**PROJECT REPORT - HUMAN RESOURCE ANALYTICS**

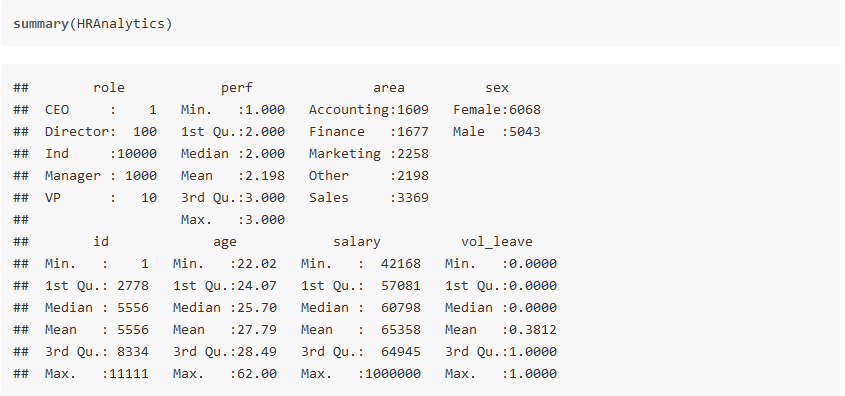
**PROBLEM STATEMENT:**

The specific goal here is to predict whether an employee will stay or x 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:**

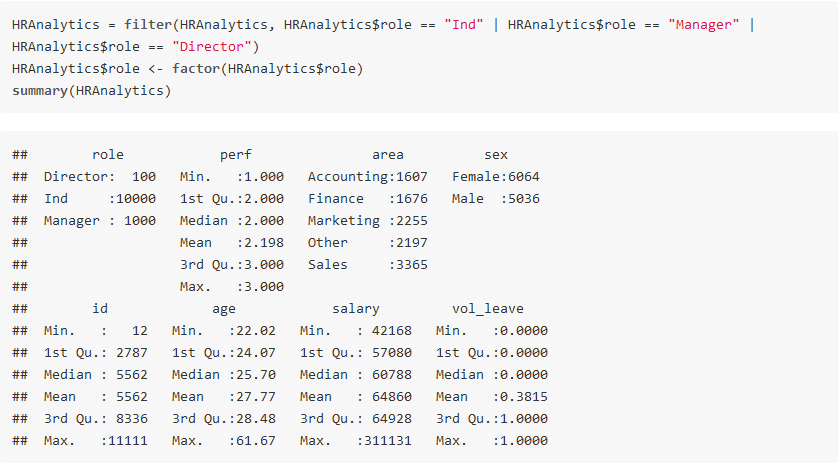
This dataset has 8 unique attributes and 11100 observations. The description of each attribute is mentioned below:

* 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



Analysis:

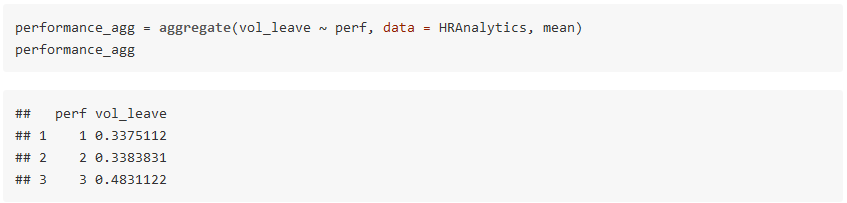
* The summary information lets us know that we have 5 fundamental roles: CEO, Director, Individual Contributors, Manager and VP.
* Since CEOs and VPs encounter an altogether different labor market than the Directors, Managers, and Individuals, incorporating them in our modeling doesn’t bode well.
* Resetting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



