PROJECT REPORT - HUMAN RESOURCE ANALYTICS

PROBLEM STATEMENT:

The specific goal here is to predict whether an employee will stay or 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.

DATA:

This dataset has 8 unique attributes and 11100 observations. The description of each attribute is mentioned below:

- Role - Specifies the designation of the employee (CEO, VP, Directors, Managers, and Individual Contributors)
- Performance - Performance scale of an employee varies from 1 (lowest) to 3 (highest)
- Area - Business department of an organization namely Sales, Finance, Accounting, Marketing and Others
- Sex - Refers to the gender of the employee, Male or 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 - Based on historical data. '0' = stay and '1' = voluntarily leave the organization

```
summary(HRAnalytics)
```

```
##       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
```

Analysis:

- The summary information lets us know that we have 5 fundamental roles: CEO, Director, Individual Contributors, Manager and VP.
- Since CEOs and VPs encounter an altogether different labor market than the Directors, Managers, and Individuals, incorporating them in our modeling doesn't bode well.
- Resetting the data

```
HRAnalytics = filter(HRAnalytics, HRAnalytics$role == "Ind" | HRAnalytics$role == "Manager" |
HRAnalytics$role == "Director")
HRAnalytics$role <- factor(HRAnalytics$role)
summary(HRAnalytics)
```

```
##       role          perf             area          sex
##  Director:  100   Min.   :1.000   Accounting:1607   Female:6064
##  Ind     :10000   1st Qu.:2.000   Finance   :1676   Male  :5036
##  Manager : 1000   Median :2.000   Marketing :2255
##                   Mean   :2.198   Other     :2197
##                   3rd Qu.:3.000   Sales     :3365
##                   Max.   :3.000
##       id             age            salary         vol_leave
##  Min.   :   12   Min.   :22.02   Min.   : 42168   Min.   :0.0000
##  1st Qu.: 2787   1st Qu.:24.07   1st Qu.: 57080   1st Qu.:0.0000
##  Median : 5562   Median :25.70   Median : 60788   Median :0.0000
##  Mean   : 5562   Mean   :27.77   Mean   : 64860   Mean   :0.3815
##  3rd Qu.: 8336   3rd Qu.:28.48   3rd Qu.: 64928   3rd Qu.:1.0000
##  Max.   :11111   Max.   :61.67   Max.   :311131   Max.   :1.0000
```

VISUALIZATION:

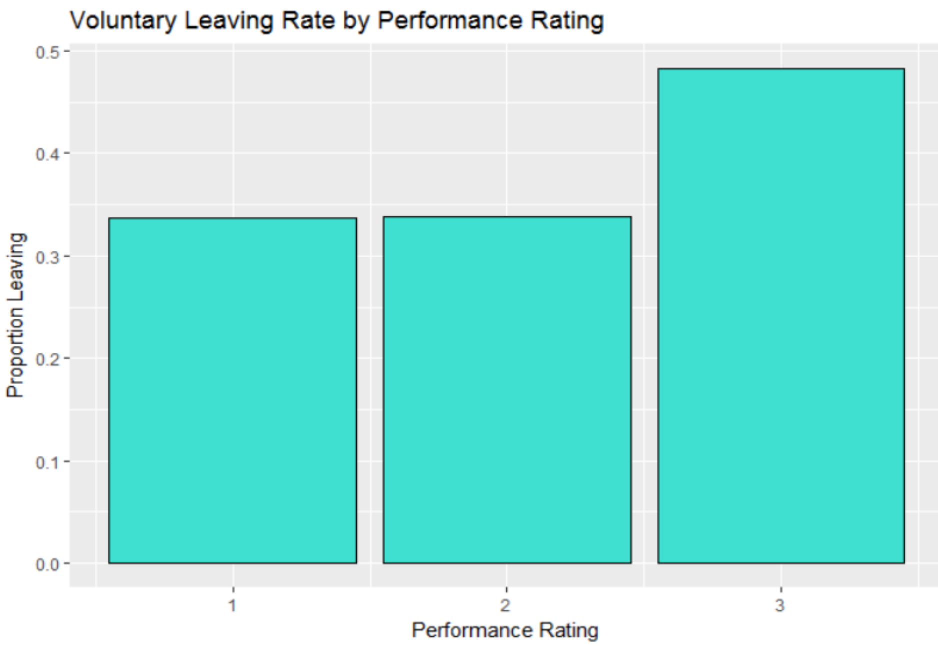
As the response output variable consist of two groups (0, 1), comparing it with other columns would be much easier if we use aggregate along with the mean function.

