ACME Employee Job Attrition

Ramyadhevi Vijayakumar

Objectives

ACME has an attrition problem. We are going to: (1) Identify the primary reasons behind attrition (2) Design machine learning algorithms to identify the 10 employees most inclined to leave (3) Indicate the likelihood of the attrition of the most important attributes

Data Structures

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Warning: package 'mdsr' was built under R version 3.5.2
```

```
## Warning: package 'mosaic' was built under R version 3.5.2
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Warning: package 'ggformula' was built under R version 3.5.2
```

```
## Loading required package: ggstance
```

```
## Warning: package 'ggstance' was built under R version 3.5.2
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Warning: package 'mosaicData' was built under R version 3.5.2
```

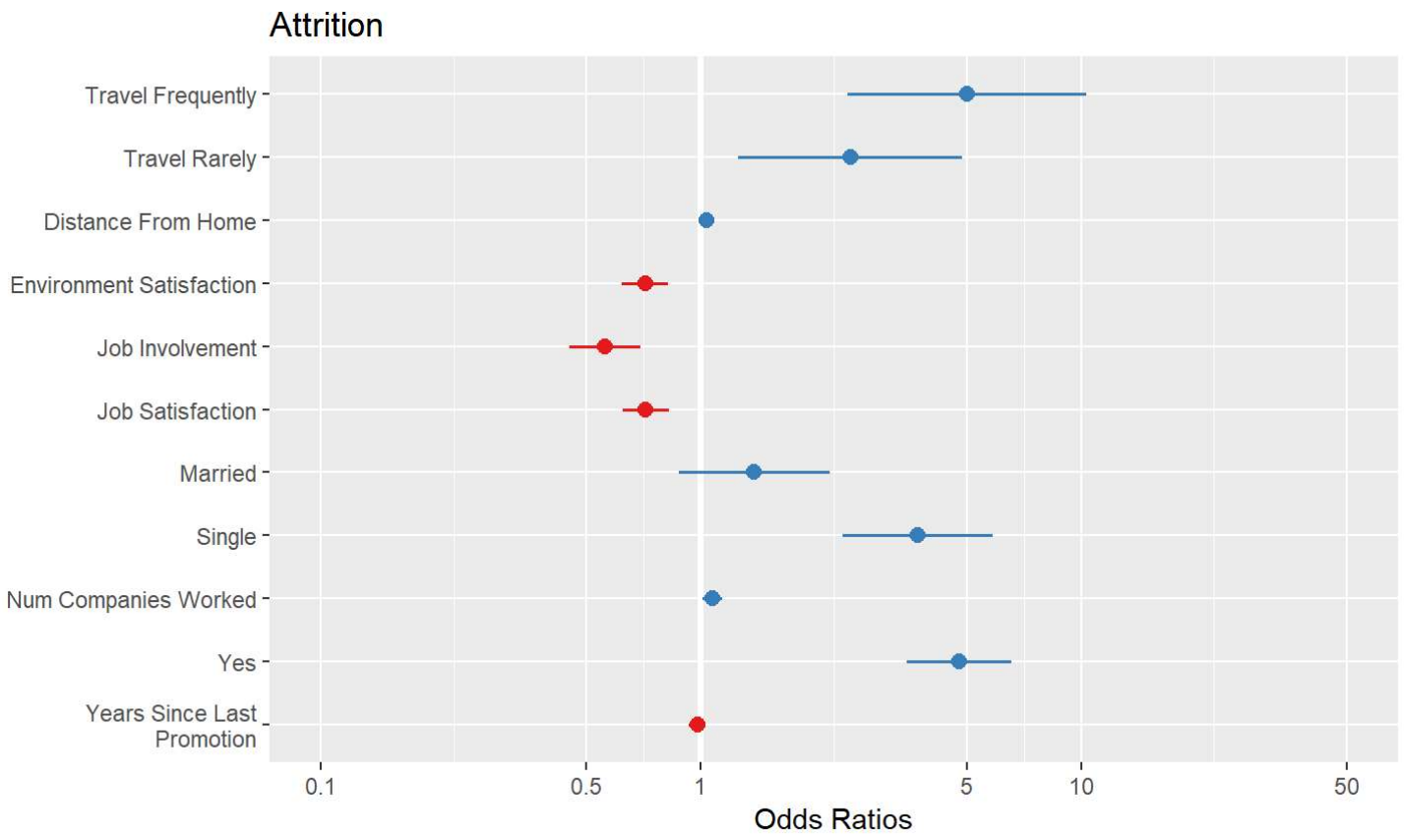
The most significant reasons for attrition are:

(i)BusinessTravel (Travel_Frequently) (ii)DistanceFromHome (iii)EnvironmentSatisfaction (iv)JobInvolvement (v)JobSatisfaction (vi)MaritalStatus (Single) (vii)NumCompaniesWorked (viii)OverTime (Yes) (ix)YearsSinceLastPromotion

Attrition

