HR Analytics Model for Organizations

Aditi Sanjay[1], Manan Rao[2]

Abstract: ***Human resource predictive analytics is a developing application field of analytics for HR Management purposes. The reason for HR Management is estimating representative execution and commitment, examining workforce cooperation designs, dissecting worker beat and turnover and demonstrating representative lifetime esteem. The thought process of applying HR analytics is to improve exhibitions and produce better degree of profitability for associations through choice making dependent on information accumulation, HR measurements and predictive models. In this paper an analysis is performed on a sample data and the main goal here is to basically predict whether an employee working in a company will Stay or Leave within the next year.***

Keywords: HR analytics, Analytics, Human Resource Management, Decision Making

1. Introduction

HR analytics is a multidisciplinary way to deal with coordinate system for improving the nature of individuals related choices so as to improve individual and hierarchical execution.[1] There are exchangeable terms utilized for HR analytics are ability analytics, individuals analytics, and workforce analytics. HR analytics assumes a job in each part of the HR work, including selecting, preparing and improvement, progression arranging, maintenance, commitment, pay, what's more, benefits. HR analytics are those that include "top of the line" predictive displaying where consider the possibility that situations conjecture the. outcomes of changing arrangements or conditions. Conventional HR analytics centers around the present, that is, things, for example, turnover and cost per contract. [2] In any case, most associations did not have a predictable and general perspective on the workforce and hence required HR analytics to perform workforce improvement and consequently it got significant for HR to create IT and money systematic abilities and capacities to create better Return on Investment (return for money invested).[3] [4]

Three huge changes that have truly made a want predictive analytics in HR and these are:[5]

1) Major boost in computing power and its affordability
2) HR big data digitally accessible via cloud storage for processing
3) Global talent war to protect and pursue talent streams.

Predictive analytics is not normal for elucidating investigation which considers outer benchmarking information and includes tables, reports, proportions, measurements, dashboards or complex maths; it is about information determined bits of knowledge that drive better choices.[4] It incorporates factual strategies, AI techniques, and information mining models that examine and concentrate existing and chronicled actualities to make expectations. It empowers associations to examine the past and anticipate spot drifts in key components identified with willful end, nonappearances and different sources of hazard. Predictive analytics includes models of hierarchical frameworks for forecast of future results and understand the significances of theoretical changes in associations.[6] Predictive analytics have prompted prescriptive analytics where HR gets choice choices to enhance execution and reshape whole HR Management basic leadership.

2. Methodology

The main 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 stands for stay, 1 stands for leave) using the other columns of data.

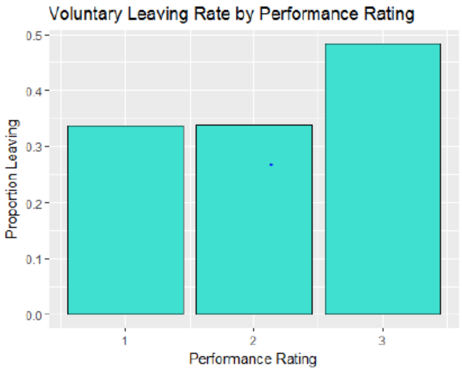
The sample data was collected [7] and saved in a .csv file format and then the analysis had been performed using R studio. You can think about this information as chronicled information which discloses to us who did and who didn't leave inside the most recent year.

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.

3. Visualization

(a) Performance v/s Voluntarily Leaving

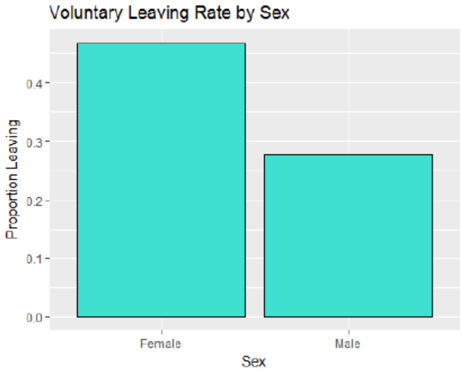
Make a variable that stores the aggregate performance value and Plot a graph using the ggplot function in R studio.



