Wichita State University

IME-780AN Big Data Analytics in Engineering

HR Analysis on Graduate Turnover

Authors:

Lakshay Arora, J436R998

Pranav Bohre, E978E879

Tej Desai A239M527

Faculty:

Dr. Wujun Si

Assistant Professor

Industrial Engineering Dept.

Contents

1 Introduction **2**

2 Data Description **3**

2.1 Analysis Objective 5

2.2 Methods Used For Predictive Analysis 5

2.3 Data Pre-Processing 6

2.4 Evaluation/Performance Analysis Aprroach 6

3 Result Analysis **8**

3.1 Logistic Regression 8

3.2 ROC Curve 10

3.3 Predictive Analysis-Decision Tree 11

4 Conclusion **13**

5 Appendix: Code **14**

References **16**

1 Introduction

Human Resources(HR) information and management information (MI) teams currently spend considerable time and effort producing descriptive reports for the data collected– monitoring them, comparing them across geographical boundaries and over time periods, but often doing very little else with the report other than producing it – again and again. Descriptive HR reports usually produced by MI teams will generally only present a picture or 'snapshot' of what is occurring in the organization at that particular time. Whilst there is little doubt that these reports are useful to the business in ensuring that managers understand what is going on within the organization, there is a real limit to what these reports can tell us. Descriptive reports do very little more than describe what is happening; they lack the capability to help understand and account for why things are happening in the organization.

An HR function that utilizes predictive HR analytics capabilities will be more reliable because the function will be able to present robust *hard* evidence to show that it has a good understanding of what makes its people tick, along with knowledge of who is likely to perform well, who is likely to leave, which parts of the organization are showing race or gender bias, which candidates are likely to be successful in the organization, and which interventions had a significant impact on the organization and which did not. The function will be able to carry out substantial *what if* scenario modelling to help build solid business cases that help the organization to make decisions around whether particular investments are likely to be worthwhile, and what the return on those investments are likely to be.

The data provided to HR is huge and very scattered. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to this rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets so that decisions can be made more accurately.

2 Data Description

The graduate employee turnover dataset consist of data consist of HR information which is collected at the time of recruitment process which contains scores and ratings taken at graduate assessment centers for a particular Financial organization firm which hires a large cohort of graduates each year. The dataset contains 23 variables for 360 graduate's students. The main objective on which dataset was collected was to predict the graduate turnover based on their personal traits and other assessment scores. The data collected includes test score gives by the interview panels, the scores obtained by the graduate's in pre assessment tests and also include the data of number of candidate attending first day of induction training and induction week. The detailed discussion about the dataset is given below:

Number of records/attributes: 23 attributes

Meaning of Records:

- GradID- Employee Identifier
- Gender Male (1) or Female (2)
- EducationHighest-Education Category:
- BSc (1) MSc (2) PhD (3)
- BAMEYN Black, Asian or Minority Ethnic: Yes (1) or No (2)
- WorkExperience- Worked Before: Yes (1) No (2)
- GradJOBfunction- Function Joined: (1) HR; (2) Finance; (3) Marketing; (4) Sales; (5) Risk; (6) Legal; (7) Operations
- ACPersonalityO -Assessment C Personality Openness (percentage)
- ACPersonalityC- Assessment C Personality Conscientious (percentage)
- ACPersonalityE- Assessment C Personality Extraversion (percentage)

- ACPersonalityA -Assessment C Personality Agreeable (percentage)
- ACPersonalityN Assessment C Personality Neuroticism (percentage)
- ACRatingINTCOMPA Assessment C Competency Technical (1–5)
- ACRatingINTCOMPB Assessment C Competency Team Player
- ACRatingINTCOMPC Assessment C Competency Critical Thought
- ACRatingINTCOMPD Assessment C Competency Business Aware
- ACRatingINTCOMPE Assessment C Competency Drive Innovation
- ACRatingAPTnumerical Numerical Aptitude Test Score
- ACRatingAPTverbal Verbal Reasoning AptitudeTest Score
- Induction Day Attend Induction Day: No (0) Yes (1)
- Induction Week Attend Induction Week: No (0) Yes (1)
- On Boarding Buddy Given Joining Buddy: No (0) Yes (1)
- Year1performanceRating Year 1 Performance Rating: (1) Fails to Meet Expectations (2) Just Meets Expectations (3) Meets Expectations (4) Exceeds Expectations (5) A Star Performer
- LeaverYr2 Leaver in First 12 Months (0) No (1) Leaver