(a) Performance v/s Voluntarily Leaving

```
performance_agg = aggregate(vol_leave ~ perf, data = HRAnalytics, mean)
performance_agg
```

```
##   perf vol_leave
## 1    1 0.3375112
## 2    2 0.3383831
## 3    3 0.4831122
```

Plotting the graph for the same



Analysis:

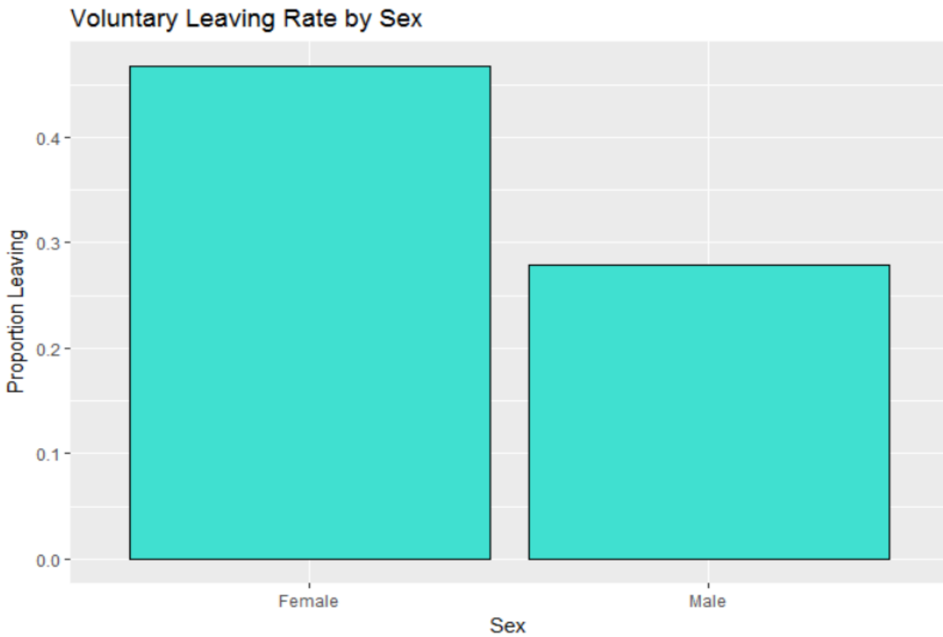
- Employees with performance rating 3 are likely to leave the company next year.

(b) Sex v/s Voluntarily Leaving

```
sex_agg = aggregate(vol_leave ~ sex, data = HRAnalytics, mean)
sex_agg
```

```
##      sex vol_leave
## 1 Female 0.4673483
## 2   Male 0.2781970
```

Plotting the graph for the same



Analysis:

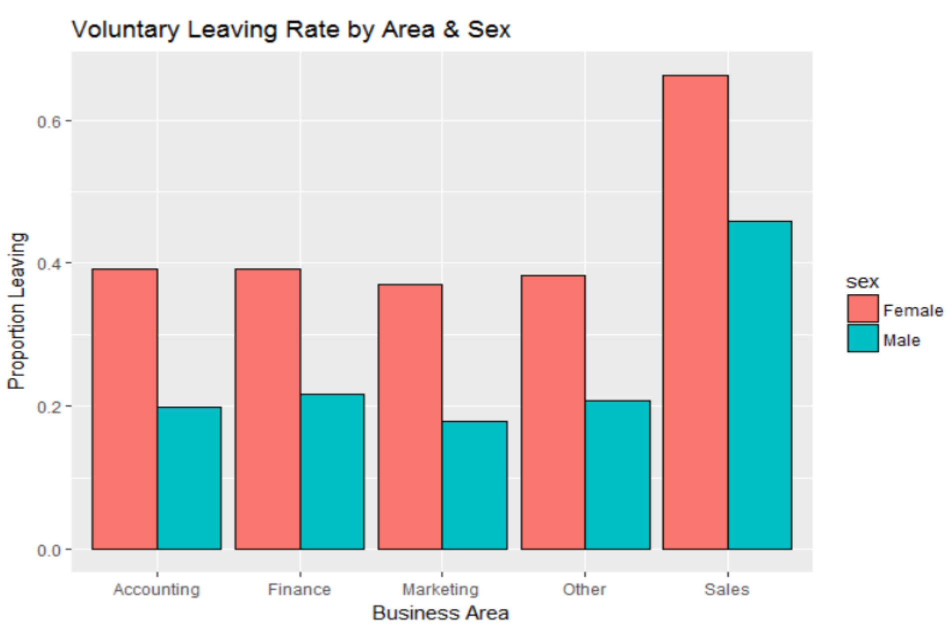
- Female attrition rate is higher than the males in the entire organization
- Females are more prone to voluntarily leaving the company