Conclusion: Employees of the company with performance rating of 3 are more likely to leave next year.

(b) Sex v/s Voluntarily Leaving

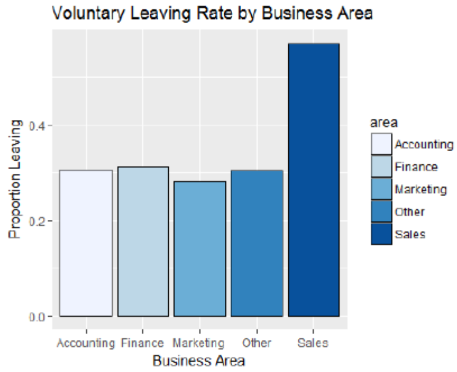
Make a variable that stores the aggregate value and Plot a graph using the ggplot function in R studio.



Conclusion: Female rate of attrition is much higher than the males in the given organization and hence with females are more likely to leave the company next year.

(c) Business Area v/s Voluntarily Leaving

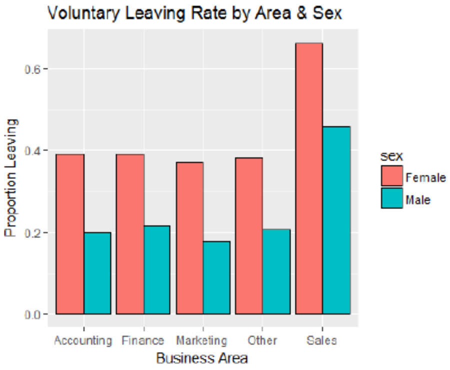
Make a variable that stores the aggregate value and Plot a graph using the ggplot function in R studio.



Conclusion: People from sales are more likely to leave their jobs next year and the contributing factors could be such as the pay is mundane and less and there are no fixed working hours, they can just hop on to other companies with much greater pay.

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

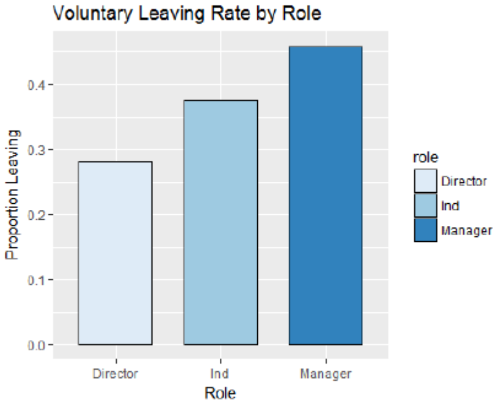
Make a variable that stores the aggregate value and Plot a graph using the ggplot function in R studio.



Conclusion: Females are more likely to leave as compared to men in all the departments and the sales department are the once going to see lot of change going around.

(e) Role v/s Voluntarily Leaving

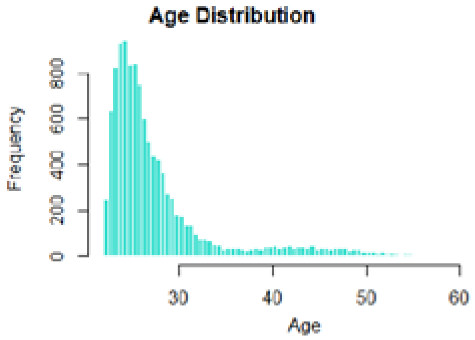
Make a variable that stores the aggregate value and Plot a graph using the ggplot function in R studio.



Conclusion: The attrition rate is higher in Managers. And on the other hand directors stay much longer.

