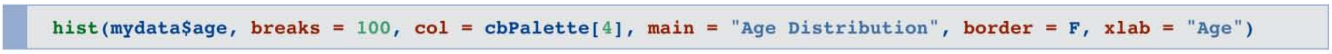
```
agg_as <- aggregate(vol_leave ~ area + sex, data = mydata, mean)
print(agg_as)
```



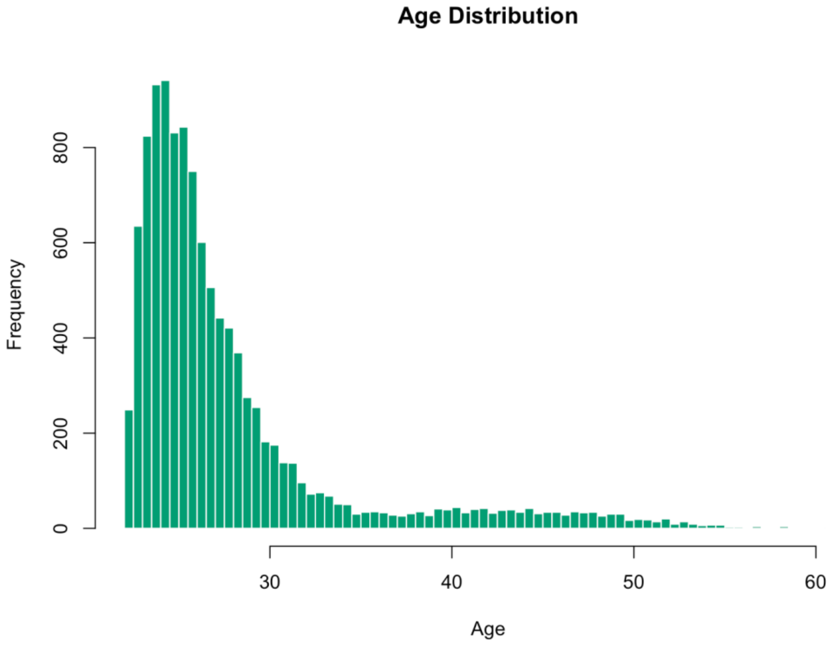
```
##          area    sex vol_leave
## 1  Accounting Female 0.3923337
## 2     Finance Female 0.3923497
## 3   Marketing Female 0.3691550
## 4       Other Female 0.3828383
## 5       Sales Female 0.6624795
## 6  Accounting   Male 0.1986111
## 7     Finance   Male 0.2168200
## 8   Marketing   Male 0.1785714
## 9       Other   Male 0.2071066
## 10      Sales   Male 0.4589309
```

While observing the plot, we can see the difference is for the most part 20% across all areas. There might a marginally hoisted contrast for females in the business assemble, yet nothing gigantic.

□ ANALYZING THE AGE OF EMPLOYEES



```
hist(mydata$age, breaks = 100, col = cbPalette[4], main = "Age Distribution", border = F, xlab = "Age")
```

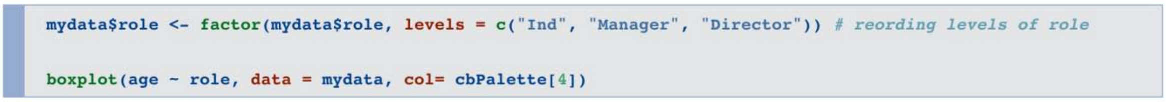


There is a solid skew here, with half of our workforce somewhere around 22 and 26 years old. We should recall however that we have three distinct levels: people, supervisors, and executives.

It will be more informative to see how those ages breakdown when we take that into account.Therefore box plots have been utilized for this purpose.

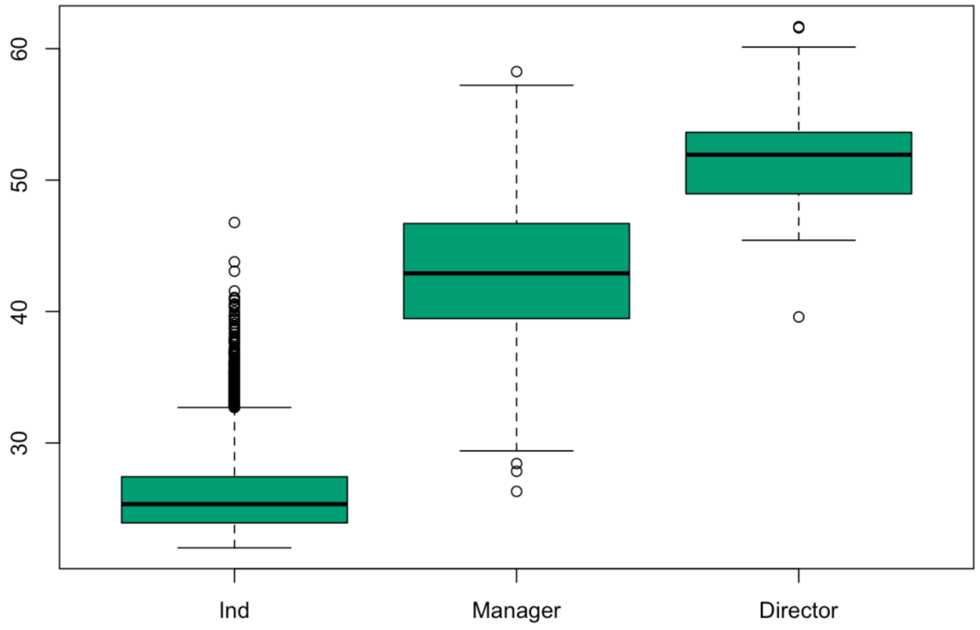
□ ANALYZING THE ROLES OF THE EMPLOYEES TO GET A BETTER IDEA OF THE AGE DISTRIBUTION

Our box plot indicates solid contrasts in age by the role with no meaningful overlap. This implies there is a solid relationship between Age and Role.



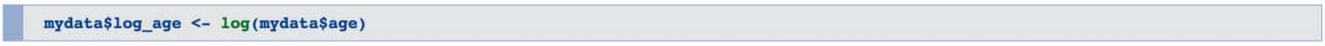
```
mydata$role <- factor(mydata$role, levels = c("Ind", "Manager", "Director")) # reording levels of role

boxplot(age ~ role, data = mydata, col= cbPalette[4])
```



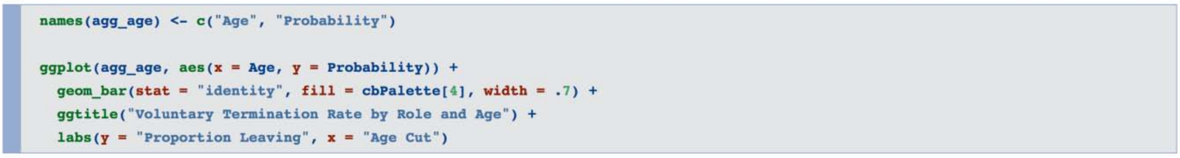
If we just held one of these factors in the model,we might mistakenly attribute turnover differences to age when they are instead more related to role, or vice versa.Still, given that age is to a great degree skewed,we'll address that by making another variable that takes the log of age. This is known as an “data transformation” and is often done to make the distributions of a continuous variable more normally distributed.