```
##
## Call:
## glm(formula = Attrition ~ BusinessTravel + DistanceFromHome +
##     EnvironmentSatisfaction + JobInvolvement + JobSatisfaction +
##     MaritalStatus + NumCompaniesWorked + OverTime + YearsSinceLastPromotion,
##     family = "binomial", data = hr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.738  -0.568  -0.385  -0.219   3.015
##
## Coefficients:
##                                Estimate Std. Error z value
## (Intercept)                   -1.02758    0.54022   -1.90
## BusinessTravelTravel_Frequently  1.61184    0.36828    4.38
## BusinessTravelTravel_Rarely      0.90421    0.34540    2.62
## DistanceFromHome                 0.03174    0.00927    3.42
## EnvironmentSatisfaction         -0.33974    0.07198   -4.72
## JobInvolvement                  -0.58238    0.10914   -5.34
## JobSatisfaction                 -0.33523    0.07082   -4.73
## MaritalStatusMarried             0.32106    0.23318    1.38
## MaritalStatusSingle              1.31232    0.23129    5.67
## NumCompaniesWorked               0.06727    0.03083    2.18
## OverTimeYes                      1.56362    0.16246    9.62
## YearsSinceLastPromotion         -0.02150    0.02563   -0.84
##                                  Pr(>|z|)
## (Intercept)                       0.05715 .
## BusinessTravelTravel_Frequently 0.000012050 ***
## BusinessTravelTravel_Rarely       0.00885 **
## DistanceFromHome                  0.00062 ***
## EnvironmentSatisfaction         0.000002359 ***
## JobInvolvement                  0.000000095 ***
## JobSatisfaction                 0.000002209 ***
## MaritalStatusMarried              0.16854
## MaritalStatusSingle             0.000000014 ***
## NumCompaniesWorked                0.02914 *
## OverTimeYes                    < 0.0000000000000002 ***
## YearsSinceLastPromotion           0.40147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.6  on 1469  degrees of freedom
## Residual deviance: 1059.7  on 1458  degrees of freedom
## AIC: 1084
##
## Number of Fisher Scoring iterations: 5
```

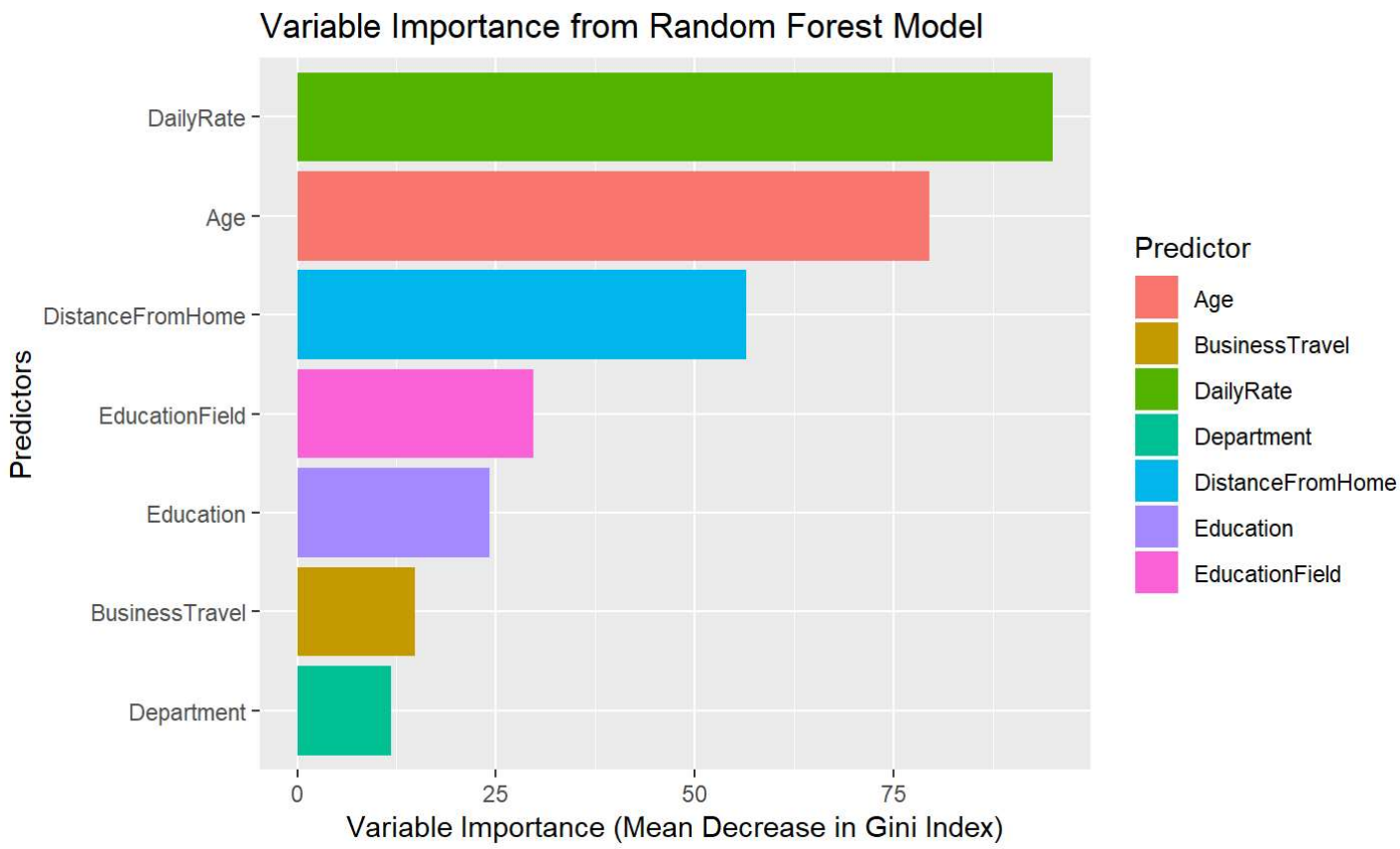
Attrition


Attrition
Travel Frequently
Travel Rarely
Distance From Home
Environment Satisfaction
Job Involvement
Job Satisfaction
Married
Single
Num Companies Worked
Yes
Years Since Last Promotion
0.1
0.5
1
5
10
50
Odds Ratios

Random Forest

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EnvironmentSatisfaction"  "Gender"
## [11] "HourlyRate"               "JobInvolvement"
## [13] "JobLevel"                 "JobRole"
## [15] "JobSatisfaction"          "MaritalStatus"
## [17] "MonthlyIncome"            "MonthlyRate"
## [19] "NumCompaniesWorked"       "Over18"
## [21] "OverTime"                 "PercentSalaryHike"
## [23] "PerformanceRating"        "RelationshipSatisfaction"
## [25] "StandardHours"            "StockOptionLevel"
## [27] "TotalWorkingYears"        "TrainingTimesLastYear"
## [29] "WorkLifeBalance"          "YearsAtCompany"
## [31] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [33] "YearsWithCurrManager"


##
##    No   Yes
## 0.841 0.159


## Attrition
##   No  Yes
## 84.1 15.9


##
## Call:
##  randomForest(formula = form, data = train, ntree = 200, mtry = 3)
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 16.58%
## Confusion matrix:
##      No Yes class.error
## No  962  27      0.0273
## Yes 168  19      0.8984


## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  240  48
##        Yes   4   2
##
```