(f) Analyzing the age of the employee

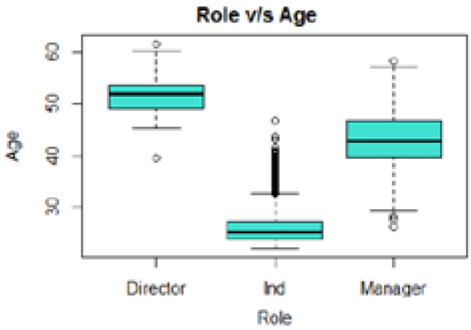
A histogram of age of the employees working in the company is to be analysed. Plot a graph using the hist function in R studio.



Conclusion: The raise or the skewness is seen with half of the employees nearly around 22 to 26 year old employees. But since there are distinct levels of employees. The plot of ages with these distinct levels would be more helpful.

Role vs Age

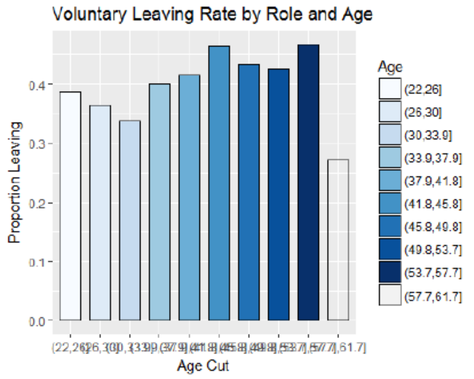
To see the age variation with levels of employees working in the organization a box plot must be more helpful. Plot a graph using the boxplot function in R studio.



Conclusion: Clearly there is a relationship between the roles and the age. Now we can find this relation useful to even check on employees who are going to leave next year.

Voluntary Leaving Rate by Role and Age

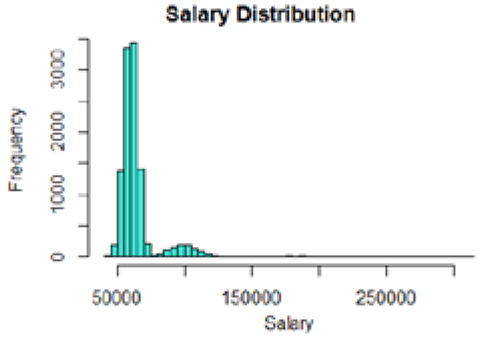
We segment the variable "age " further to get important insights.



Conclusion: This shows People inside 34-54 age gathering leave the organization more probable than the individuals inside 22-34 who may be individual employees/ representatives. Age gathering of 54-62 is at Director level and the whittling down is least in that age gathering.

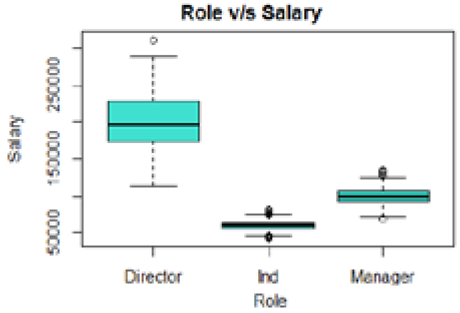
(g) Analyzing the salary pattern

A histogram of salary of the employees working in the company is to be analysed. Plot a graph using the hist function in R studio.



Conclusion: The middle pay is 60800, with the maximum being 1000000 and the min being 42170. Pay variable is profoundly skewed with practically 80% of the individuals acquiring till $66173.65. Segmenting pay division dependent on job.

Role vs Salary



Conclusion: Clearly there is a relationship between the roles and the Salary. Director have a much higher pay than others.

I. Data Modeling

Before we start making models, we have to part our information into a training set and a test set. We will use 66% of the information for training and model development and 33% of the information for testing the models. We set the random seed to a specific number so we can basically replicate our outcomes. We use set. seed(n) function in R studio to set the random seed for replication.