This will give us a less skewed value and reduce any problems for inclusion in our later model.



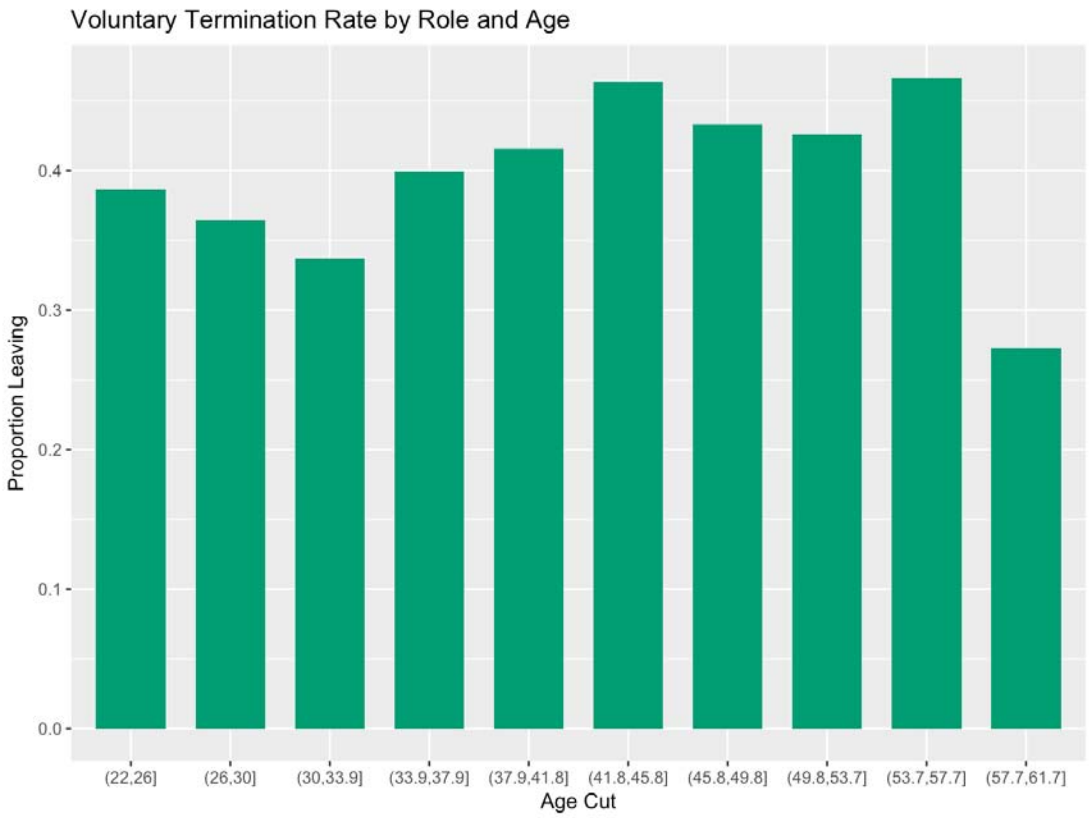
```
mydata$log_age <- log(mydata$age)
```

ANALYZING VOLUNTARY TERMINATION BY AGE AND ROLE



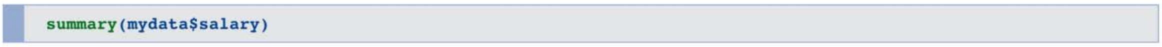
```
names(agg_age) <- c("Age", "Probability")

ggplot(agg_age, aes(x = Age, y = Probability)) +
  geom_bar(stat = "identity", fill = cbPalette[4], width = .7) +
  ggtitle("Voluntary Termination Rate by Role and Age") +
  labs(y = "Proportion Leaving", x = "Age Cut")
```

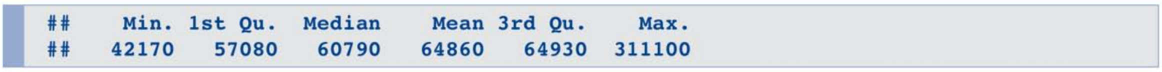


The plot proposes a steep increment in turnover rate between age 34 and 46, for the most part hoisted rate until 58, and followed by a precarious drop-off.

□ ANALYZING THE SALARY PATTERN



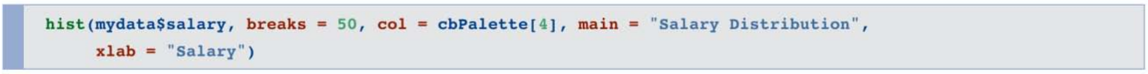
```
summary(mydata$salary)
```



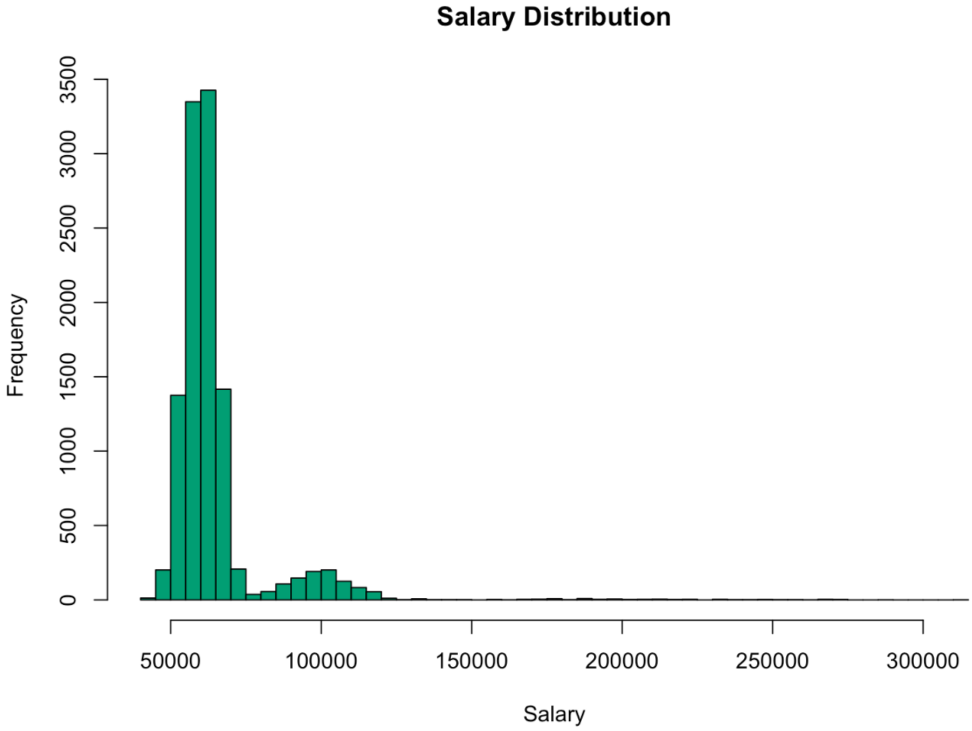
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    42170   57080   60790   64860   64930  311100
```

The median salary is $60800 with the max being $1000000 and the min being $42170.

Plotting the Histogram for the same

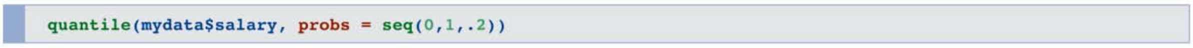