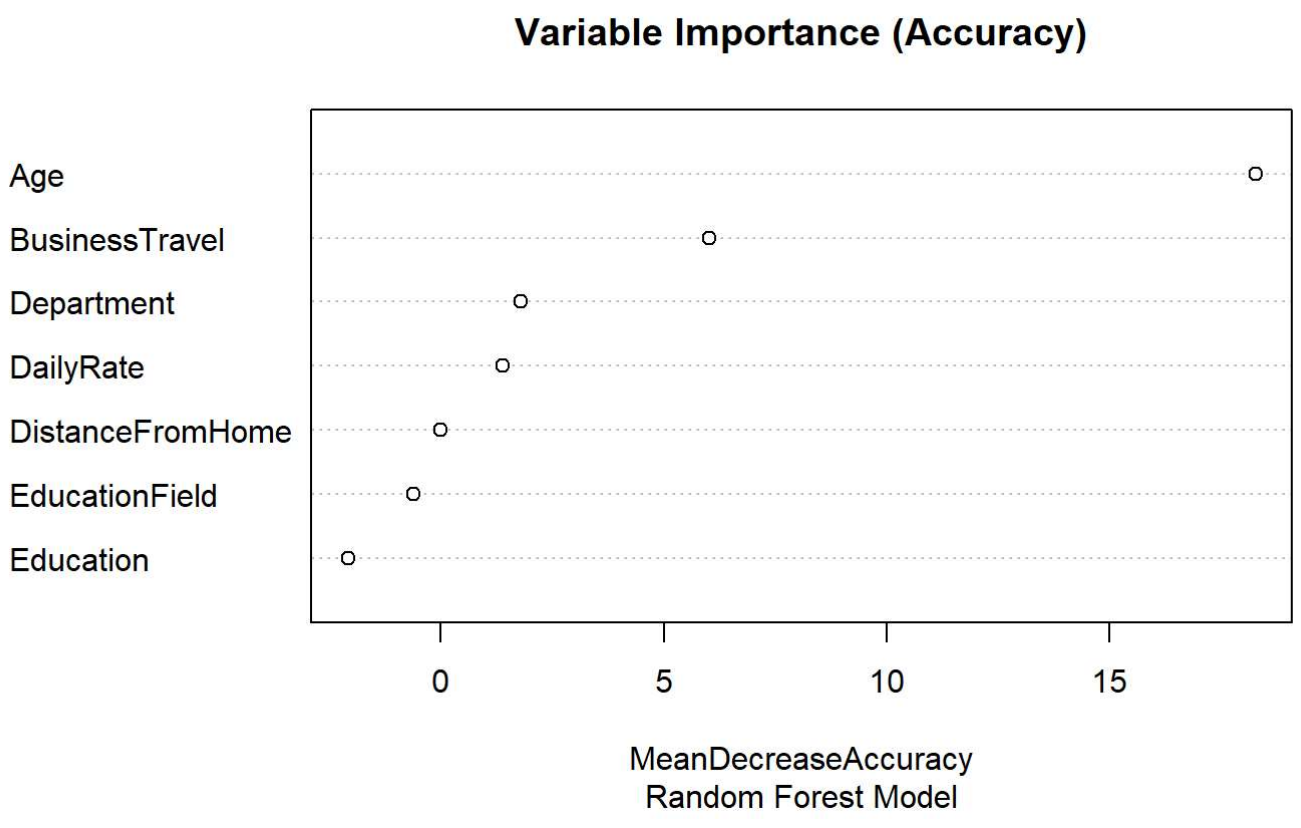
Random Forest



Important Variables

```
##
## Call:
##  randomForest(formula = form, data = train, ntree = 500, importance = TRUE,      na.action = na.
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 16.07%
## Confusion matrix:
##      No Yes class.error
## No  977  12     0.01213
## Yes 177  10     0.94652
```

Important Variable



Confusion Matrix

note: accuracy 83.7%

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction  No Yes
##        No  243  46
##        Yes   1   4
##
##               Accuracy : 0.84
##                 95% CI : (0.793, 0.88)
##    No Information Rate : 0.83
##    P-Value [Acc > NIR] : 0.354
##
##                  Kappa : 0.118
##  Mcnemar's Test P-Value : 0.000000000138
##
##               Precision : 0.8000
##                  Recall : 0.0800
##                      F1 : 0.1455
##              Prevalence : 0.1701
##          Detection Rate : 0.0136
##    Detection Prevalence : 0.0170
##       Balanced Accuracy : 0.5380
##
##        'Positive' Class : Yes
##
```

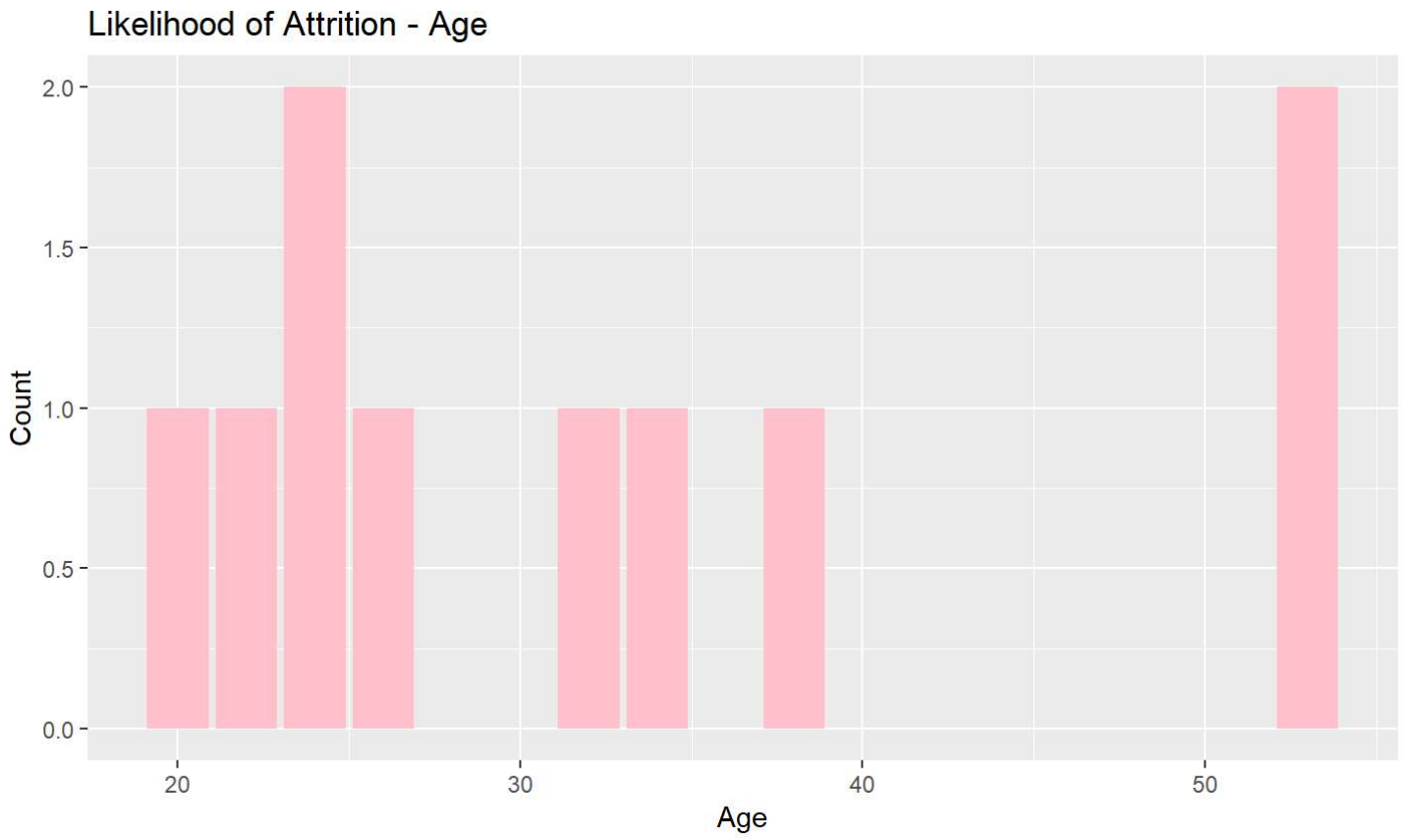
Top 10 Employee likely to abscond

These are the 10 employees most likely to leave according to the prediction analysis.