**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.

1. **Performance v/s Voluntarily Leaving**



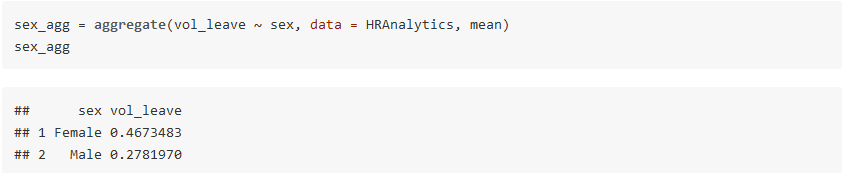
Plotting the graph for the same



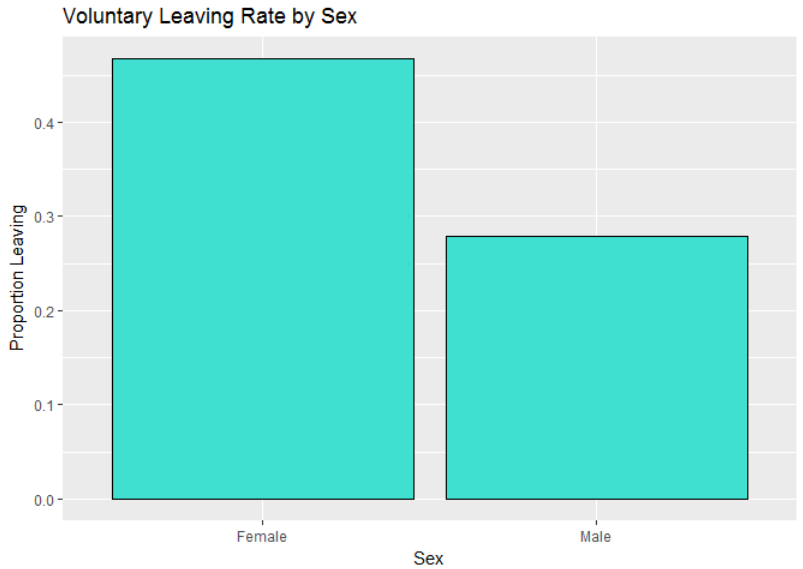
Analysis:

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

1. **Sex v/s Voluntarily Leaving**



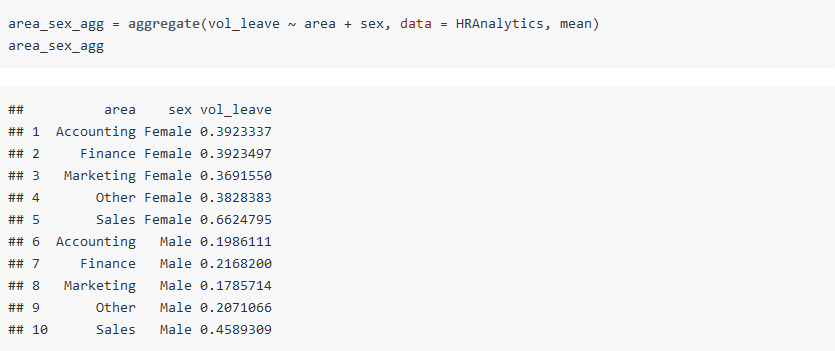
Plotting the graph for the same



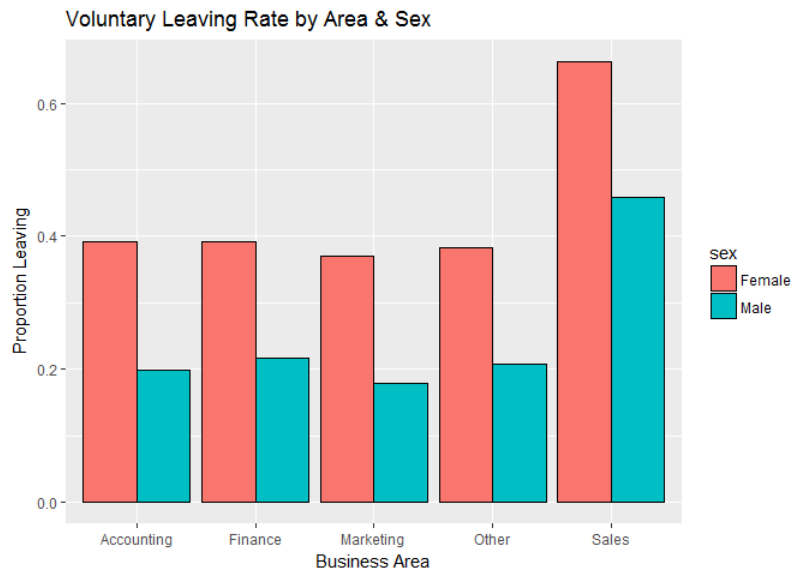
Analysis:

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

1. **Business Area and Gender v/s Voluntarily Leaving**



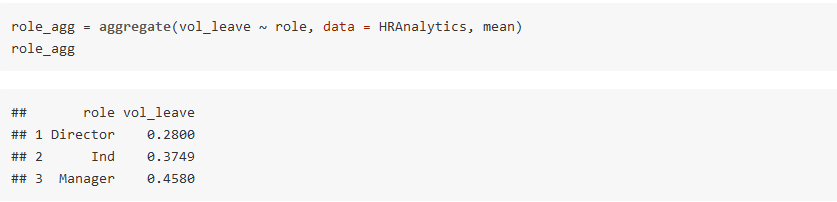
Plotting the graph for the same



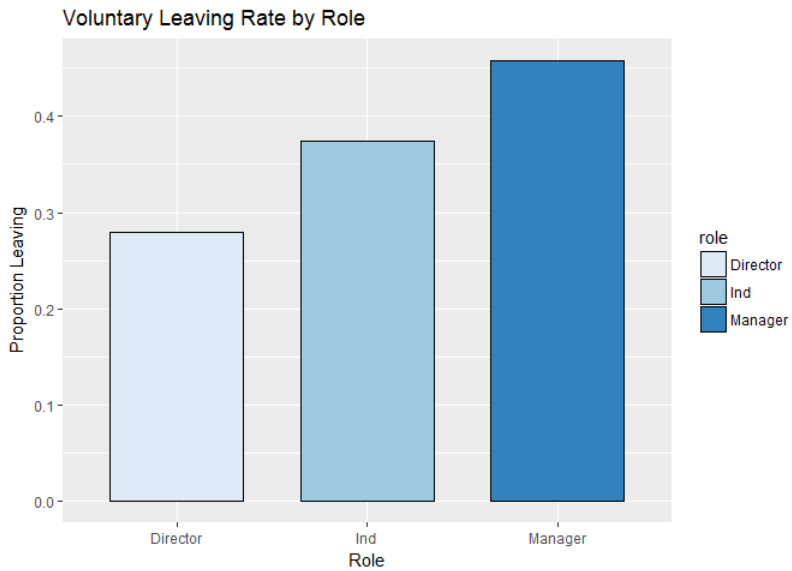
Analysis:

* 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

1. **Role v/s Voluntarily Leaving**



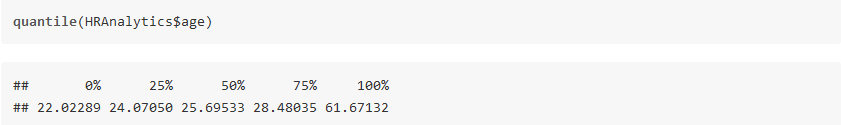
Plotting the graph for the same

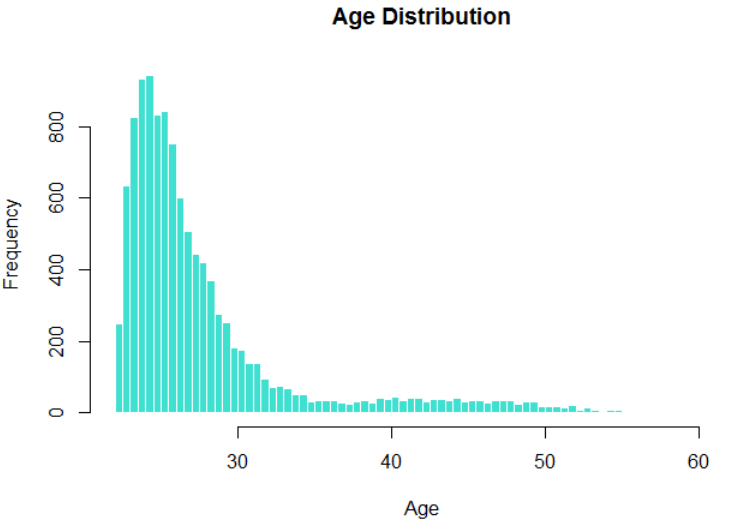


Analysis:

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

1. **Analyzing the age of the employee**

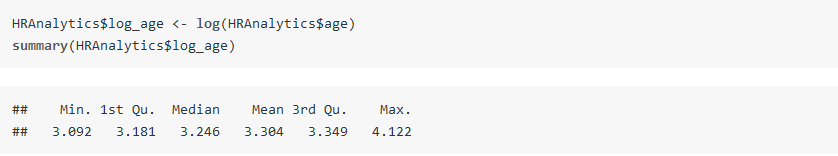




Analysis:

* 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.

