Attrition Rate Analysis

Anastasia Volokhova

2025-05-06

```
library(correlationfunnel)
```

```
## Warning: package 'correlationfunnel' was built under R version 4.4.3
```

```
## == correlationfunnel Tip #2 ========================================================
## Clean your NA's prior to using `binarize()`.
## Missing values and cleaning data are critical to getting great correlations. :)
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.3
```

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

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
data <- read.csv('attrition.csv')
summary(data)
```

```
##       Age          Attrition         BusinessTravel       DailyRate
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0
##  Median :36.00   Mode  :character   Mode  :character   Median : 802.0
##  Mean   :36.92                                         Mean   : 802.5
##  3rd Qu.:43.00                                         3rd Qu.:1157.0
##  Max.   :60.00                                         Max.   :1499.0
##   Department        DistanceFromHome   Education     EducationField
##  Length:1470        Min.   : 1.000   Min.   :1.000   Length:1470
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000   Class :character
##  Mode  :character   Median : 7.000   Median :3.000   Mode  :character
##                     Mean   : 9.193   Mean   :2.913
##                     3rd Qu.:14.000   3rd Qu.:4.000
##                     Max.   :29.000   Max.   :5.000
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
##  Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1470
##  1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000           Class :character
##  Median :1     Median :1020.5   Median :3.000           Mode  :character
##  Mean   :1     Mean   :1024.9   Mean   :2.722
##  3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Max.   :1     Max.   :2068.0   Max.   :4.000
##    HourlyRate     JobInvolvement    JobLevel       JobRole
##  Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1470
##  1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character
##  Median : 66.00   Median :3.00   Median :2.000   Mode  :character
##  Mean   : 65.89   Mean   :2.73   Mean   :2.064
##  3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
##  Max.   :100.00   Max.   :4.00   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Min.   :1.000   Length:1470        Min.   : 1009   Min.   : 2094
##  1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.: 8047
##  Median :3.000   Mode  :character   Median : 4919   Median :14236
##  Mean   :2.729                      Mean   : 6503   Mean   :14313
##  3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:20462
##  Max.   :4.000                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18            OverTime         PercentSalaryHike
##  Min.   :0.000      Length:1470        Length:1470        Min.   :11.00
##  1st Qu.:1.000      Class :character   Class :character   1st Qu.:12.00
##  Median :2.000      Mode  :character   Mode  :character   Median :14.00
##  Mean   :2.693                                           Mean   :15.21
##  3rd Qu.:4.000                                           3rd Qu.:18.00
##  Max.   :9.000                                           Max.   :25.00
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.000     Median :3.000            Median :80    Median :1.0000
##  Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
```

```
##  3rd Qu.:3.000     3rd Qu.:4.000           3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.000     Max.   :4.000           Max.   :80    Max.   :3.0000
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000
##  Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 3.000     Median : 1.000          Median : 3.000
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :18.000     Max.   :15.000          Max.   :17.000
```

```
str(data)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel_Freque
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Researc
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : chr  "Female" "Male" "Male" "Female" ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : chr  "Sales Executive" "Research Scientist" "Laboratory Technician" "Res
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : chr  "Single" "Married" "Single" "Married" ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
```

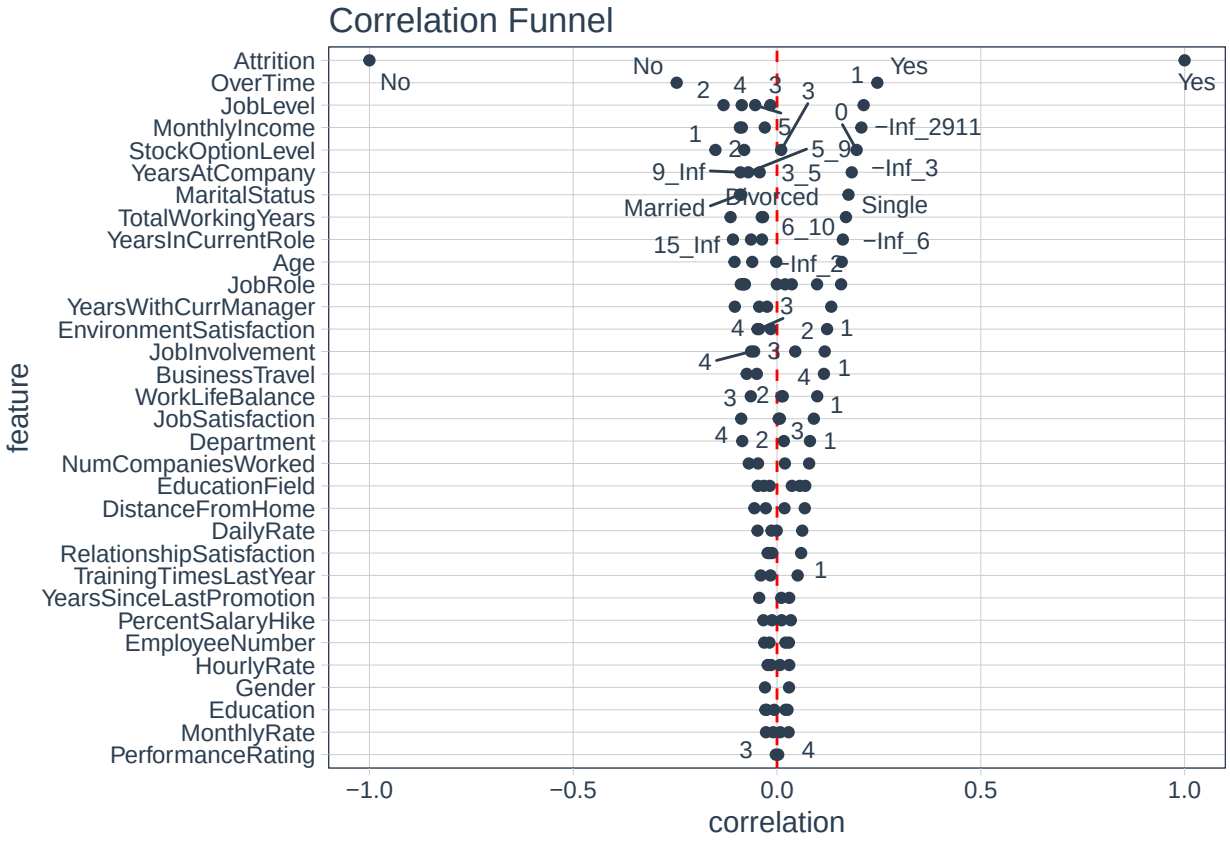
```
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```
# Overall company attrition rate
attrition_rate <-mean(data$Attrition == 'Yes')
print(paste('Attrition Rate: ', round(attrition_rate *100,2), '%'))
```