```
##    EmployeeNumber Age    BusinessTravel              Department
## 1              72  26     Travel_Rarely                   Sales
## 2             217  22     Travel_Rarely Research & Development
## 3             621  34     Travel_Rarely                   Sales
## 4             632  24 Travel_Frequently                   Sales
## 5             893  38     Travel_Rarely                   Sales
## 6             901  53     Travel_Rarely Research & Development
## 7            1050  53 Travel_Frequently                   Sales
## 8            1226  20     Travel_Rarely                   Sales
## 9            1746  24 Travel_Frequently         Human Resources
## 10           2010  32     Travel_Rarely Research & Development
##    EducationField DailyRate DistanceFromHome Education
## 1       Marketing      1443               23         3
## 2         Medical      1256               19         1
## 3   Life Sciences       258               21         4
## 4         Medical       535               24         3
## 5       Marketing       395                9         3
## 6   Life Sciences       102               23         4
## 7       Marketing       124                2         3
## 8       Marketing       654               21         3
## 9         Medical       897               10         3
## 10  Life Sciences       267               29         4
```

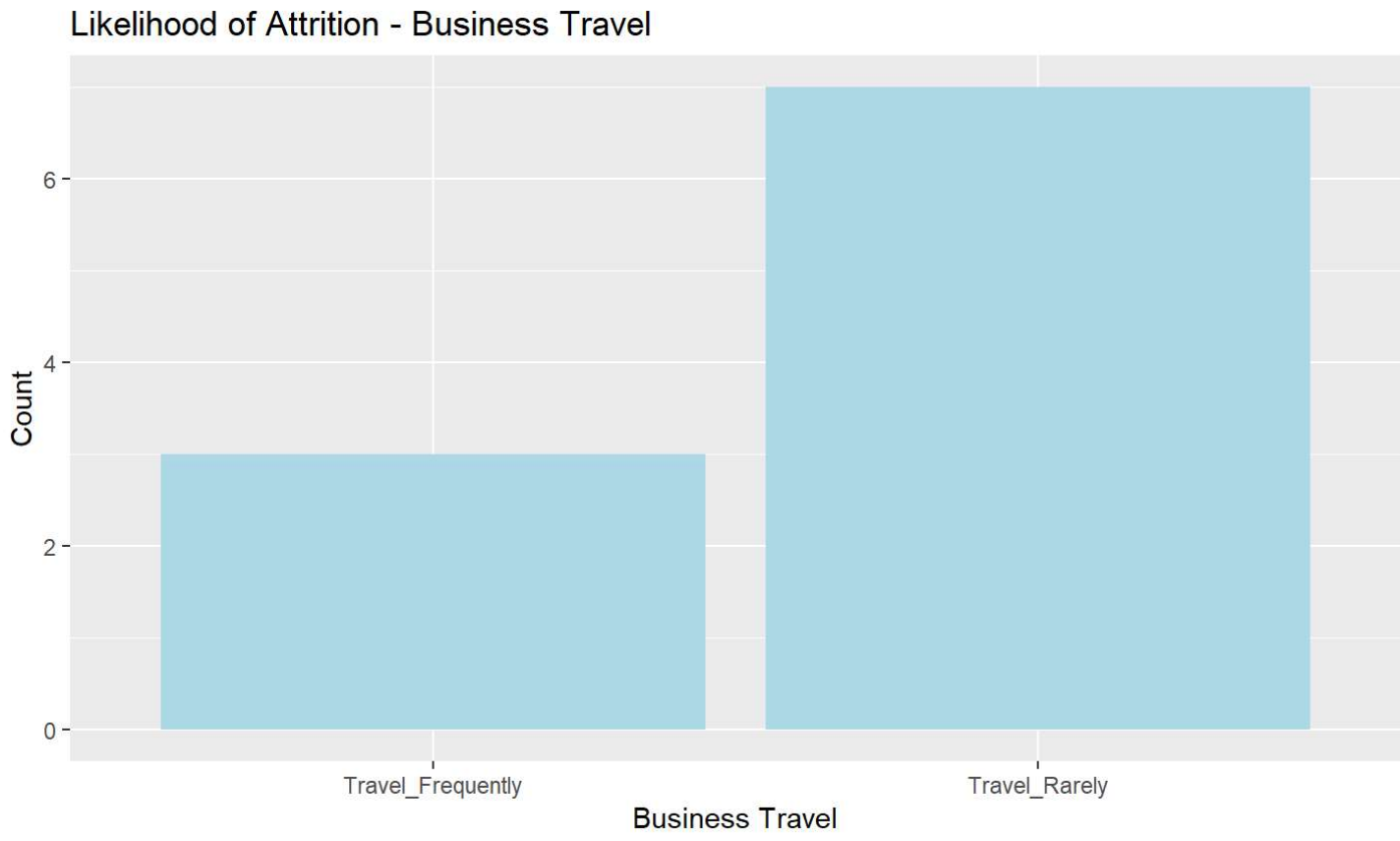
Likelihood of Attrition - Age

Of the 10 employees most likely to abscond, half of them were below age 30



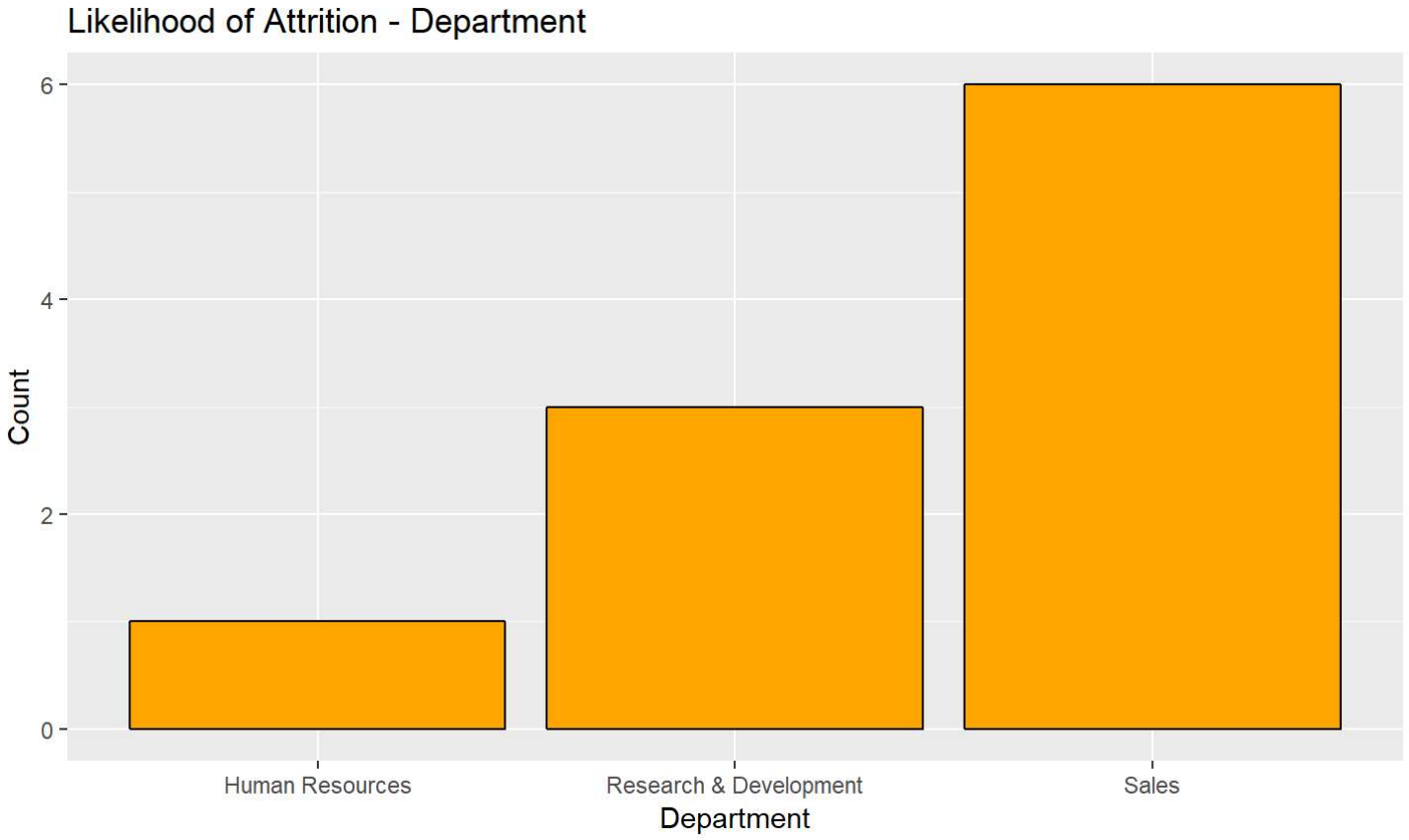