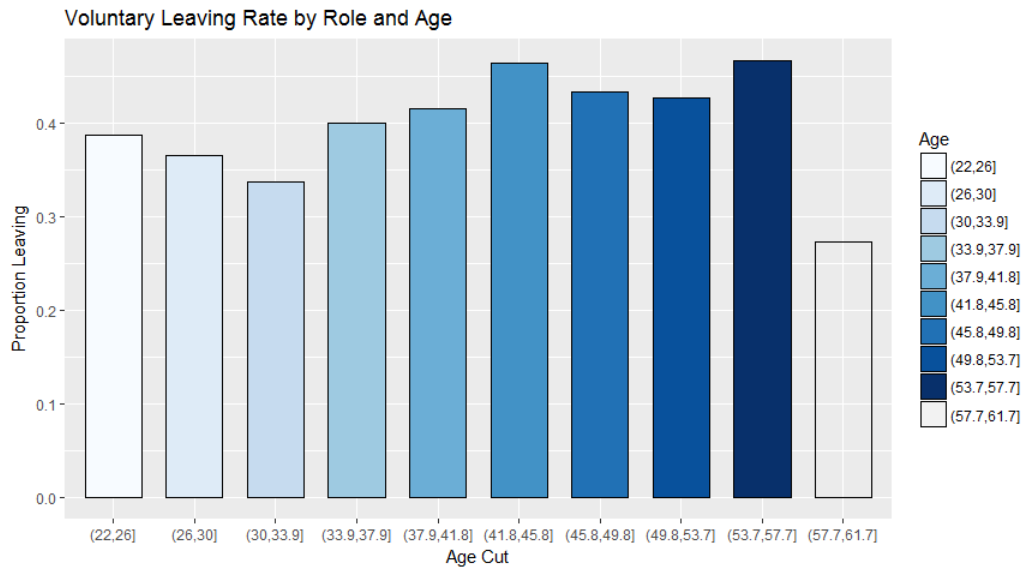
Since the variable age is skewed, we will take the log of age while fitting a model.



Segmenting age variable even further to get proper insights



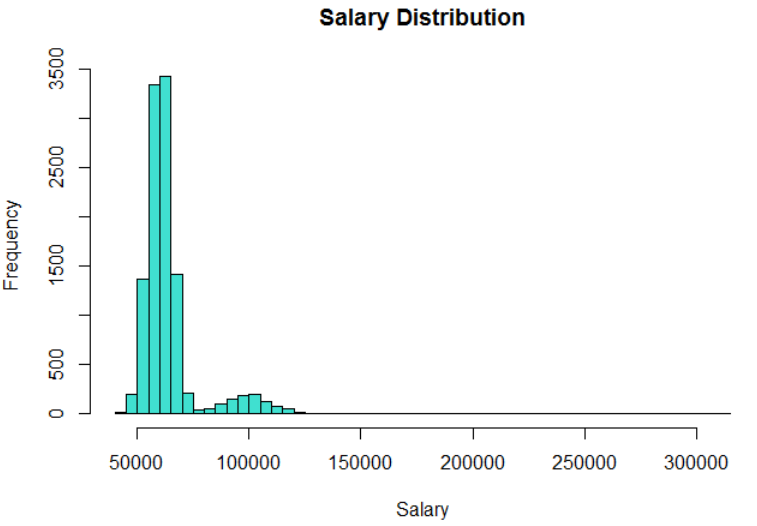
Plotting the graph for the same



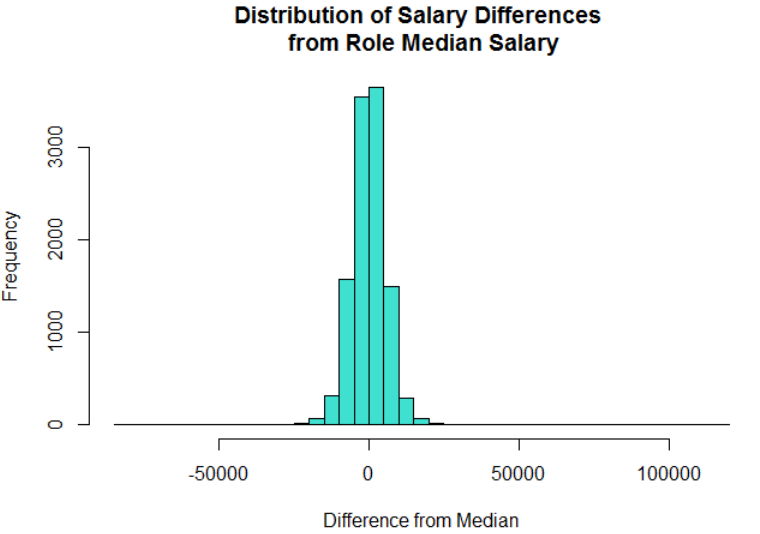
Analysis:

* 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.