The 30 variable in this dataset contains some CV related information such as graduate's gender, highest education, BAME, Work Experience. Then the data set contains the personality test scores given by the interviewer at the time of recruiting based on openness, conscientiousness, extraversion, agreeableness and neuroticism. Then we also have score for competency rating based on technical skills, Team player, Critical thinking, Business awareness and Drive Innovation. Then we have score for numerical and verbal reasoning and then we have HR collected data on presence of employee on induction data and induction week and how many people were given "on boarding buddy" and graduates leaving the company in first 12 months.

2.1 Analysis Objective

The main objective which we took for the analysis is to predict which factors mostly influences the graduates to turnover within first 12 months (Leaver Yr2) and predict a criteria of how much a graduate is likely to turnover based on his test scores, work experience, and other significant factors which will help a HR to decide the recruitment criteria which will solve the problem of graduates leaving each year and will also help company to dig down to know the reasons from employee to make changes in working environment and culture to avoid spending of money and time on hiring new graduates and training them each year.

2.2 Methods Used For Predictive Analysis

1. **Logistic Regression**: Logistic regression is a statistical model method for analyzing a dataset in which there are one or more independent variables that determine outcome. Logistic regression is used when the dependent variable (target) is categorical or binary. Thus we performed our predictive analysis using logistic regression as it is appropriate analysis technique for performing predictive analysis when dependent variable is dichotomous (binary). Also the logistic regression is used to describe the data and to explain the relationship between one dependent binary variable and one or more independent variables. Logistic regression generates the coefficients (and its standard errors and significant level) of formula to predict a logit transformation of the probability of presence of characteristic of interest.

2. **Decision Tree**: In order to help the HR to set the baseline for recruitment process for the significant factors which will help the HR to predict in future recruitment process. The prediction can be achieved by constructing a decision tree with test points and branches. Due to its easy visualization and flexibility we preferred to use decision tree.

2.3 Data Pre-Processing

From the initial screening made by our team on the dataset, we found that there are no missing values or data from the dataset so we did not do much of the pre-processing work. Only thing that we did at the start of our data was to convert the variable into factors using *as.factor* function.

```
project_data_1$GradJOBfunction= as.factor(project_data_1$GradJOBfunction
    )
project_data_1$EducationHighest= as.factor(
    project_data_1$EducationHighest)
project_data_1$Year1performanceRating= as.factor(
    project_data_1$Year1performnceRating)
```

2.4 Evaluation/Performance Analysis Aprroach

Based on the concepts taught in the class, we used some basic data exploration analysis using R language. The graphs for the basic analysis and the codes used are given below. We installed "ggplot2" package to perform some initial data visualization. We tried to explore the number of graduate employee involved in different Job function number of employee leaving after working for 12 months:

```
a= ggplot(project_data_1,aes(x=GradJOBfunction))+ geom_bar()+ ggtitle("
    Employees in diffrent job function: 1=HR,2=FINANCE,3=MARKETING,4=
    SALES,5= RISK,6= LEGAL,7= OPERATIONS")

prop.table(table(project_data_1$GradJOBfunction))
```

From the above plot we analysed that there are more number of graduates working in marketing function of the financial firm.

We also visualized on how many people leave the job after first 12 months using the following code:

```
b= ggplot(project_data_1,aes(x=LeaverYr2))+ geom_bar()+theme_light()+
    labs(y="Employeer count", title ="Number of employee left in first
    12 months" )
```