```
hist(mydata$salary, breaks = 50, col = cbPalette[4], main = "Salary Distribution",
     xlab = "Salary")
```

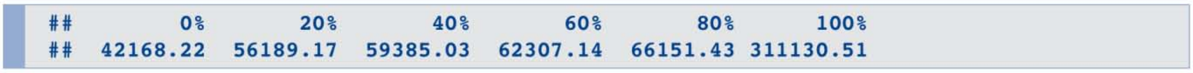


RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

□ FINDING THE SALARY DISTRIBUTION BASED ON QUANTILE



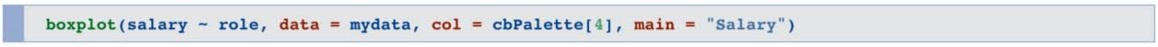
```
quantile(mydata$salary, probs = seq(0,1,.2))
```



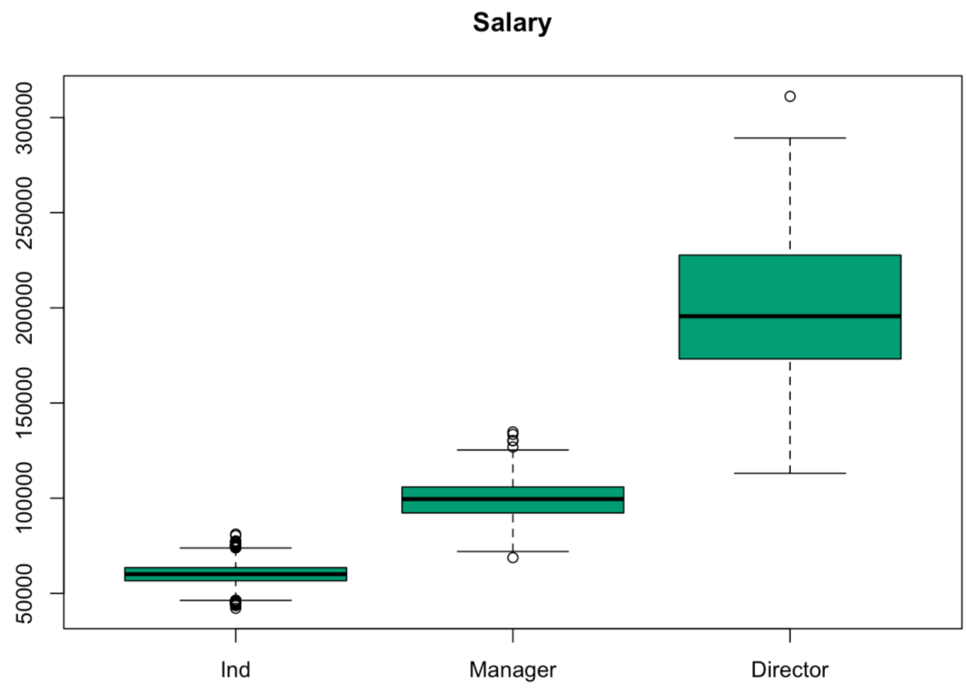
```
##       0%      20%      40%      60%      80%      100%
## 42168.22 56189.17 59385.03 62307.14 66151.43 311130.51
```

Salary variable is highly skewed with almost 80% of the people earning till $66173.65

• PLOTTING THE BOXPLOT TO SHOW SALARY DISTRIBUTION BASED ON ROLES



```
boxplot(salary ~ role, data = mydata, col = cbPalette[4], main = "Salary")
```



From the above graph we can observe three distinct pieces, relating firmly to the role. Given the tremendous contrasts in pay,we decided to find some way to represent salary that is more meaningful across the roles.

For example, an individual contributor would be delighted to make $80,000 considering the median for that group. A manager, on the other hand, might not given the median manager salary of roughly $100,000.

This recommends taking a look at compensation in respect to the middle of a representative's part bodes well.

To do this, we’ll first calculate the median salary for each group and then create a new salary measure that measures a salary relative to that employee’s role median.

For this situation, we'll basically get the distinction between the worker's compensation and that of the median.By this new metric, salaries lower than the median of the relevant role will be negative, those higher than the median will be positive.This will put the salaries on a relative footing.

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

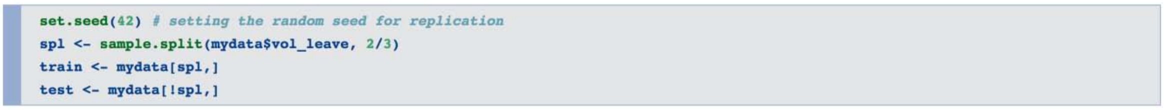
DATA MODELING

- DATA SPLITTING

Before we start creating models, we need to split our data into a training set and a test set.

We will utilize two-thirds of the data for training and model development and one third of the data for testing the models.

The splitting strategy here is random with the limitation that both sets will have a similar extent of leavers and stayers. We set the random seed to a particular number so we can simply replicate our outcomes.



```
set.seed(42) # setting the random seed for replication
spl <- sample.split(mydata$vol_leave, 2/3)
train <- mydata[spl,]
test <- mydata[!spl,]
```

- OUR MODELS: A BRIEF INTRODUCTION

Logistic regression builds a condition that as a result predicts the probability of a two-class result (staying or leaving) utilizing the chosen indicators. Each of the indicators are connected with a "significance"" pointer that lets you know whether the indicator is helpful or not.

By contrast, decision trees work by utilizing the indicators to part the data into buckets using a set of decision rules.

In our present data for instance, the general likelihood of leaving is 38% and we are attempting to make distinctive buckets in which a few containers of workers are a great deal more inclined to leave (say, 70%) andother buckets have employees are much more likely to stay (say, 80%).

- THE LOGISTIC REGRESSION MODEL

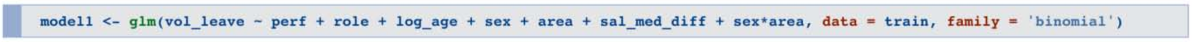
With our dataset now split, we'll utilize the glm function to make our logistic regression model. The plan beneath demonstrates that we are foreseeing the vol_leave variable with the chose set of indicators.

In addition to including each of the variables we discussed in the descriptive analyses above, we are also including the “sex*area” interaction term.

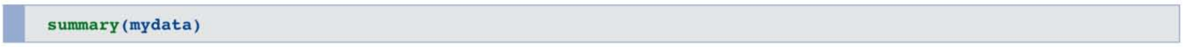
This gives a test to whether the effect of the male/female variable on turnover varies as indicated by the business region (or equally, whether the effect of business region on turnover contrasts for males v/s females).

The "family" argument of the function is set to "binomial", indicating to the model that we have a 0/1 response outcome.

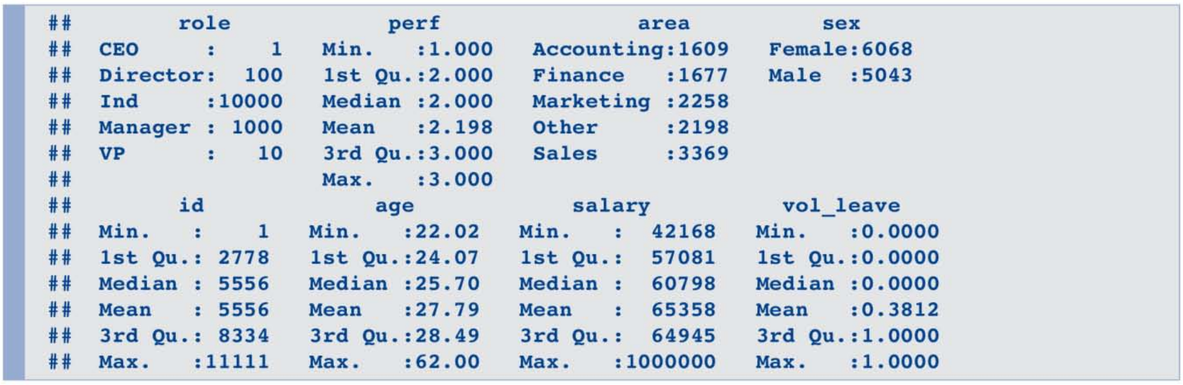
- INTERPRETING THE MODEL OUTPUT