We will utilize two systems,
(a) Logistic Regression
(b) Decision Tree

Logistic regression constructs a condition that subsequently predicts the probability of a two-class result (staying or leaving) using the picked indicators. Every one of the indicators are associated with a "significance" pointer that tells you whether the marker is useful or not. On the other hand, decision trees work by using the indicators to part the information into buckets utilizing a lot of decision rules.

(a) Logestic Regression

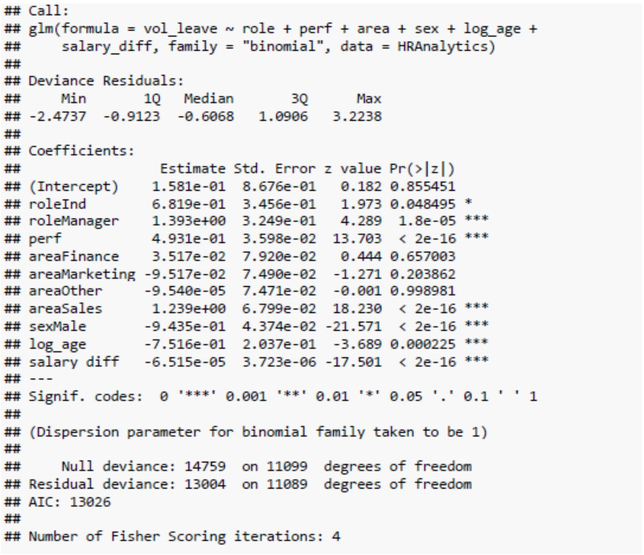
Run the following code on R studio:
test_mean =**mean**(test$vol_leave)
train_mean = **mean**(train$vol_leave)
print the two statements. This will give you the mean of both test leave and train leave values.

Output:
[1] 0.3816216 0.3814865

Fit the model

Use the glm function in R studio to set the model of the algorithm.
Run the following code:
summary(fit)

Output:



```
## 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
```

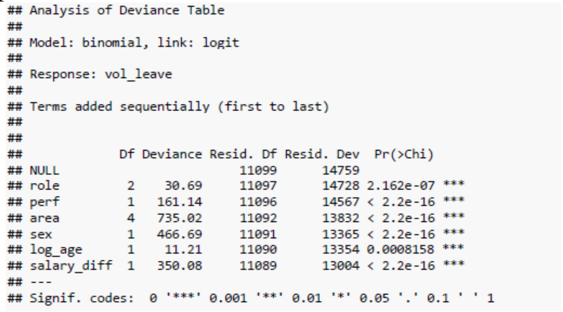
Conclusion:
Here we see that areaFinance, areaOther and areaMarketing is not exactly statistically significant.

And all the statistically significant variables, areaSales, salary and perf has the lowest p-value suggesting a strong association of these variable with the prob of leaving the organization. Now we can execute the anova() function for Chi Square test on the model to analyze the table of deviance

Chi Square Test

Execute the following statement in R studio.
anova(fit, test = "Chisq")

Output:



```
## 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
```

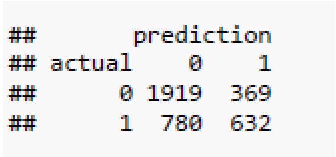
Conclusion:
The evident difference between the null deviance and the residual deviance depicts how the model is doing against the null model. The wider the gap is, the better.

A smaller p-value here indicates that all the variables in the model are significant

Assessing the predictive ability of the model

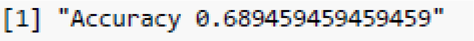
Calculate the fitted result using predict function and then perform an ifelse function on it and find the mean value which would be utilized in making he confusion matrix.