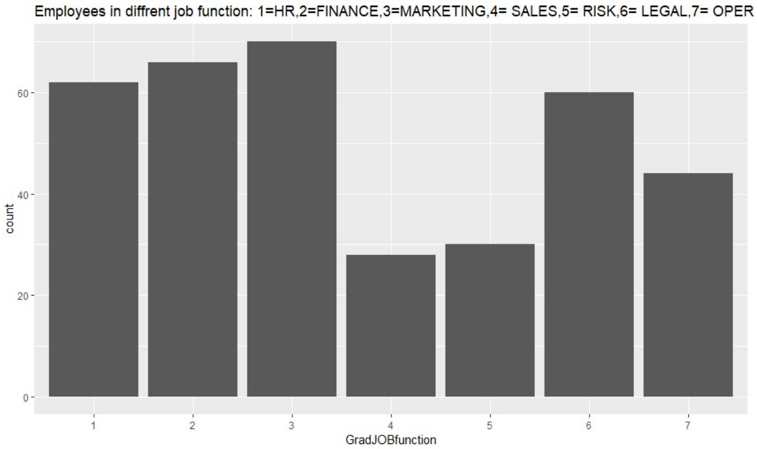


Figure 1: Jobs

```
prop.table(table(project_data_1$LeaverYr2))
```

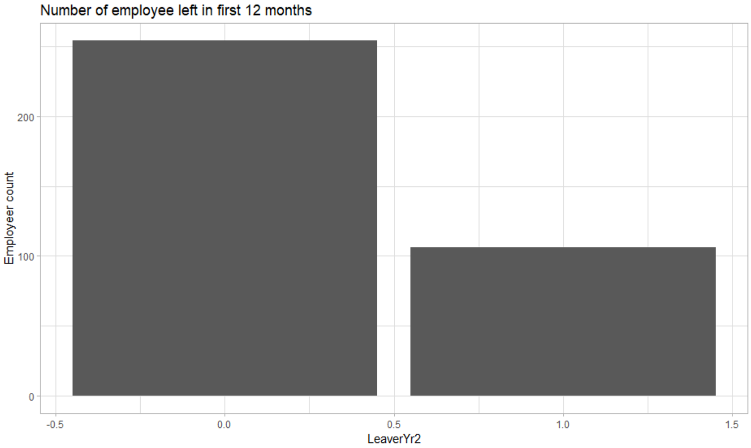


Figure 2: Number of employee leaving in first 12 months

From the above plot we can see that the not much number of employee left after completion of 1 year. To explore more factors that affects the graduate turnover we analysed our dataset using regression.