```
model1 <- glm(vol_leave ~ perf + role + log_age + sex + area + sal_med_diff + sex*area, data = train, family = 'binomial')
```



```
summary(mydata)
```



```
##        role             perf                 area            sex      
##  CEO      :    1   Min.   :1.000   Accounting:1609   Female:6068  
##  Director:  100   1st Qu.:2.000   Finance    :1677   Male   :5043  
##  Ind      :10000   Median :2.000   Marketing :2258                
##  Manager : 1000   Mean   :2.198   Other      :2198                
##  VP       :   10   3rd Qu.:3.000   Sales      :3369                
##                    Max.   :3.000                                  
##        id             age             salary          vol_leave     
##  Min.   :    1   Min.   :22.02   Min.   :  42168   Min.   :0.0000  
##  1st Qu.: 2778   1st Qu.:24.07   1st Qu.:  57081   1st Qu.:0.0000  
##  Median : 5556   Median :25.70   Median :  60798   Median :0.0000  
##  Mean   : 5556   Mean   :27.79   Mean   :  65358   Mean   :0.3812  
##  3rd Qu.: 8334   3rd Qu.:28.49   3rd Qu.:  64945   3rd Qu.:1.0000  
##  Max.   :11111   Max.   :62.00   Max.   :1000000   Max.   :1.0000  
```

The Performance Variable

Not surprisingly, we see that those rated more highly on performance are significantly more likely to leave, consistent with our descriptive analyses.

The Role Variable

In the case of Managers, we see that the estimate is positive, meaning that Managers are significantly more likely to leave. Surprisingly though, the designation of Director did not significantly predict turnover.

Given the marked drop-off in turnover we saw for Directors in our descriptive analyses, how can this be?

The short answer is that when we include other variables such as age and performance, the impact of Director per se disappears.

The Age Variable

Our transformed age variable is significant and negative. This tells us that older employees are less likely to leave.

The Male/ Female Variable

We see a significant (p < .05) and negative parameter value males. This means males are less likely to leave than females.

The Area Variable

The interpretation for Area is a bit more complex. The bottom line is that those in the Sales are significantly more likely to leave (positive and statistically significant parameter value) while there are no significant differences associated with the other areas.

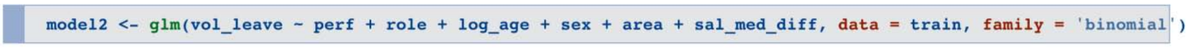
This is consistent with our descriptive analysis above. Again, to keep this primer focused, we will just leave the Area values in as they are. In principle, though, one might wish to simply code area as just two categories (Sales v. Non-Sales) to reduce the model complexity.

The Area X Sex Interaction Variable

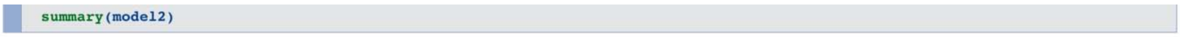
Finally, let’s look at our interaction terms. Each term tells us whether being male impacts the contribution of area to retention. Across the board, we see that none of these are significant.

We will therefore drop this interaction term from our final model because it is not helping us predict which employees will leave and which will stay.

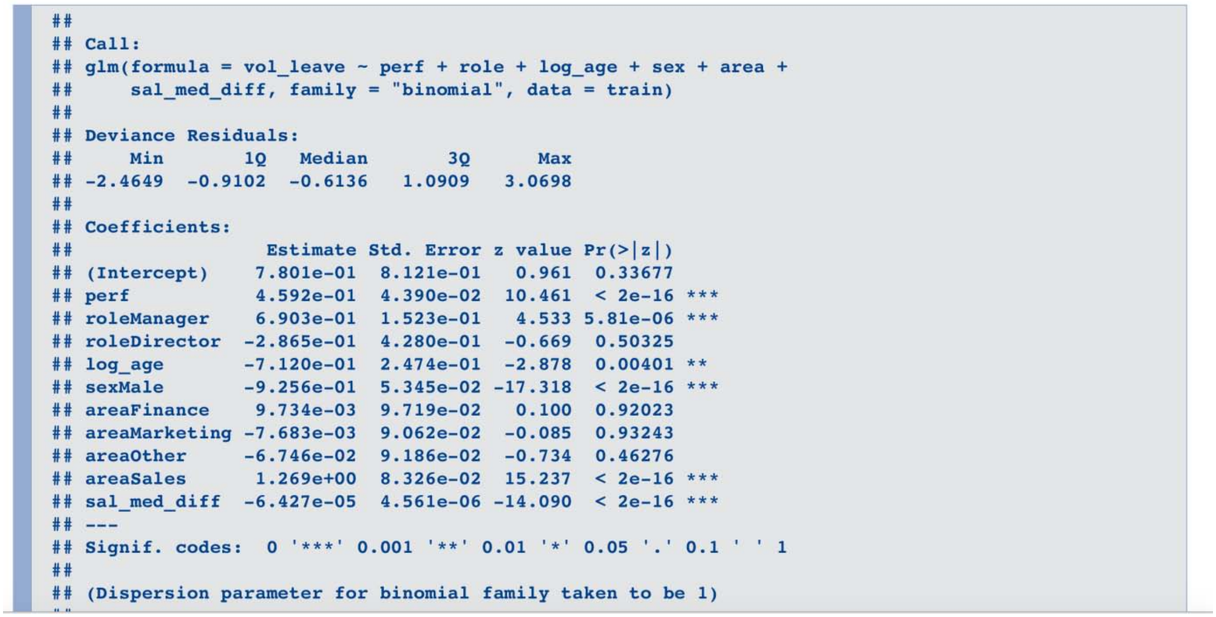
- FINAL LOGISTIC MODEL



```
model2 <- glm(vol_leave ~ perf + role + log_age + sex + area + sal_med_diff, data = train, family = 'binomial')
```