Likelihood of Attrition - Business Travel

Of the 10 employees most likely to abscond, 7 travel rarely and 3 travel frequently



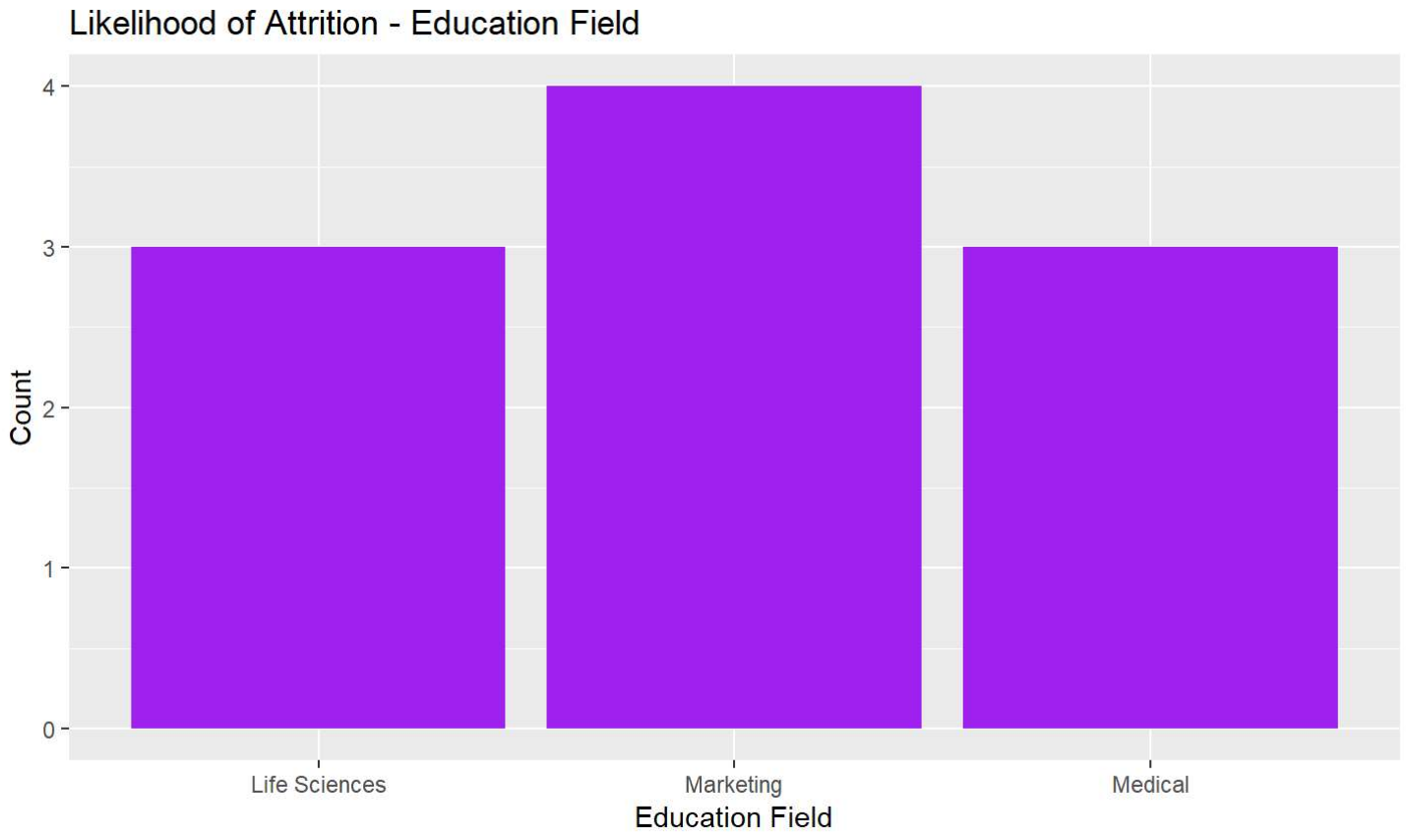
Likelihood of Attrition - Department

Of the 10 employees most likely to abscond, most were found in the Sales department follwed with a 3 in Research & Development and the rest were in the Human Resources department.



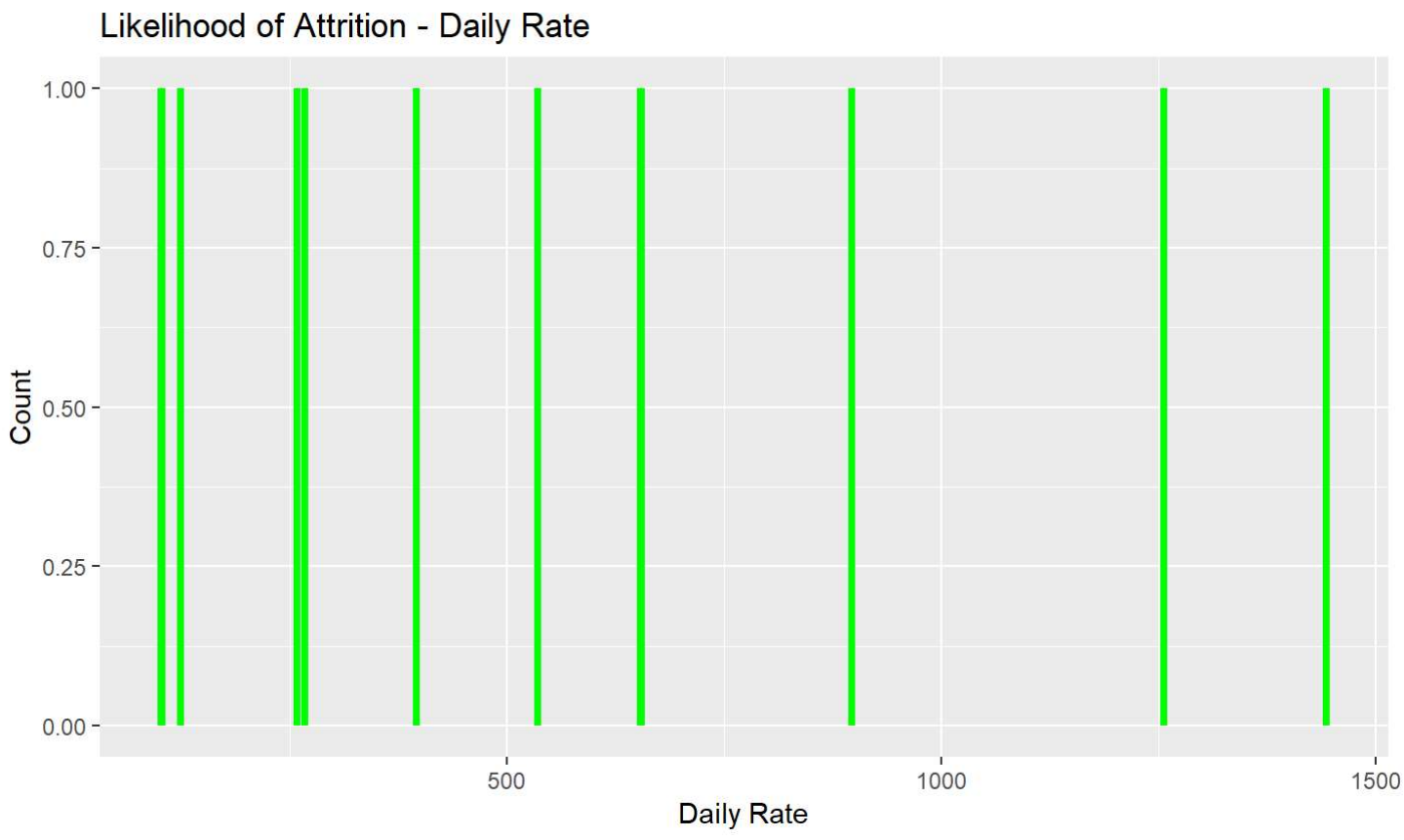
Likelihood of Attrition - Education Field

Of the 10 employees most likely to abscond, most had their education field as Marketing and the rest in Life Science and Medical



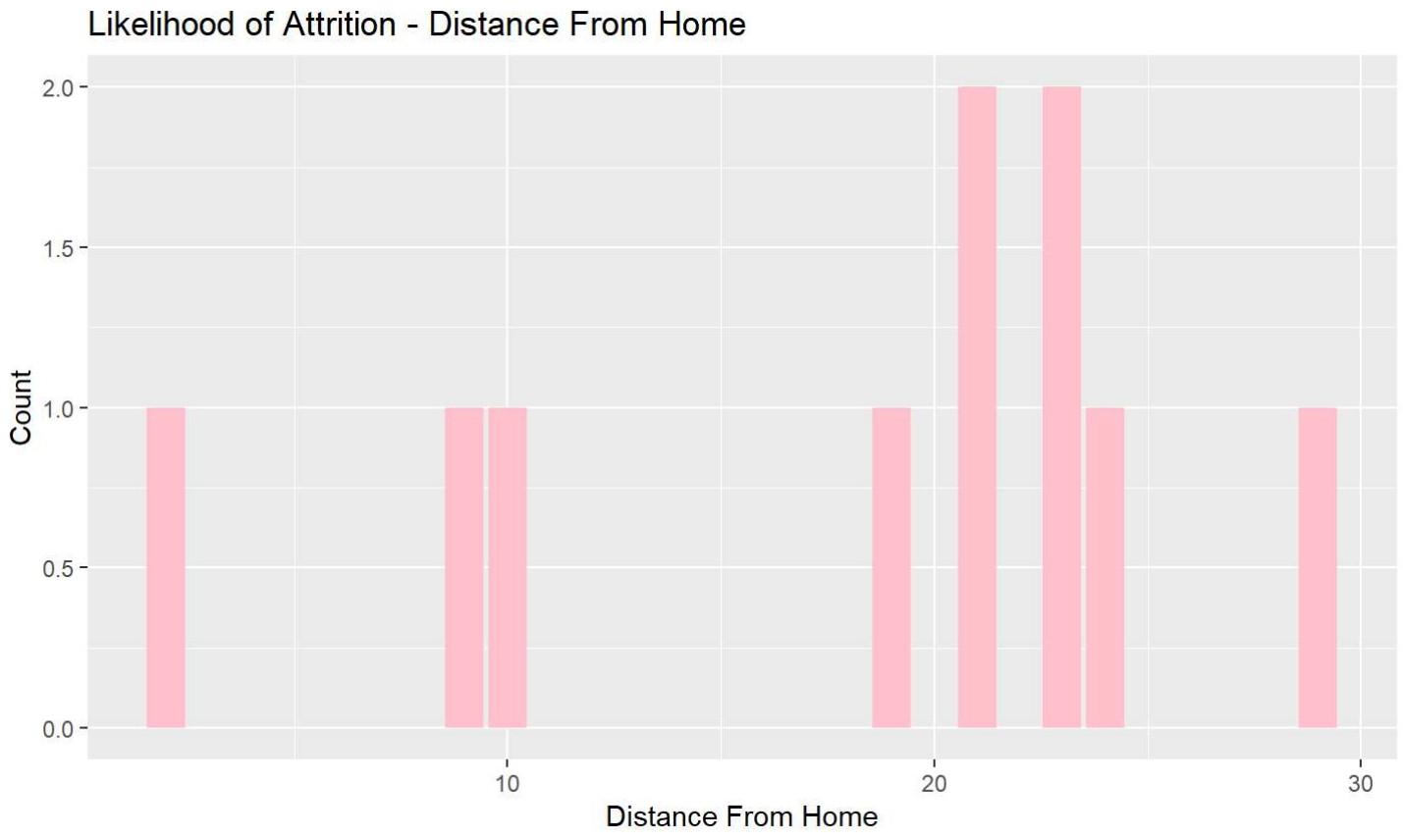
Likelihood of Attrition - Daily Rate

Of the 10 employees most likely to abscond, half the employees had a daily rate less than 500.