```
## [1] "Attrition Rate:  16.12 %"
```

```
# Correlation funnel
data <- data %>%
  mutate(Attrition = as.factor(Attrition))
data_binary <- data %>% binarize(n_bins = 4)
data_corr <- data_binary %>%
  correlate(target = Attrition__Yes)
data_corr %>%
  plot_correlation_funnel()
```

```
## Warning: ggrepel: 81 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
# Attrition prediction using logistic regression
data$Attrition <- as.factor(data$Attrition)
model <- glm(Attrition ~ Department + OverTime + JobLevel + MonthlyIncome,
             data = data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Attrition ~ Department + OverTime + JobLevel +
##     MonthlyIncome, family = binomial, data = data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -8.046e-01  3.791e-01  -2.122   0.0338 *
## DepartmentResearch & Development -4.566e-01  3.560e-01  -1.283   0.1997
## DepartmentSales                  3.012e-01  3.674e-01   0.820   0.4123
## OverTimeYes                      1.402e+00  1.519e-01   9.227   <2e-16 ***
## JobLevel                        -5.888e-01  2.353e-01  -2.503   0.0123 *
## MonthlyIncome                   -1.822e-05  5.804e-05  -0.314   0.7535
## ---
## 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: 1143.9  on 1464  degrees of freedom
## AIC: 1155.9
##
## Number of Fisher Scoring iterations: 5
```

```
exp(cbind(Odds_Ratio = coef(model), confint(model)))
```

```
## Waiting for profiling to be done...
```

```
##                                 Odds_Ratio     2.5 %    97.5 %
## (Intercept)                      0.4472513 0.2054513 0.9169235
## DepartmentResearch & Development  0.6334441 0.3243668 1.3231689
## DepartmentSales                  1.3514746 0.6762846 2.8825562
## OverTimeYes                      4.0635809 3.0202593 5.4820549
## JobLevel                         0.5549940 0.3485301 0.8772282
## MonthlyIncome                    0.9999818 0.9998672 1.0000950
```

```
# Predict probabilities
data$predicted_prob <- predict(model, type = "response")

# Now classify (AFTER predicted_prob exists)
data$predicted_class <- ifelse(data$predicted_prob > 0.5, "Yes", "No")
table(data$predicted_class)
```