```
summary(model2)
```



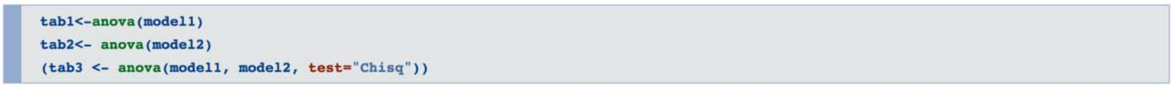
```
##
## Call:
## glm(formula = vol_leave ~ perf + role + log_age + sex + area +
##     sal_med_diff, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4649  -0.9102  -0.6136   1.0909   3.0698
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    7.801e-01  8.121e-01   0.961  0.33677
## perf           4.592e-01  4.390e-02  10.461  < 2e-16 ***
## roleManager    6.903e-01  1.523e-01   4.533 5.81e-06 ***
## roleDirector  -2.865e-01  4.280e-01  -0.669  0.50325
## log_age       -7.120e-01  2.474e-01  -2.878  0.00401 **
## sexMale       -9.256e-01  5.345e-02 -17.318  < 2e-16 ***
## areaFinance    9.734e-03  9.719e-02   0.100  0.92023
## areaMarketing -7.683e-03  9.062e-02  -0.085  0.93243
## areaOther     -6.746e-02  9.186e-02  -0.734  0.46276
## areaSales      1.269e+00  8.326e-02  15.237  < 2e-16 ***
## sal_med_diff  -6.427e-05  4.561e-06 -14.090  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

- INTERPRETING MODEL OUTCOME

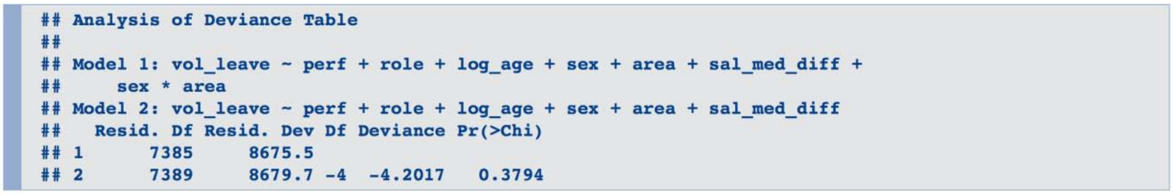
With our final model determined, it’s time to actually see the predicted output from the fitted values in the model output. The fitted qualities give the anticipated likelihood of leaving for every person.

This individual-level information is truly valuable in this present reality since we can then compute probabilities of leaving for any set of arbitrarily defined groupsafter creating out model. For instance, we can aggregate our expectation to anticipate "Females in Sales", "Directors in Accounting", or even "low performing guys under 30 in Sales".

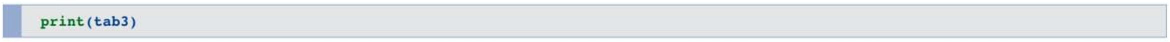
- COMPARING TWO MODELS



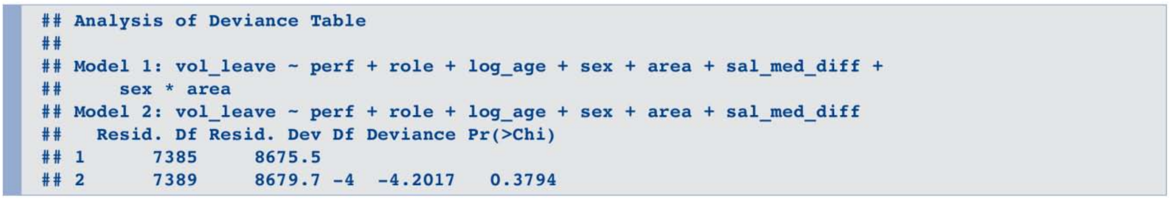
```
tab1<-anova(model1)
tab2<- anova(model2)
(tab3 <- anova(model1, model2, test="Chisq"))
```



```
## Analysis of Deviance Table
##
## Model 1: vol_leave ~ perf + role + log_age + sex + area + sal_med_diff +
##     sex * area
## Model 2: vol_leave ~ perf + role + log_age + sex + area + sal_med_diff
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7385     8675.5
## 2      7389     8679.7 -4  -4.2017   0.3794
```



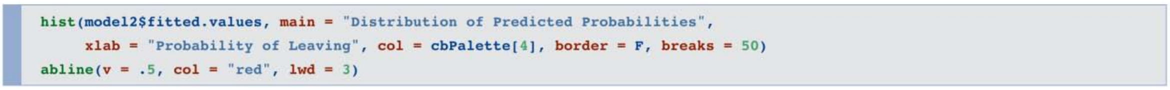
```
print(tab3)
```



```
## Analysis of Deviance Table
##
## Model 1: vol_leave ~ perf + role + log_age + sex + area + sal_med_diff +
##     sex * area
## Model 2: vol_leave ~ perf + role + log_age + sex + area + sal_med_diff
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7385     8675.5
## 2      7389     8679.7 -4  -4.2017   0.3794
```

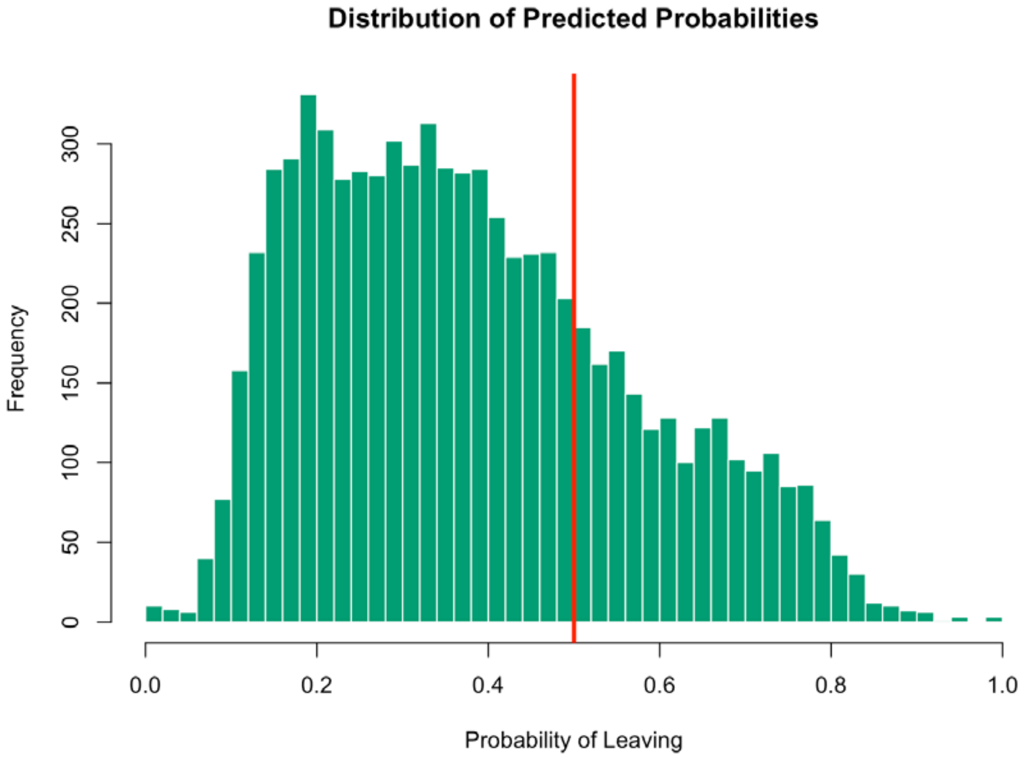
The final logistic model with interaction term has reduced the deviance by 2.4939 at the expense of 4 degrees of freedom and that is why we are going forward with the final logistic model.