(c) Business Area and Gender v/s Voluntarily Leaving

```
area_sex_agg = aggregate(vol_leave ~ area + sex, data = HRAnalytics, mean)
area_sex_agg
```

```
##          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
```

Plotting the graph for the same



Analysis:

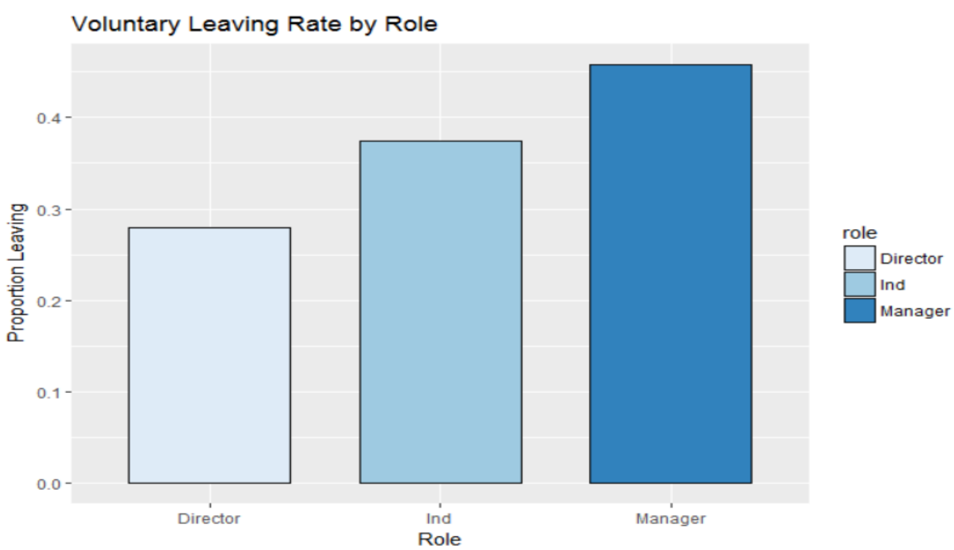
- Voluntary termination is higher in females.
- Under sales department, all employees are nearly unhappy
- People working in Sales department are much more likely to leave the job reason being: Most sales jobs are paid less and mundane, No fixed working hours, Work timing extends to late nights as well.
- Whereas, people working in Marketing department are likely to stay

(d) Role v/s Voluntarily Leaving

```
role_agg = aggregate(vol_leave ~ role, data = HRAnalytics, mean)
role_agg
```

```
##       role vol_leave
## 1 Director    0.2800
## 2      Ind    0.3749
## 3  Manager    0.4580
```