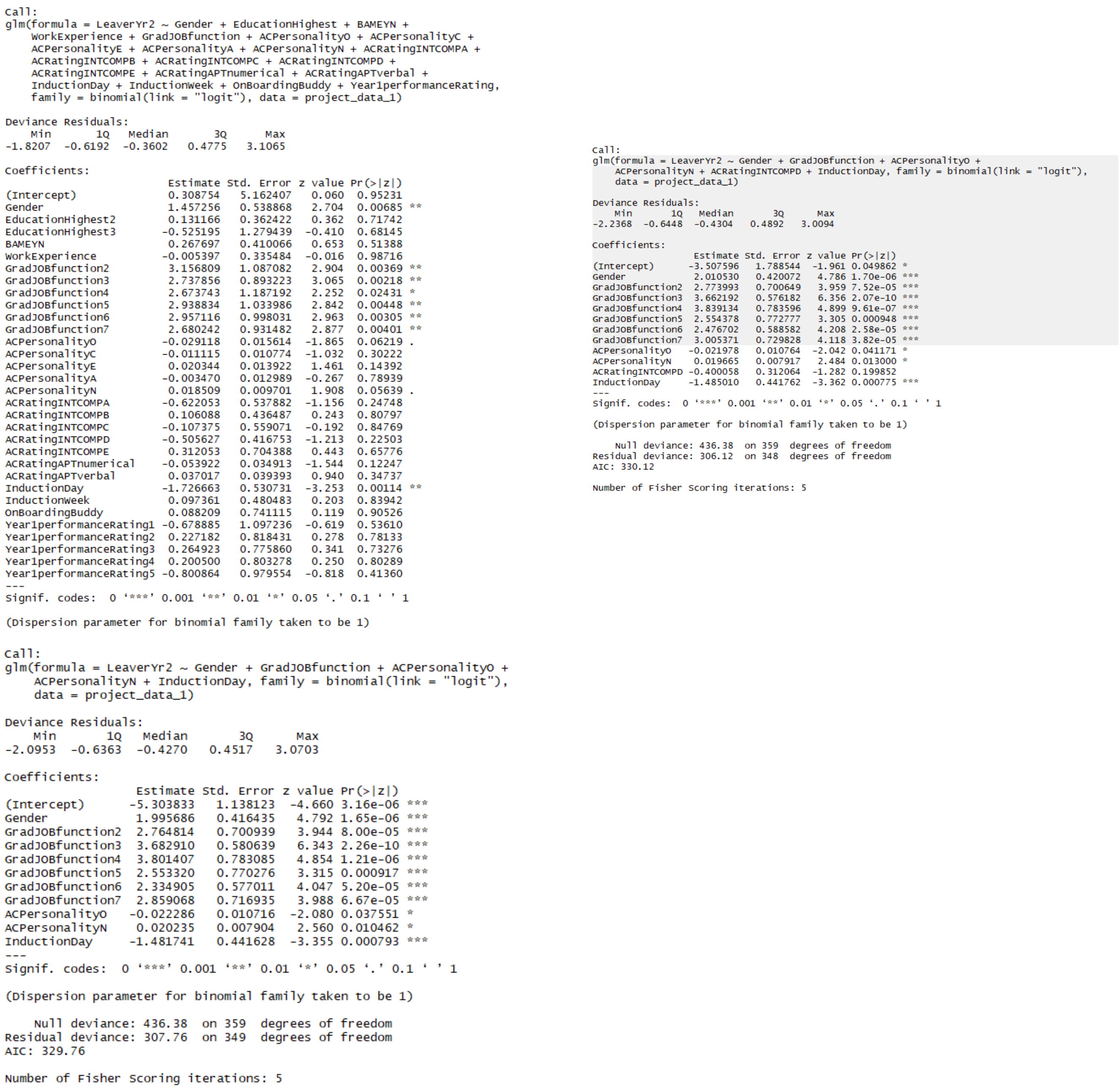
3 Result Analysis

3.1 Logistic Regression

To perform predictive analysis on our data selected we first analysed our model using logistic regression we performed 3 iteration where we eliminated non-significant factors based on their p values. From our analysis we observed that there are 11 significant factors which are gender, Grad job function, and Personality trait score of Openness and Neuroticism and Induction day training. These variable connect to the dependent variable leaver in first 12 month in statistical significant way. The negative value of coefficient on ACPersonalityO and Induction Day indicates that the event identified by Dependent variable (i.e.LeaverYr2) in our case decreases as the value of AC personality and Induction Day (attend) increases. Also for the values of positive estimate value of grad job functions, gender and personality test score of Neuroticism indicates that leaverYr (Dependent variable) probability increases with positive change in these variables. So it can be observed that with increase in score of openness and attendance on induction day training the graduate employee are less likely to leave the company and with higher score of neuroticism the chances of employee to be leave the company also increase. Thus the person with higher neuroticism score are more likely to be moody and to experience negative feelings which can lead to leaving the company.

Regression equation is given by:

$$\begin{aligned} Y = -5.30 + 1.99 * gender + 2.76 * gradjobfunction2 \\ +3.68 * gradjobfunction3 + 3.80 * gradjobfunction4 \\ +2.55 * gradjobfunction5 + 2.33 * gradjobfunction6 \\ +2.85 * gradjobfunction7 - 0.02 * ACpersonalityO \\ +0.02 * ACpersonalityN - 1.48 * InductionDay \end{aligned} \tag{1}$$



```
Call:
glm(formula = LeaverYr2 ~ Gender + EducationHighest + BAMEYN +
    WorkExperience + GradJOBfunction + ACPersonalityO + ACPersonalityC +
    ACPersonalityE + ACPersonalityA + ACPersonalityN + ACRatingINTCOMPA +
    ACRatingINTCOMPB + ACRatingINTCOMPC + ACRatingINTCOMPD +
    ACRatingINTCOMPE + ACRatingAPTnumerical + ACRatingAPTverbal +
    InductionDay + InductionWeek + OnBoardingBuddy + Year1performanceRating,
    family = binomial(link = "logit"), data = project_data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8207  -0.6192  -0.3602   0.4775   3.1065

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.308754   5.162407   0.060  0.95231
Gender                   1.457256   0.538868   2.704  0.00685 **
EducationHighest2        0.131166   0.362422   0.362  0.71742
EducationHighest3       -0.525195   1.279439  -0.410  0.68145
BAMEYN                   0.267697   0.410066   0.653  0.51388
WorkExperience          -0.005397   0.335484  -0.016  0.98716
GradJOBfunction2         3.156809   1.087082   2.904  0.00369 **
GradJOBfunction3         2.737856   0.893223   3.065  0.00218 **
GradJOBfunction4         2.673743   1.187192   2.252  0.02431 *
GradJOBfunction5         2.938834   1.033986   2.842  0.00448 **
GradJOBfunction6         2.957116   0.998031   2.963  0.00305 **
GradJOBfunction7         2.680242   0.931482   2.877  0.00401 **
ACPersonalityO          -0.029118   0.015614  -1.865  0.06219 .
ACPersonalityC          -0.011115   0.010774  -1.032  0.30222
ACPersonalityE           0.020344   0.013922   1.461  0.14392
ACPersonalityA          -0.003470   0.012989  -0.267  0.78939
ACPersonalityN           0.018509   0.009701   1.908  0.05639 .
ACRatingINTCOMPA        -0.622053   0.537882  -1.156  0.24748
ACRatingINTCOMPB         0.106088   0.436487   0.243  0.80797
ACRatingINTCOMPC        -0.107375   0.559071  -0.192  0.84769
ACRatingINTCOMPD        -0.505627   0.416753  -1.213  0.22503
ACRatingINTCOMPE         0.312053   0.704388   0.443  0.65776
ACRatingAPTnumerical    -0.053922   0.034913  -1.544  0.12247
ACRatingAPTverbal        0.037017   0.039393   0.940  0.34737
InductionDay            -1.726663   0.530731  -3.253  0.00114 **
InductionWeek            0.097361   0.480483   0.203  0.83942
OnBoardingBuddy          0.088209   0.741115   0.119  0.90526
Year1performanceRating1 -0.678885   1.097236  -0.619  0.53610
Year1performanceRating2  0.227182   0.818431   0.278  0.78133
Year1performanceRating3  0.264923   0.775860   0.341  0.73276
Year1performanceRating4  0.200500   0.803278   0.250  0.80289
Year1performanceRating5 -0.800864   0.979554  -0.818  0.41360
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
Call:
glm(formula = LeaverYr2 ~ Gender + GradJOBfunction + ACPersonalityO +
    ACPersonalityN + InductionDay, family = binomial(link = "logit"),
    data = project_data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0953  -0.6363  -0.4270   0.4517   3.0703

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.303833   1.138123  -4.660 3.16e-06 ***
Gender           1.995686   0.416435   4.792 1.65e-06 ***
GradJOBfunction2 2.764814   0.700939   3.944 8.00e-05 ***
GradJOBfunction3 3.682910   0.580639   6.343 2.26e-10 ***
GradJOBfunction4 3.801407   0.783085   4.854 1.21e-06 ***
GradJOBfunction5 2.553320   0.770276   3.315 0.000917 ***
GradJOBfunction6 2.334905   0.577011   4.047 5.20e-05 ***
GradJOBfunction7 2.859068   0.716935   3.988 6.67e-05 ***
ACPersonalityO  -0.022286   0.010716  -2.080 0.037551 *
ACPersonalityN   0.020235   0.007904   2.560 0.010462 *
InductionDay    -1.481741   0.441628  -3.355 0.000793 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 436.38  on 359  degrees of freedom
Residual deviance: 307.76  on 349  degrees of freedom
AIC: 329.76

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = LeaverYr2 ~ Gender + GradJOBfunction + ACPersonalityO +
    ACPersonalityN + ACRatingINTCOMPD + InductionDay, family = binomial(link = "logit"),
    data = project_data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2368  -0.6448  -0.4304   0.4892   3.0094

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.507596   1.788544  -1.961 0.049862 *
Gender            2.010530   0.420072   4.786 1.70e-06 ***
GradJOBfunction2  2.773993   0.700649   3.959 7.52e-05 ***
GradJOBfunction3  3.662192   0.576182   6.356 2.07e-10 ***
GradJOBfunction4  3.839134   0.783596   4.899 9.61e-07 ***
GradJOBfunction5  2.554378   0.772777   3.305 0.000948 ***
GradJOBfunction6  2.476702   0.588582   4.208 2.58e-05 ***
GradJOBfunction7  3.005371   0.729828   4.118 3.82e-05 ***
ACPersonalityO   -0.021978   0.010764  -2.042 0.041171 *
ACPersonalityN    0.019665   0.007917   2.484 0.013000 *
ACRatingINTCOMPD -0.400058   0.312064  -1.282 0.199852
InductionDay     -1.485010   0.441762  -3.362 0.000775 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 436.38  on 359  degrees of freedom
Residual deviance: 306.12  on 348  degrees of freedom
AIC: 330.12

Number of Fisher Scoring iterations: 5
```