- RUNNING THE MODEL ON TRAINING DATA

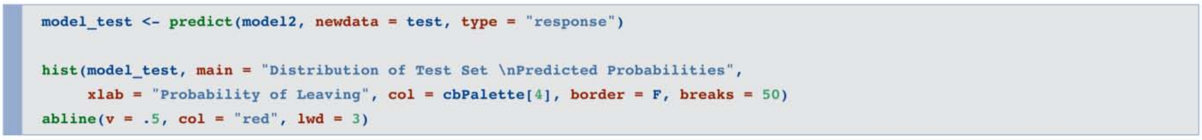


```r
hist(model2$fitted.values, main = "Distribution of Predicted Probabilities",
     xlab = "Probability of Leaving", col = cbPalette[4], border = F, breaks = 50)
abline(v = .5, col = "red", lwd = 3)
```

Our histogram (and calculations) for the training data show us that we have roughly 25% of our employees with a probability of leaving at 50% or higher and 60% of our employees with a probability of leaving greater than 30%.

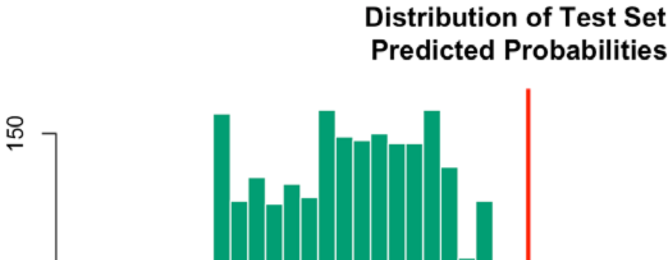


- RUNNING THE MODEL ON TEST DATA

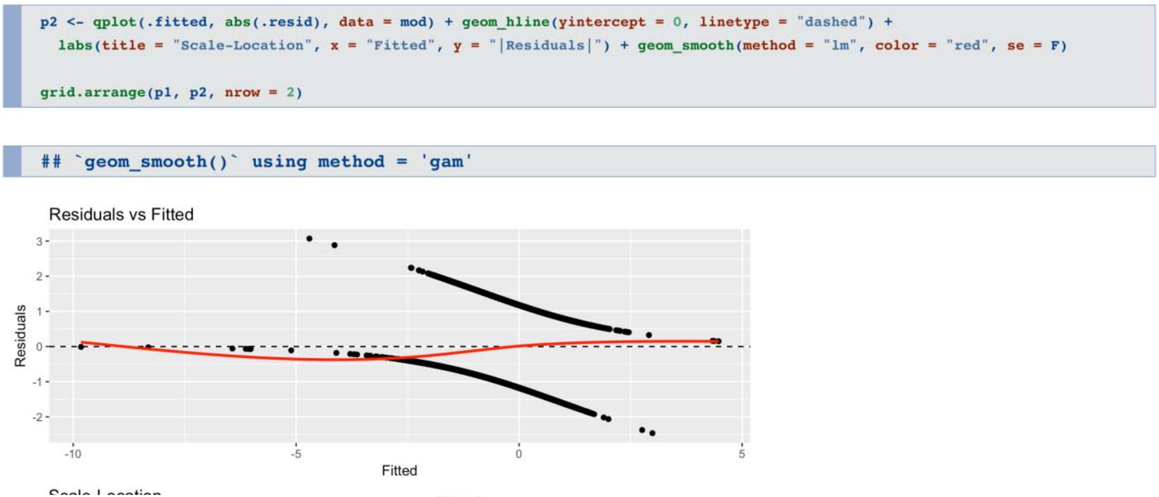


```r
model_test <- predict(model2, newdata = test, type = "response")

hist(model_test, main = "Distribution of Test Set \nPredicted Probabilities",
     xlab = "Probability of Leaving", col = cbPalette[4], border = F, breaks = 50)
abline(v = .5, col = "red", lwd = 3)
```



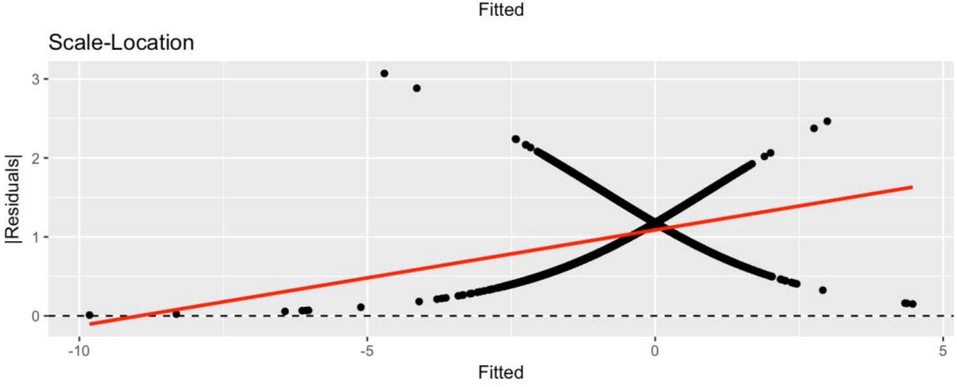
Similar to our training set predictions, we see that 27% have a predicted probability greater than .5 and 63% with a predicted probability of greater than .3. Note again though that the model has provided the individual probabilities, not a 0/1 response.



```
p2 <- qplot(.fitted, abs(.resid), data = mod) + geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Scale-Location", x = "Fitted", y = "|Residuals|") + geom_smooth(method = "lm", color = "red", se = F)

grid.arrange(p1, p2, nrow = 2)
```

```
## `geom_smooth()` using method = 'gam'
```