Plotting the graph for the same



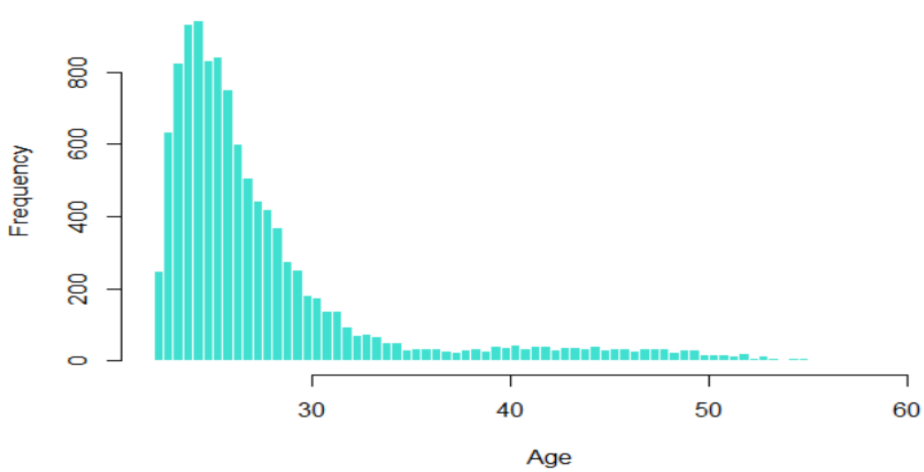
Analysis:

- Managers have higher attrition rate. Directors have a longer run at a company

(e) Analyzing the age of the employee

```
quantile(HRAnalytics$age)
```

```
##       0%      25%      50%      75%     100%
## 22.02289 24.07050 25.69533 28.48035 61.67132
```



Analysis:

- Skewness is present here with half of our workforce somewhere around 22 and 26 years old.
- However there are 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.

Since the variable age is skewed, we will take the log of age while fitting a model.

```
HRAnalytics$log_age <- log(HRAnalytics$age)
summary(HRAnalytics$log_age)
```

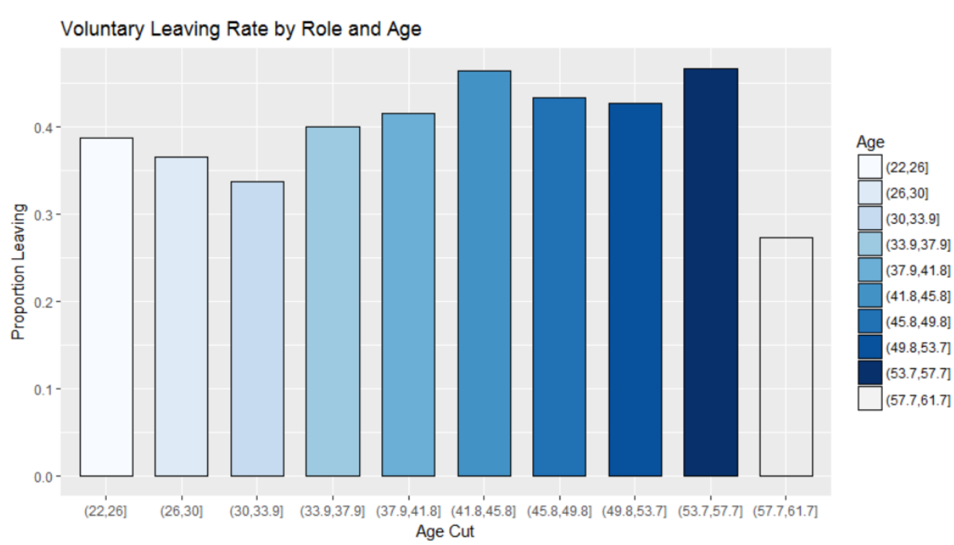
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.092   3.181   3.246   3.304   3.349   4.122
```

Segmenting age variable even further to get proper insights

```
age_agg = aggregate(x = HRAnalytics$vol_leave, by = list(cut(HRAnalytics$age, 10)), mean)
age_agg
```