Figure 3: Logistic Regression Results

3.2 ROC Curve

In a ROC curve the true positive rate is plotted in function of the false positive rate for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two groups.

This is used to evaluate and interpret logistic regression results. The usefulness of this plot is that the preferred outcome of a classifier is to have a low FPR and a high TPR. So, when moving from left to right on the FPR axis, a good model/classifier has the TPR rapidly approach values near 1, with only a small change in FPR. The closer the ROC curve tracks along the vertical axis and approaches the upper-left hand of the plot, near the point (0,1), the better the model/classifier performs. Thus, a useful metric is to compute the area under the ROC curve (AUC). By examining the axes, it can be seen that the theoretical maximum for the area is 1. From the curve, area under the curve came out to be 0.8372 for our logistic regression analysis which is close to 1.

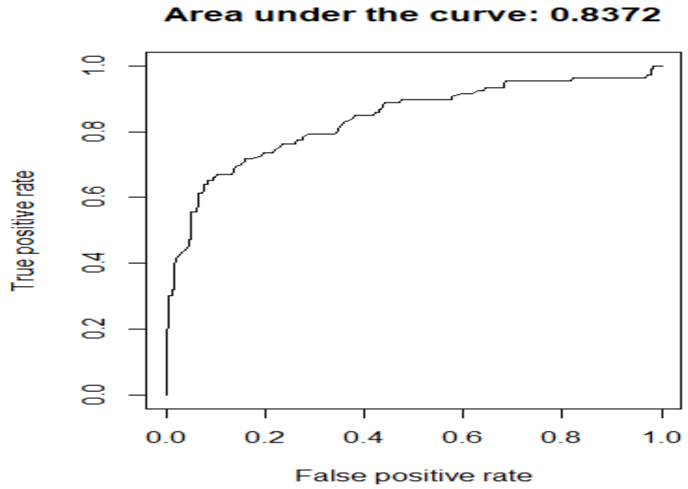


Figure 4: ROC curve for HR prediction

For the purpose of our classifier, threshold value is taken as 0.2. This results in TPR = 0.7925 and FPR = 0.311.

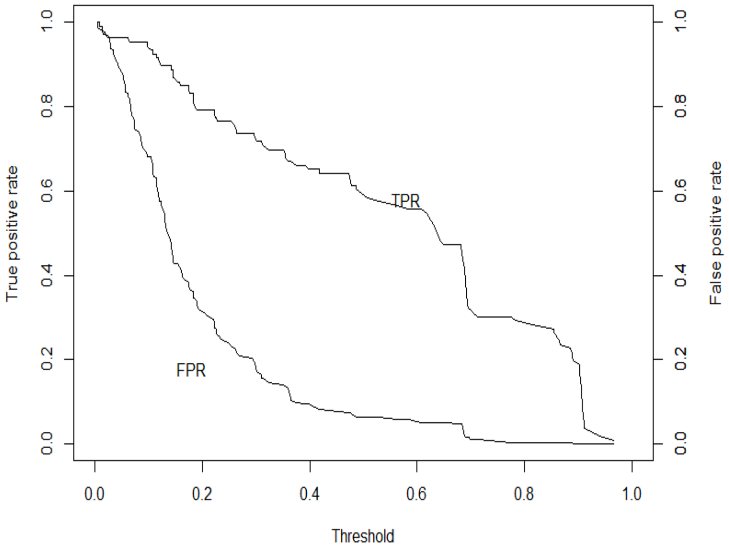


Figure 5: Effect of threshold value

This represents that 80% of the graduates who will leave are properly identified but at a cost of misclassifying 31% of the graduates who won't.