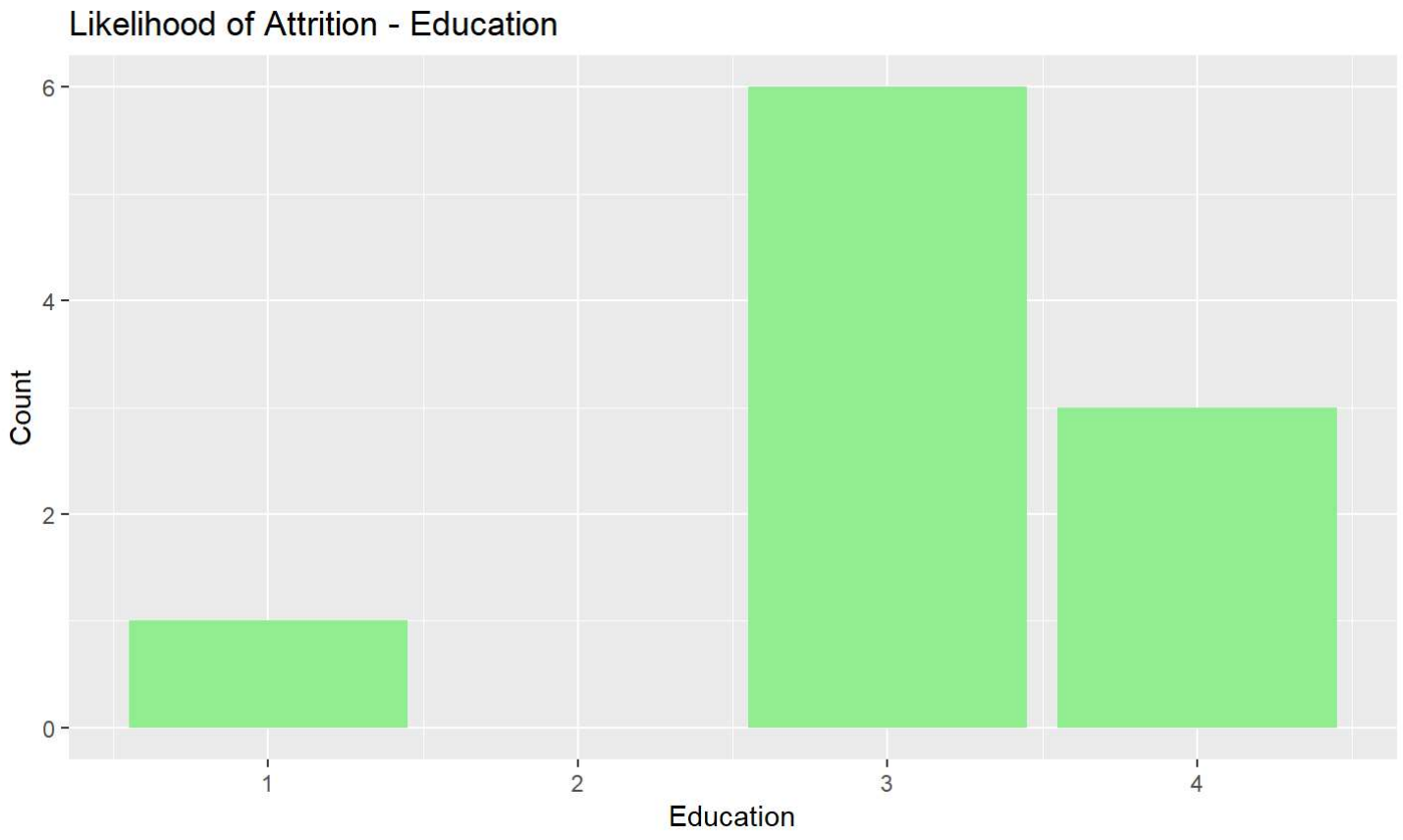
Likelihood of Attrition - Distance From Home

Of the 10 employees most likely to abscond, majority had to work at a distance over 20 from home.



Likelihood of Attrition - Education

Of the 10 employees most likely to abscond, 3 have an education level of 4, 6 have an education level of 3, and 1 has an education level of 1.

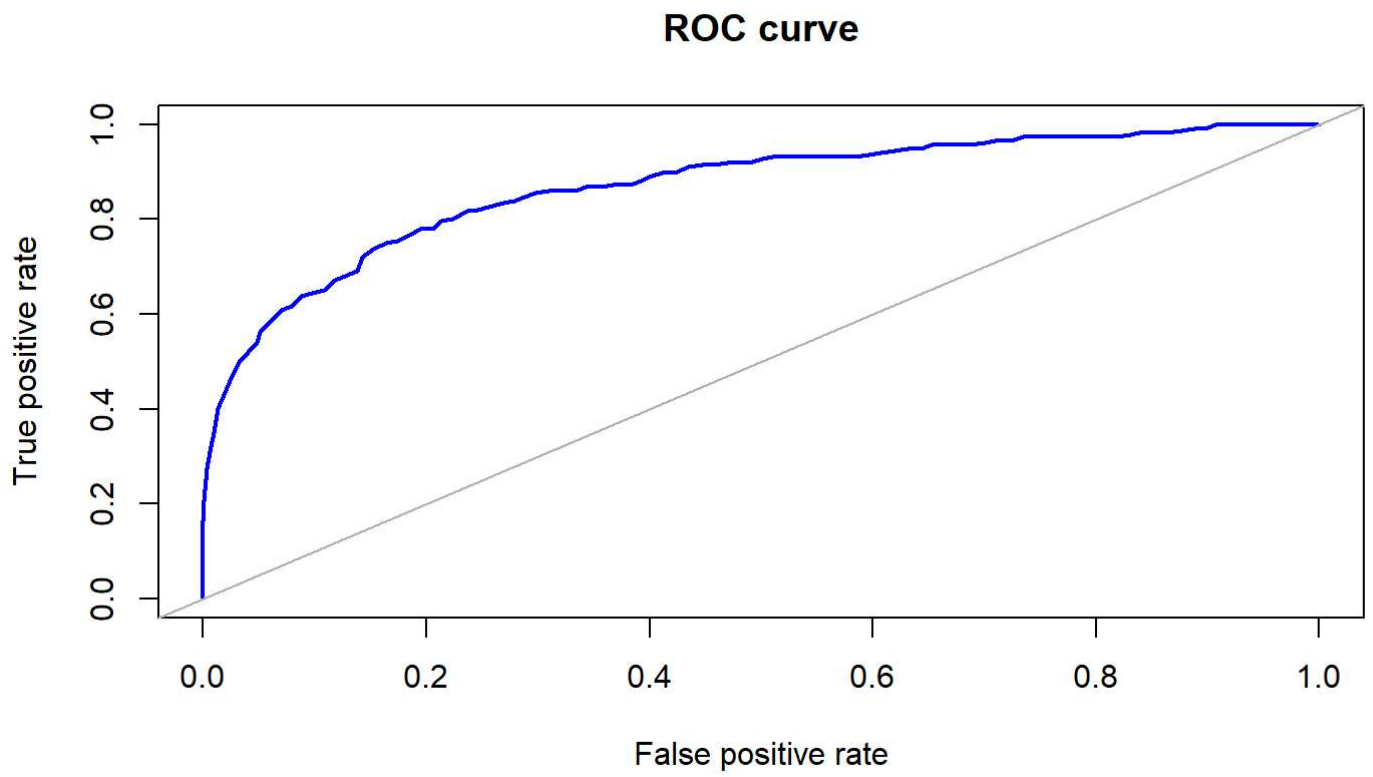


ROC CONFUSION MATRIX

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   No  Yes
##        No  1200  126
##        Yes   33  111
##
##               Accuracy : 0.892
##                 95% CI : (0.875, 0.907)
##    No Information Rate : 0.839
##    P-Value [Acc > NIR] : 0.000000003865643
##
##                  Kappa : 0.525
##  Mcnemar's Test P-Value : 0.000000000000296
##
##             Sensitivity : 0.973
##             Specificity : 0.468
##          Pos Pred Value : 0.905
##          Neg Pred Value : 0.771
##              Prevalence : 0.839
##          Detection Rate : 0.816
##    Detection Prevalence : 0.902
##       Balanced Accuracy : 0.721
##
##        'Positive' Class : No
##
```

ROC CURVE

The ROC Curve shows that the attrition prediction model can correctly distinguish between a true positive and false positive rate with an accuracy of 87.8%.




```
## Area under the curve (AUC): 0.869

##
## Call:
## accuracy.meas(response = as.factor(hr$Attrition), predicted = logit_predict,
##     threshold = 0.5)
##
## Examples are labelled as positive when predicted is greater than 0.5
##
## precision: 0.771
## recall: 0.468
## F: 0.291
```