```
##        Group.1         x
## 1      (22,26] 0.3866177
## 2      (26,30] 0.3645902
## 3    (30,33.9] 0.3374536
## 4  (33.9,37.9] 0.3992806
## 5  (37.9,41.8] 0.4155405
## 6  (41.8,45.8] 0.4640288
## 7  (45.8,49.8] 0.4333333
## 8  (49.8,53.7] 0.4260870
## 9  (53.7,57.7] 0.4666667
## 10 (57.7,61.7] 0.2727273
```

Plotting the graph for the same

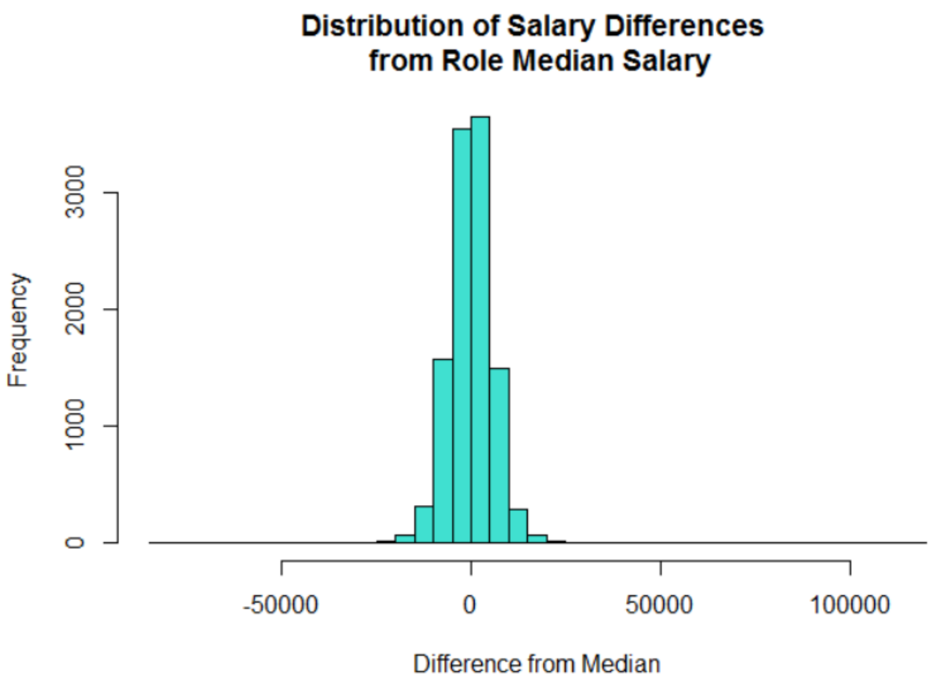


Analysis:

- This shows that employee within 34 - 54 age groups will leave the company more likely than the ones within 22 - 34 who might be individual employees.
- Age group of 54 - 62 is at director level and the attrition is least in that age group.

(f) Analyzing salary pattern

Normalizing the salary variable



DATA MODELING:

Before we start creating models, we need to split our data into a training set and a test set. Utilize two-thirds of the data for training and model development and one third of the data for testing the models.

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

We will be using two techniques,

a. Logistic Regression
b. Decision Tree

(a) LOGISTIC REGRESSION

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

(1) Fit the model

```
fit = glm(vol_leave ~ role + perf + area + sex + log_age + salary_diff, data = HRAnalytics, family =
'binomial')
summary(fit)
```

```
##
## Call:
## glm(formula = vol_leave ~ role + perf + area + sex + log_age +
##     salary_diff, family = "binomial", data = HRAnalytics)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4737  -0.9123  -0.6068   1.0906   3.2238
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.581e-01  8.676e-01   0.182 0.855451
## roleInd        6.819e-01  3.456e-01   1.973 0.048495 *
## roleManager    1.393e+00  3.249e-01   4.289  1.8e-05 ***
## perf           4.931e-01  3.598e-02  13.703  < 2e-16 ***
## areaFinance    3.517e-02  7.920e-02   0.444 0.657003
## areaMarketing -9.517e-02  7.490e-02  -1.271 0.203862
## areaOther     -9.540e-05  7.471e-02  -0.001 0.998981
## areaSales      1.239e+00  6.799e-02  18.230  < 2e-16 ***
## sexMale       -9.435e-01  4.374e-02 -21.571  < 2e-16 ***
## log_age       -7.516e-01  2.037e-01  -3.689 0.000225 ***
## salary_diff   -6.515e-05  3.723e-06 -17.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14759  on 11099  degrees of freedom
## Residual deviance: 13004  on 11089  degrees of freedom
## AIC: 13026
##
## Number of Fisher Scoring iterations: 4
```

Analysis:

- First of all, we can see that areaFinance, areaMarketing and areaOther are not statistically significant.
- As for the statistically significant variables, salary, areaSales and perf has the lowest p-value suggesting a strong association of these variables with the probability of leaving the company.

Now we can run the anova() function on the model to analyze the table of deviance.

(2) Chi-Square Test