3.3 Predictive Analysis-Decision Tree

We are using Decision tree to predict the criteria for test scores which will help to set the baseline of scores for recruitment process which will help HR to tackle the problem of graduate employee turnover.

Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables. The prediction between input and output can be achieved by constructing a decision tree with test points and branches. At each test point, a decision is made to pick a specific branchand traverse down the tree. Eventually, a final point is reached, and a prediction can be made. Each testpoint in a decision tree involves testing a particular input variable (or attribute), and each branch representsthe decision being made. Due to its flexibility and easy visualization, decision trees are commonly deployed in data mining applications for classification purposes.

For the current problem, since HR needs to predict based on behavioral traits of the graduates and induction day attendance which is why Induction Day and Personality Test scores for Openness and Neuroticism are used as the input decision variables.

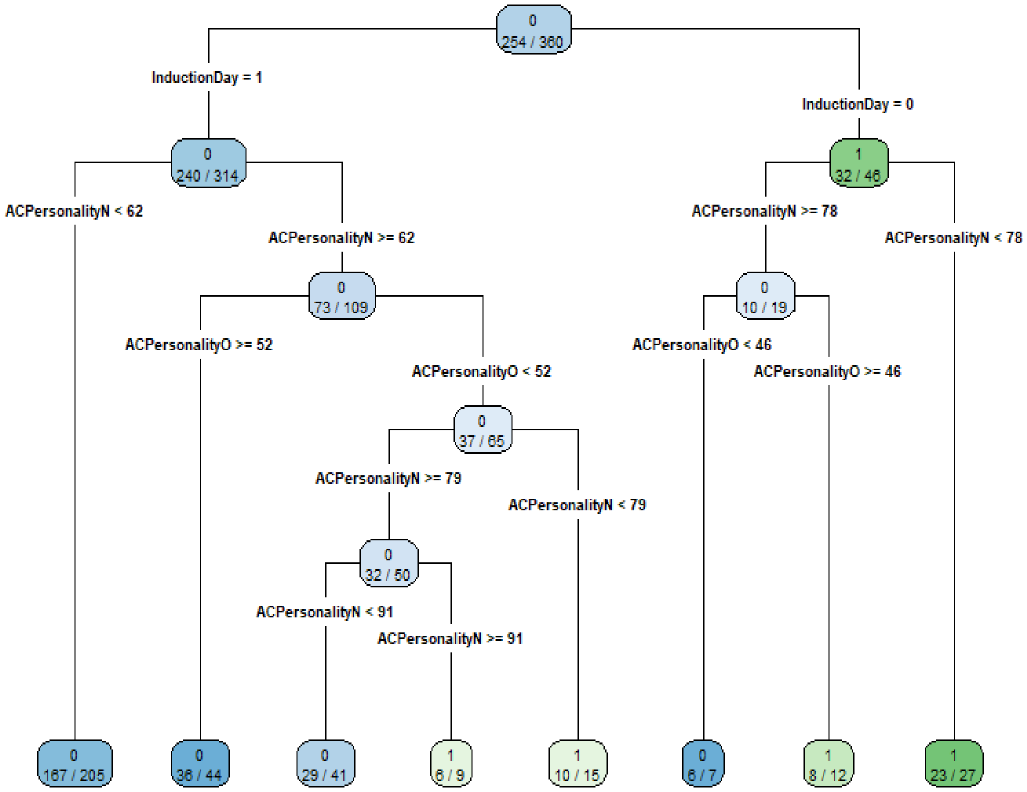


Figure 6: Decision Tree

From the figure , it is clear that 240 graduates attended(InductionDay=1) the induction day and 32 who did not attend it(InductionDay=0). Then the number of graduates is given which are classified on the basis of the Personality Test Scores for Openness and Neuroticism.

4 Conclusion

From the analysis of the data set we took into consideration and by doing the predictive analysis, we found that the personality traits plays a major role in graduate employee turnover. Their presence on induction day also has significant effect on their probability of leaving, so in our conclusion we would like to recommend that at the time of recruitment candidates with score of above 52 for personality trait score of openness and score of less than 62 for neuroticism should be given more opportunity as there would be less chances of their turnover.