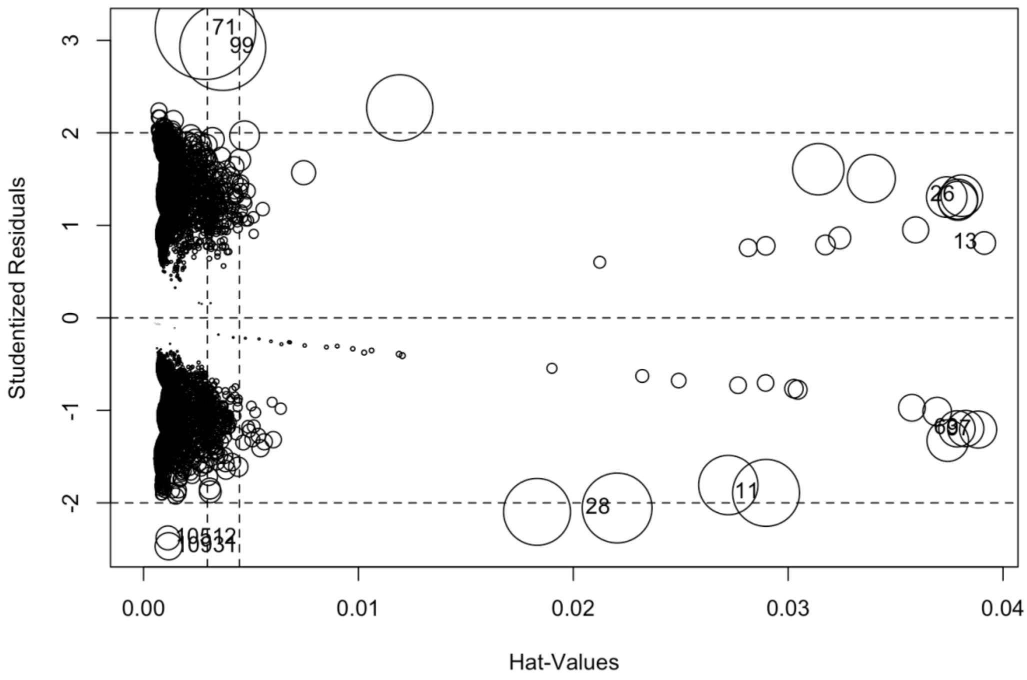
- TEST FOR HETERSKEDASTICITY

We can see the red line is not flat and there is definitely heterskedasticity (non constant variance).

- DETERMINING INFLUENTIAL OUTLIERS

```
influencePlot(model2, id.method=cooks.distance(model2), id.n=4)
```

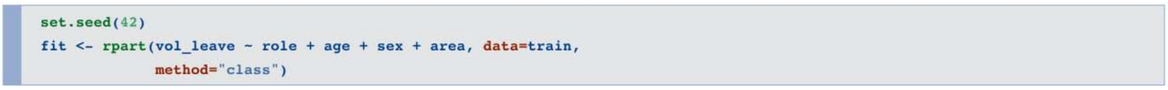
```
## Warning in if (id.method != "identify") {: the condition has length > 1 and
## only the first element will be used
```



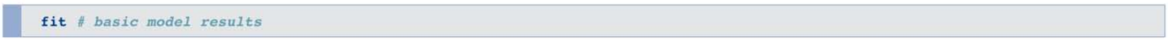
The diameter of the circle determines its influence on the data. Hence the outliers number 71,99 are the most impactful on the given data. The outliers number 28, 11 insignificantly impact data.

THE DECISION TREE MODEL

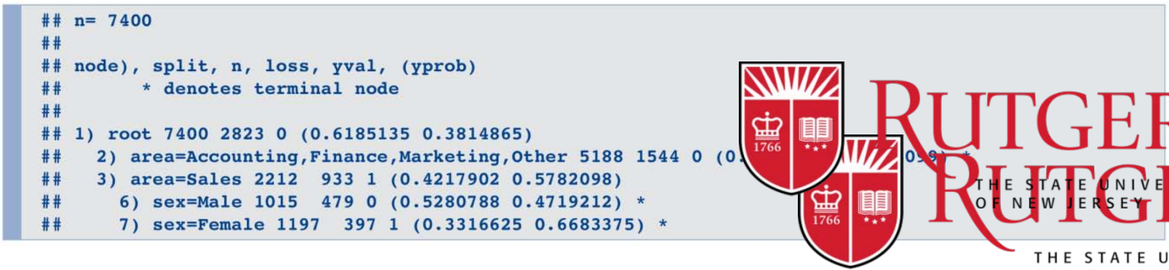
Decision Trees (DT) make everything easier by telling us which predictors are useful. More technically, decision trees are robust to issues like predictors with skewed distributions because the underlying math does not depend on estimating parameter values and having nice, normally distributed predictors. we will create our model using the training data and then use the resulting model to make predictions with our test data.



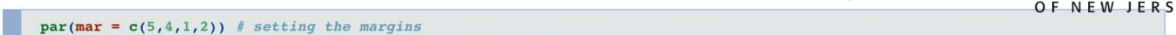
```
set.seed(42)
fit <- rpart(vol_leave ~ role + age + sex + area, data=train,
             method="class")
```



```
fit # basic model results
```



```
## n= 7400
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 7400 2823 0 (0.6185135 0.3814865)
##   2) area=Accounting,Finance,Marketing,Other 5188 1544 0 (0.[illegible]09) *
##   3) area=Sales 2212  933 1 (0.4217902 0.5782098)
##     6) sex=Male 1015  479 0 (0.5280788 0.4719212) *
##     7) sex=Female 1197  397 1 (0.3316625 0.6683375) *
```



```
par(mar = c(5,4,1,2)) # setting the margins
```

- INTERPRETING THE MODEL OUTPUT

The first node is alluded to as the root. The "0" alludes to the dominate case. Here, 62% of those in our training data have 0 (stayers) for the response variable and 38% have a 1 (leavers).

Below that, we see our first decision node. In the event that our workers are in the Accounting, Finance, Marketing, or Other regions, then we say "yes" and take the left branch. On the off chance that the answer is no (i.e. they are in deals), then we take the right branch.

After the left branch, we see that it ends into a solitary node. Think of this node like a bucket for all of those who are not in Sales.For all of these people, the most common (mode) response is “0”, with 70% people as stayers and only 30% in this bucket leaving. The “70%” reported in the bottom of the node tells us that this single bucket accounts for 70% of the total sample we are modeling.

On following the right branch,we see here that the most well-known (mode) reaction is "1" for the leavers. We know this in light of the fact that there is a "1" at the highest point of the node and it is shaded blue rather than green. Moreover, the node is likewise letting us know 42% of those in this bucket are stayers while 58% are leavers.

At long last, the node lets us know that 30% of the aggregate test data we are modeling falls into this group. The 30% here in addition to the 70% from the ending left node gives us 100% of this training test.
"Directors in Accounting", or even "low performing guys under 30 in Sales".

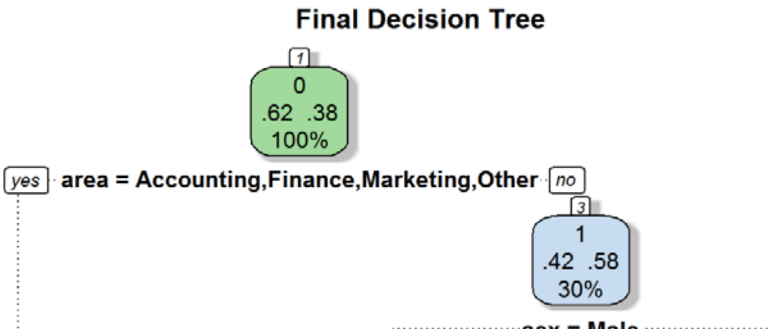
Proceeding with the rightnode, we now get another decision point. This time, if the worker is a male, we say "yes" and go to left side. On the off chance that the worker is female, we go to right side.

For the guys, we wind up in a terminating node (light green) with a prevailing 0 response (stay), 53% to 47% for those in this bucket. This bucket represents 14% of the aggregate example in our training set.

For females, we wind up in an terminatingnode that has a dominant response of 1 (33% stayers v. 67% leavers). This ending node represents 16% of the aggregate populace.