```
anova(fit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vol_leave
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        11099      14759
## role         2    30.69     11097      14728 2.162e-07 ***
## perf         1   161.14     11096      14567 < 2.2e-16 ***
## area         4   735.02     11092      13832 < 2.2e-16 ***
## sex          1   466.69     11091      13365 < 2.2e-16 ***
## log_age      1    11.21     11090      13354 0.0008158 ***
## salary_diff  1   350.08     11089      13004 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis:

- The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better.
- A smaller p-value here indicates that all the variables in the model are significant

(3) Assessing the predictive ability of the model

Set the parameter type = 'response', R will output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y = 0.

```
fitted.results = predict(fit, test, type = 'response')
fitted.results = ifelse(fitted.results > 0.5,1,0)
misClasificError = mean(fitted.results != test$vol_leave)

# Confusion Matrix
table(actual = test$vol_leave, prediction = fitted.results)
```

```
##       prediction
## actual    0    1
##      0 1919  369
##      1  780  632
```

```
# Accuracy
print(paste('Accuracy', 1 - misClasificError))
```

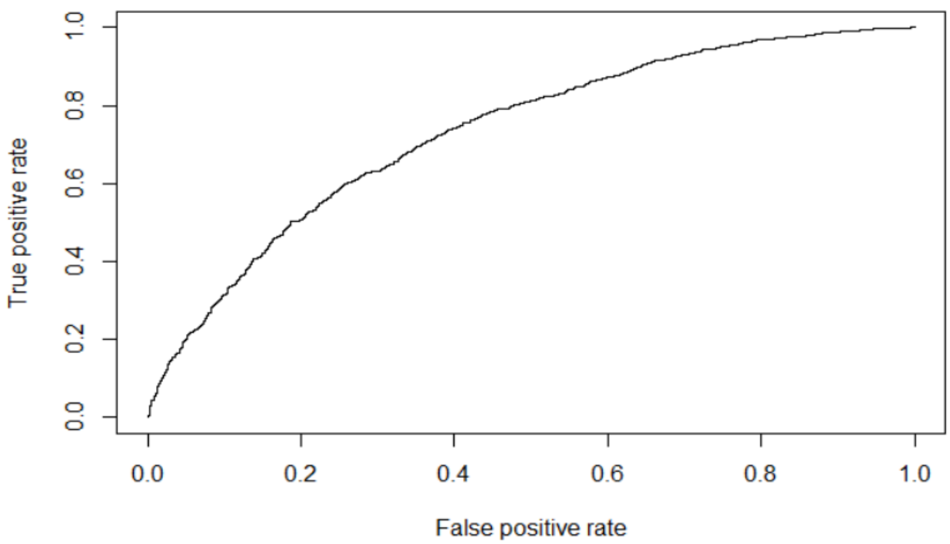
```
## [1] "Accuracy 0.689459459459459"
```

Analysis:

- The accuracy of our model is approximately 68%

(4) ROC & AUC(Area Under Curve)

Plot the ROC curve and calculate the AUC which are typical performance measurements for a binary classifier.




```
auc = performance(pr, measure = "auc")
auc = auc@y.values[[1]]
auc
```

```
## [1] 0.7326298
```

Analysis:

- The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC is the area under the ROC curve.
- As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5.
- Based on the value of AUC for our dataset, we can say that it has good predictive ability.