1. **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.

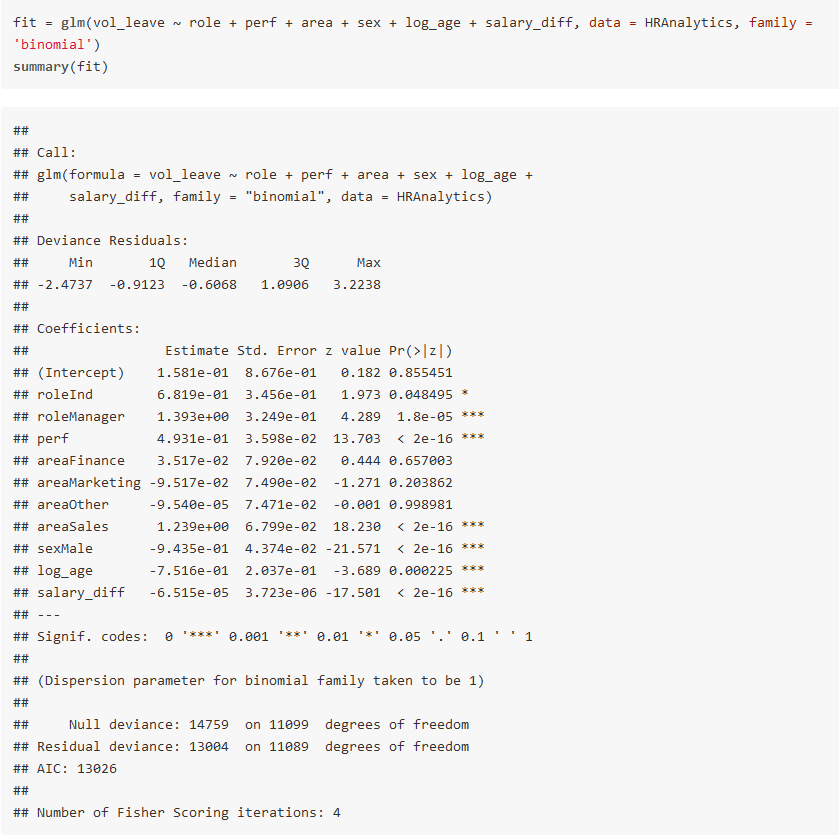


We will be using two techniques,

1. Logistic Regression
2. Decision Tree
3. **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**

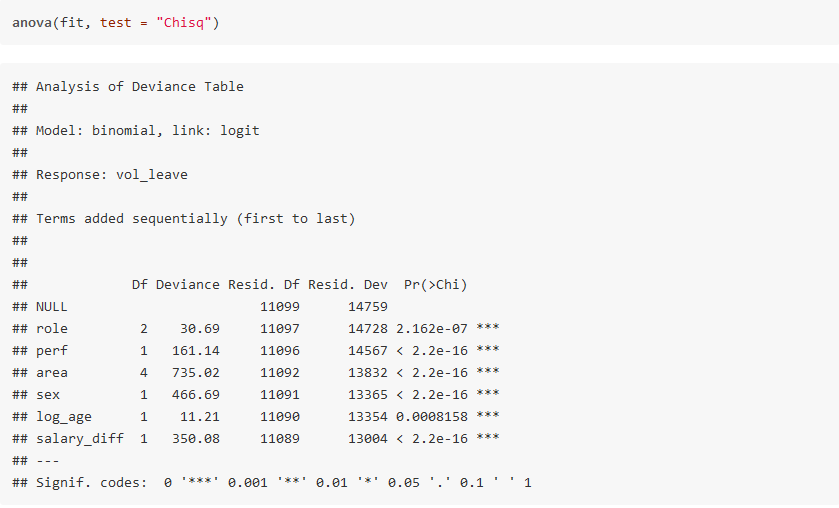


Analysis:

* 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.

1. **Chi-Square Test**



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

1. **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.