See that including the greater part of the last node rates gives us 100%. Each terminating node then lets us know initially what number of individuals are in what buckets.

- **FINAL DECISION TREE AND OUTPUT**



There a couple of things to observe about the bigger model. In the first place, despite the fact that we incorporated a few more predictors, just the difference from mediansalary wound up being held in the last model created.

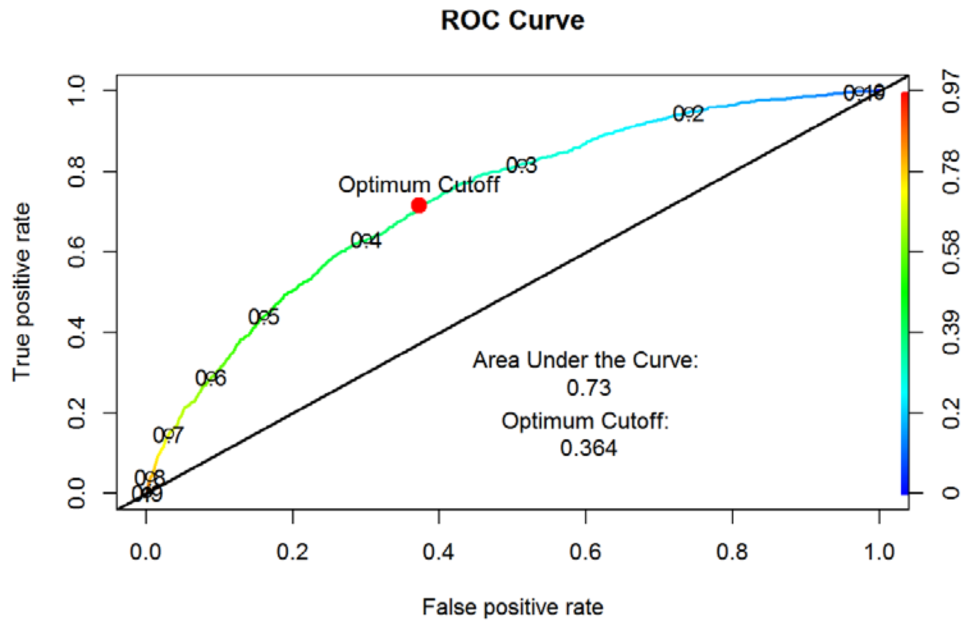
Furthermore, take note of that this predictor just develops AFTER the "yes, male"decision node. You can think about this similar to a type of association in which the effect of salary with respect to the rolemedian mattered only for the males in this simulated information.

In this model, if that distinction is more prominent than or equivalent to -$1023 dollars, we answer"yes" and branch off to left side and end up in a 0 dominated node. On the off chance that the salary difference is more than $1023 below the role median, we branch right and wind up in a mode commanded by leavers (42% stay, 58% leave).Those earning significantly less than the median for their role are seeking greener pastures.

- PREDICTING WITH NEW DATA

We will use the predict function, taking our final decision tree model and apply it to the test data.Our accuracy is about 68%, in line with that from the logistic regression model.

ROC CURVES



- <span>LOGISTIC REGRESSION ROC CURVE</span>

The above roc curve plot depicts three important things:

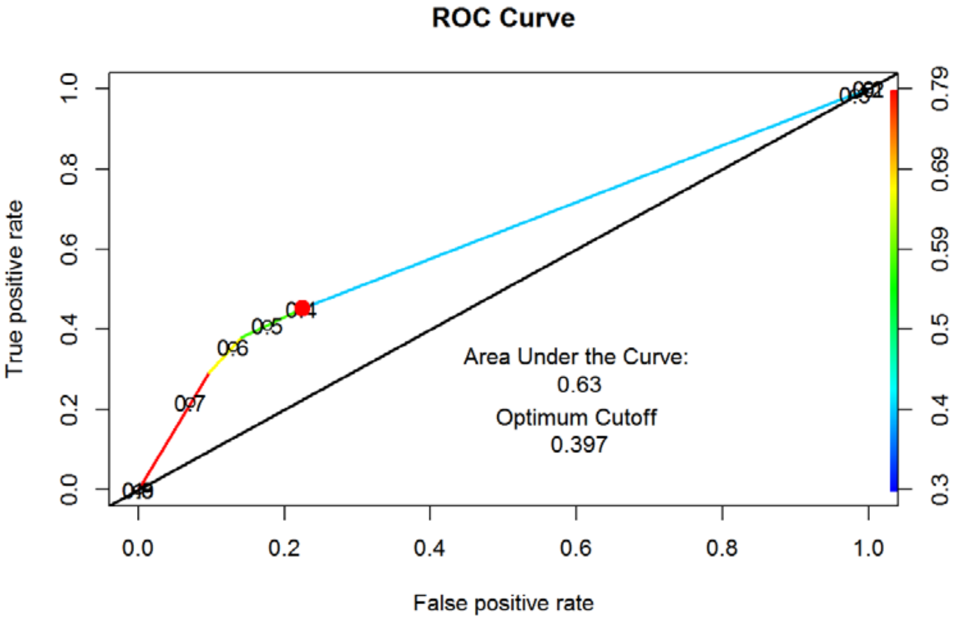
1. The model is certainly superior to anything the no data reference point since it curvessignificantly over the diagonal line.

2. The most ideal cutoff value is .364. That value is the point along the curve that is nearest to the perfect (0,1) area in the upper left.

To get the best tradeoff between False Positives and True Positives, we would classify everybody beneath .364 as a "0" (stayer) and everybody at or above it as 1 (leaver).

The ideal model has an AUC of 1 and a non-predictive model embracing the askew line would have an AUC of .5 (since it slices our space down the middle).Our logistic regression model falls into the fair (read: “Not that bad”) category.

DECISION TREE ROC CURVE



The ROC curve for the decision tree is not remotely smooth like the logistic regression ROC curve. Instead it has just has a few sharp cuts.

Our tree model, on the other hand, only produced 4 probability values, one for each of the possible buckets in our final model.

If we look at our pretty tree plot again, we see employees can only go into four possible buckets; their individual probabilities of leaving are set by the probability of leaving for their specific bucket. There are only 4 buckets so we only have 4 possible cutoffs to choose from.

Conclusion From ROC Curves

Putting it all together see that while overall accuracy of our two models were similar, the ROC curve analysis revealed a clear and strong difference. Our logistic regression model is better.

CONCLUSION

To play a more important and vital part in the organization, the HR function needs to move past beyond mere reporting to precise expectation. Rather than simply creating receptive reports, it needs to grasp advancedanalytics and predictive techniques that bolster key organizational objectives.

Utilization of predictive analysis in HR involves using important information to take care of particular business issues. Predictive analyses helps associations contain HR-related expenses while building up a high performing workforce. The bits of knowledge inferred can enhance business execution and additionally employee engagement and satisfaction.

Predictive Analysis may be an unknown territory for HR, therefore to fully realize its benefits, HR personnel need to collaborate with other business units and customer-facing functions to understand how they leverage data and analytics to create value. By doing so, HR departments can facilitate superlative employee experiences that lead to sustained long-term benefits for the organization.