(b) DECISION TREE

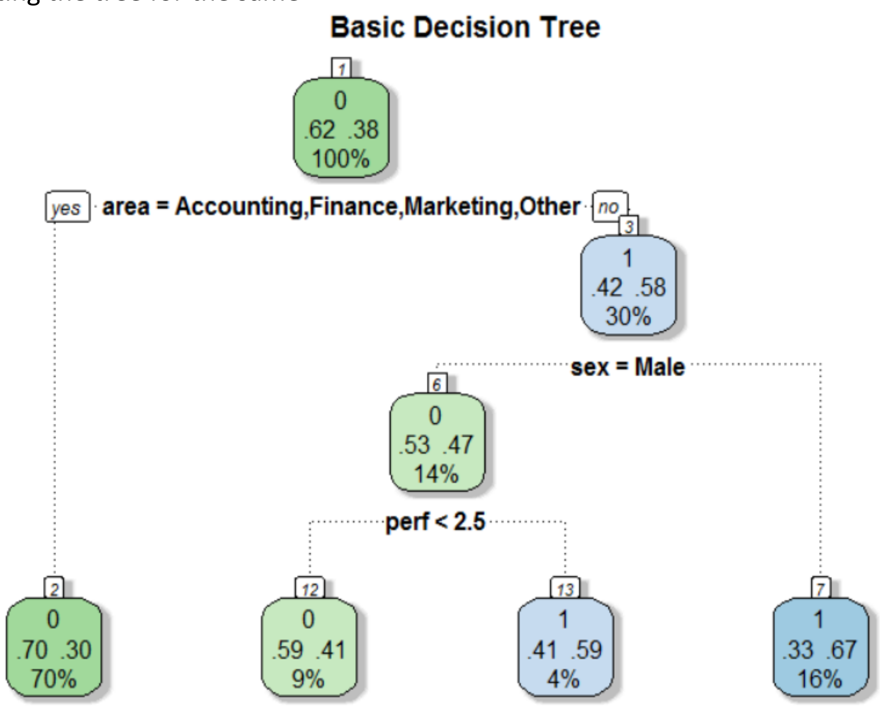
We have already divided our dataset into training and testing. So we proceed further by making the decision tree.

(1) Fit the model

```
set.seed(42)
decision_fit = rpart(vol_leave ~ role + perf + age + sex + area + salary, data = train, method =
"class")
decision_fit
```

```
## 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 (0.7023901 0.2976099) *
##   3) area=Sales 2212  933 1 (0.4217902 0.5782098)
##     6) sex=Male 1015  479 0 (0.5280788 0.4719212)
##      12) perf< 2.5 682  281 0 (0.5879765 0.4120235) *
##      13) perf>=2.5 333  135 1 (0.4054054 0.5945946) *
##     7) sex=Female 1197  397 1 (0.3316625 0.6683375) *
```

Plotting the tree for the same



Analysis:

- The first node is the root. The ‘0’ alludes to the dominate case. Here, 62% of those in our training data have 0 (Stay) for the response variable and 38% have a 1 (Leave).
- 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 else we go right.
- 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 response is ‘0’ (Stay), with 70% employee who will stay in the company and only 30% in this bucket will leave the company. 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 that the most well-known reaction is ‘1’ for the employee who will leave the company. Moreover, the node is likewise letting us know 42% of employees in this bucket will stay while 58% will leave.
- Proceeding with the right branch is further, if the worker is male, we say ‘yes’ and go to the left side. On the off chance that the worker is female, we go right.

- For females, we wind up in a terminating node that has a dominant response of 1 (33% - Stay and 67% - Leave). This ending node represents 16% of the aggregate populace.
- For male, we further go down to performance variable. If the performance is less than 2.5 we go left else we go right.
- For performance less than 2.5, we wind up in a terminating node that has a dominant response of 0 (59% - Stay and 41% - Leave). This ending node represents 16% of the aggregate populace.
- For performance greater than 2.5, we wind up in a terminating node that has a dominant response of 1 (33% - Stay and 67% - Leave). This ending node represents 4% of the aggregate populace.

(2) Assessing the predictive ability of the model

```
t_pred = predict(decision_fit, test, type = 'class')

# Confusion Matrix
confMat = table(actual = test$vol_leave, prediction = t_pred)
confMat
```

```
##       prediction
## actual    0    1
##      0 2006  282
##      1  930  482
```

```
# Accuracy
accuracy = sum(diag(confMat))/sum(confMat)
accuracy
```

```
## [1] 0.6724324
```

CONCLUSION:

- Logistic regression is better than decision tree in predicting the output response variable.
- To play 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 advanced analytics and predictive techniques that bolster key organizational objectives.