Confusion Matrix
Make the confusion matrix table using table function in R studio



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

Accuracy
Print the accuracy of the matrix
Output:

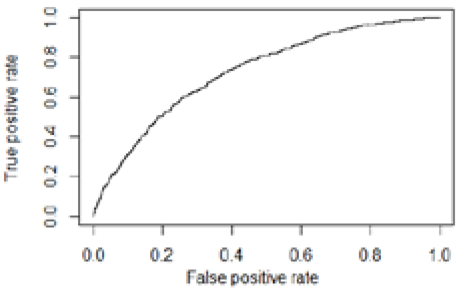


```
[1] "Accuracy 0.689459459459459"
```

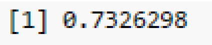
Conclusion: As we can see there is 68% accuracy in this model.

ROC Curve & AUC
Now comes the final step which is we plot the ROC curve and then calc the AUC i.e, area under the curve which are a performance measurements typical to a binary classifier.
Use the predict function to calculate the values and then plot the graph using the plot function.

Output:



Also calculate the performance using the performance function in R Studio. Now print the value:



```
[1] 0.7326298
```

Conclusion:
The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. The AUC is the area under the ROC curve. As a rule of thumb, a model with great predictive ability ought to have an AUC more like (1 is perfect) than to 0.5. In view of the estimation of AUC for our dataset, we can say that it has great predictive ability.

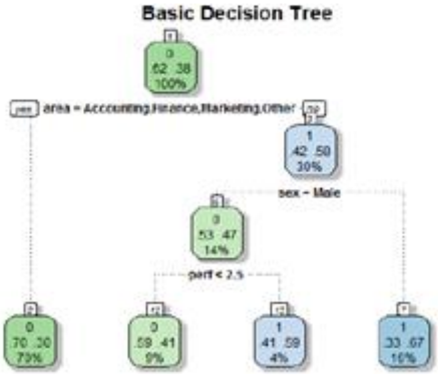
(b) Decision Tree
We have already divided our dataset into training and testing. So we proceed further by making the decision tree

Fit the model
Use the set seed function once again. Now perform a rpart function on role, perg, age, sex and area and also salary and then plot the tree**.**

Plot the tree

Use the function par to plot the tree with c(5,4,1,2) and fancy Rpart Plot the decision_fit value that we found above.

Output:



Conclusion: 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 (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. On the off chance that the answer is 'no' (i.e. they are in Sales), then we take the right branch.

After the left branch, we see that it ends into a solitary node. Think of these node like a say bucket for 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%' revealed in the base of the node discloses to us that this single bucket represents 70% of the absolute example 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.

Assessing the predictive ability of the modelCalculate the fitted result using predict function and then perform an ifelse function on it and find the mean value which would be utilized in making he confusion matrix.

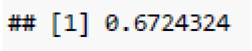
Confusion Matrix

Make the confusion matrix table using table function in R studio

Accuracy

Print the accuracy of the matrix

Output:



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

Conclusion: As we can see there is 68% accuracy in this model.

4. Conclusion

It is clear that businesses can't make due over the long haul if they don't have predictive examination abilities from the human asset the board. The helpfulness of predictive examination is more extensive and henceforth application in every single related area of HRM is fundamental.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.

References

[1] Kirsten Edwards and Martin Edwards, Predictive HR Analytics: Mastering the HR Metric
[2] Sujeet N. Mishra, Dev Raghvendra Lama, Yogesh Pal, Human Resource Predictive Analytics (HRPA) For HR Management In Organizations, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 5, ISSUE 05, MAY 2016
[3] L. Bassi, “Raging Debates in HR Analytics”, People & Strategy, Vol. 34, Issue 2, 2011
[4] M. Molefe, “From Data to Insights : HR Analytics in Organizations,” Gordon Institute of Business Science, University of Pretoria, 11 Nov. 2013
[5] K. Ladimeji, “5 Things that HR Predictive Analytics will Actually Predict.” Recruiter (Jan. 23, 2013), sec. I p.1.
[6] J Fitz-enz and J. R. Mattrox II, “Predictive Analytics for Human Resource.” Wiley Publication, SAS Institute Inc., Cary, North America, USA, 2014 pp. 2-3
[7] https://www.kaggle.com/