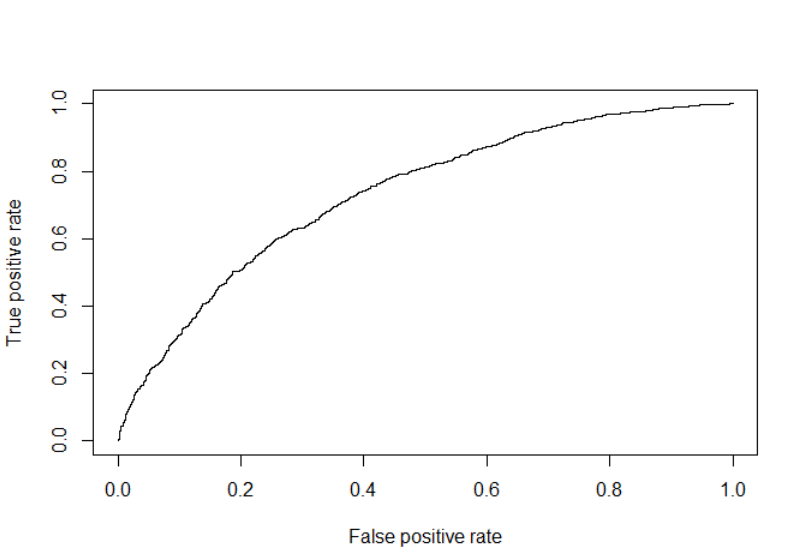


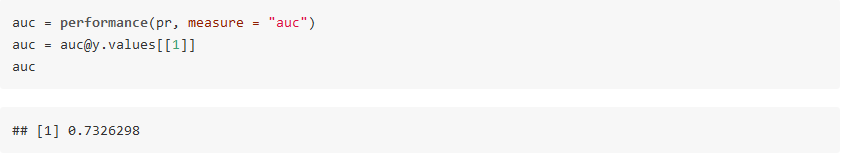
Analysis:

* The accuracy of our model is approximately 68%

1. **ROC & AUC(Area Under Curve)**

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





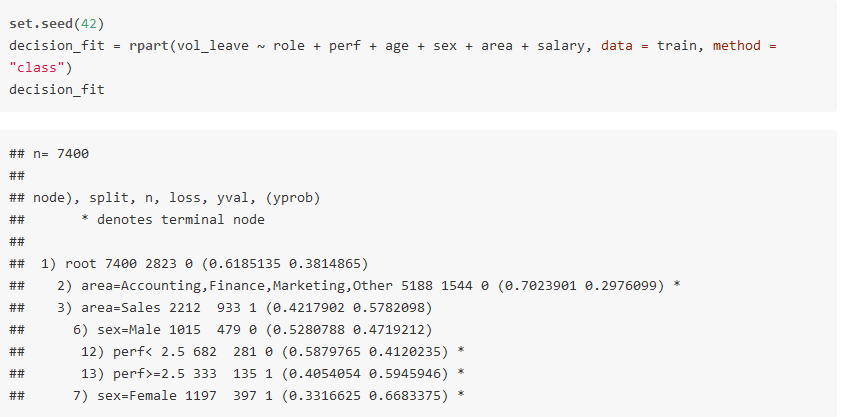
Analysis:

* 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.

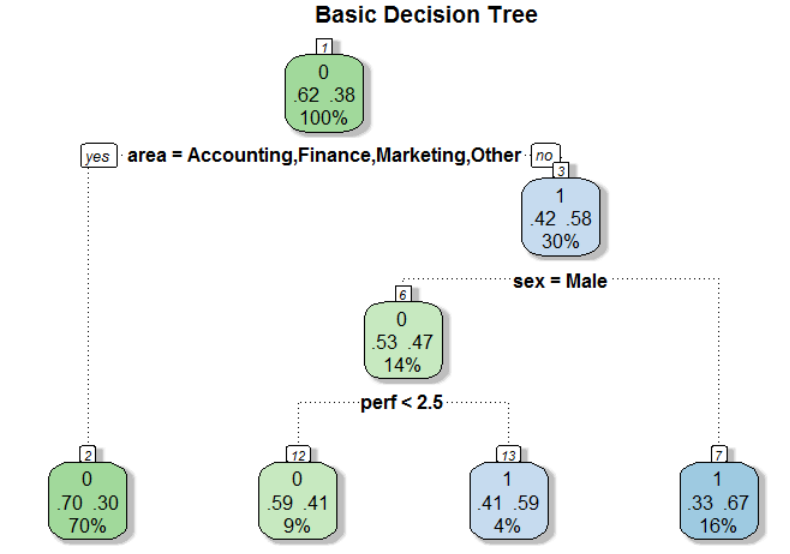
1. **DECISION TREE**

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

1. **Fit the model**



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.

1. **Assessing the predictive ability of the model**



**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.