5 Appendix: Code

```


library(readxl)
project_data_1 <- read_excel(”C:/Users/adm/Desktop/project data 1.xlsx”)
View(project_data_1)
project_data_1$GradJOBfunction= as.factor(project_data_1$GradJOBfunction)
project_data_1$EducationHighest= as.factor(project_data_1$EducationHighest)
project_data_1$Year1performanceRating= as.factor(project_data_1$Year1performanceRating)
Project_datalogistic01= glm(LeaverYr2~Gender+EducationHighest+BAMEYN+WorkExperience+GradJOBfunction+ACPersonalityO+ACPersonalityC+ACPersonalityE+ACPersonalityA+ACPersonalityN+ACRatingINTCOMPA+ACRatingINTCOMPB+ACRatingINTCOMPC+ACRatingINTCOMPD+ACRatingINTCOMPE+ACRatingAPTnumerical+ACRatingAPTverbal+InductionDay+InductionWeek+OnBoardingBuddy+Year1performanceRating,data = project_data_1,family = binomial(link = ”logit”))
summary(Project_datalogistic01)
Project_datalogistic02 = glm(LeaverYr2~Gender+GradJOBfunction+ACPersonalityO+ACPersonalityN+ACRatingINTCOMPD+InductionDay,data = project_data_1,family = binomial(link = ”logit”))
summary(Project_datalogistic02)
Project_datalogistic03 = glm(LeaverYr2~Gender+GradJOBfunction+ACPersonalityO+ACPersonalityN+InductionDay,data = project_data_1,family = binomial(link = ”logit”))
summary(Project_datalogistic03)
library(ROCR)
pred= predict(Project_datalogistic03,type = ”response”)
 predobj1= prediction(pred,project_data_1$LeaverYr2)
 rocObj= performance(predobj1,measure = ”tpr”,x.measure = ”fpr”)
 aucObj= performance(predobj1,measure = ”auc”)
 plot(rocObj,main= paste(”Area under the curve:”,round(aucObj@y.values[[1]],4)))
```

```r
threshold= round(as.numeric(unlist(rocObj@alpha.values)),4)
fpr= round(as.numeric(unlist(rocObj@x.values)),4)
tpr= round(as.numeric(unlist(rocObj@y.values)),4)
par(mar= c(5,5,2,5))
plot(threshold,tpr, xlab="Threshold",xlim= c(0,1), ylab= "True positive rate",type= "l")
par(new="True")
plot(threshold, fpr, xlab=" ", ylab= " ", axes= F , xlim= c(0,1),type="l")
axis(side = 4)
axis(side = 4)
mtext(side = 4, line = 3,"False positive rate")
text(0.18,0.18,"FPR")
text(0.58,0.58,"TPR")
a= which(round(threshold,2) == 0.20)
paste("Threshold=" , (threshold[a]), "TPR=", tpr[a], "FPR=", fpr[a])
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
Fit1= rpart(LeaverYr2~ACPersonalityO+ACPersonalityN+InductionDay,method = "class", data = project_data_1,control = rpart.control(minsplit = 1), parms = list(split='information'))
summary(Fit1)
rpart.plot(Fit1,type = 4,extra = 1)
rpart.plot(Fit1,type=4,extra = 2,clip.right.labs = FALSE, varlen = 0, faclen = 0)
Fit2= rpart(LeaverYr2~ACPersonalityO+ACPersonalityN,method = "class", data = project_data_1,control = rpart.control(minsplit = 1), parms = list(split='information'))
summary(Fit2)
rpart.plot(Fit2,type = 4,extra = 1)
rpart.plot(Fit2,type=4,extra = 2,clip.right.labs = FALSE, varlen = 0, faclen = 0)
```

References

[1] Nada Elgendy and Ahmed Elragal. Big data analytics: A literature review paper. In Petra Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, pages 214–227, Cham, 2014. Springer International Publishing.

[2] Nada Elgendy and Ahmed Elragal. Big data analytics: a literature review paper. In *Industrial Conference on Data Mining*, pages 214–227. Springer, 2014.

[3] EMC Education Services. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* January 2015.