```
##
##   No  Yes
## 1448   22
```

```
# View first few predictions
head(data$predicted_prob)
```

```
## [1] 0.40415522 0.07362469 0.38082733 0.37731441 0.12862018 0.12943936
```

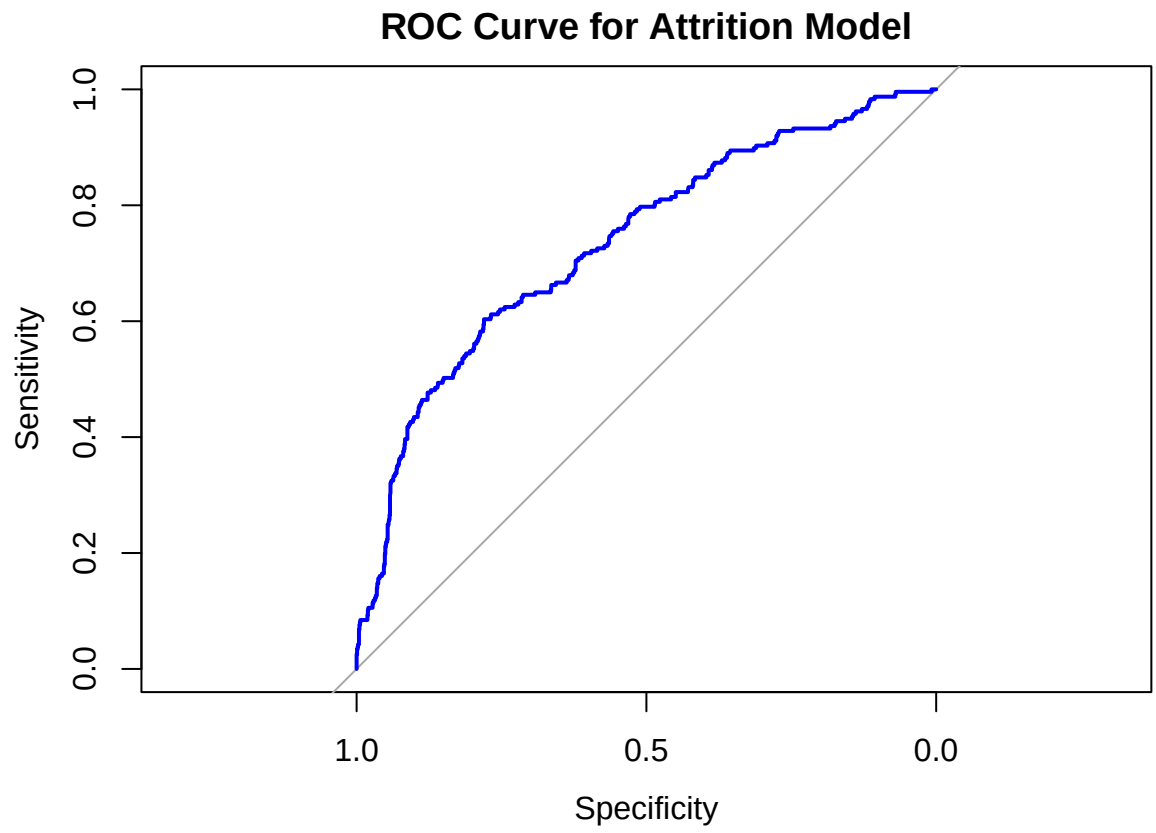
```r
# Predict on the same data
data$predicted_prob <- predict(model, type = "response")

# Create ROC curve object
roc_obj <- roc(data$Attrition, data$predicted_prob)
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
# Plot ROC curve
plot(roc_obj, col = "blue", main = "ROC Curve for Attrition Model")
```




```r
# Get AUC (Area Under the Curve)
auc(roc_obj)
```

```
## Area under the curve: 0.7349
```

```r
# Find best threshold (cutoff)
best_cutoff <- coords(roc_obj, "best", ret = "threshold")
print(best_cutoff)
```

```
##   threshold
## 1 0.1506612
```

```r
# Classify using the best cutoff (0.1507)
data$predicted_class_best <- ifelse(data$predicted_prob > 0.1507, "Yes", "No")

# Check counts
table(data$predicted_class_best)
```

```
##
##   No  Yes
## 1056  414
```

```r
# Confusion Matrix with best cutoff (0.1507)
conf_matrix <- confusionMatrix(
  factor(data$predicted_class_best, levels = c("Yes", "No")),
  factor(data$Attrition, levels = c("Yes", "No")),
  positive = "Yes"
)

# View the confusion matrix
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes 143 271
##        No   94 962
##
##                Accuracy : 0.7517
##                  95% CI : (0.7288, 0.7736)
##     No Information Rate : 0.8388
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2947
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.60338
##             Specificity : 0.78021
##          Pos Pred Value : 0.34541
##          Neg Pred Value : 0.91098
##              Prevalence : 0.16122
##          Detection Rate : 0.09728
##    Detection Prevalence : 0.28163
##       Balanced Accuracy : 0.69179
##
##        'Positive' Class : Yes
##
```

```r
# Extract Precision, Recall, and F1 Score from the confusion matrix
precision <- conf_matrix$byClass["Pos Pred Value"]
recall <- conf_matrix$byClass["Sensitivity"]
f1_score <- 2 * (precision * recall) / (precision + recall)
```

```r
# Print the results
cat("Precision:", precision, "\n")
```

```
## Precision: 0.3454106
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.6033755
```

```r
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.4393241
```

Company Attrition rate = 16.2%

Correlation funnel plot serves as a vusual guide to discern variables strongly correlated with attrition. The top variables OverTime, JobLevel, MonthlyIncome and StockOptionLevel. Employees who work overtime are more likely to leave the company. Junior-level employees, level 1 on a scale 1-5, exhibit higher likelihood of leaving - company may need to consider swift promotion of talented employees to mitigate attrition. Employees with monthly income of $2911 or lower are more likely to leave.

Area Under the Curve Interpretation: The AUC score of 0.7349 indicates that our model has moderate predictive power between attrition vs. non-attrition.

Threshold: The optimal threshold = 0.1507. This is the cut-off point above which predictions are classified as positive, meaning attrition will occur.

Model catches 60% of actual leavers(recall).When it flags someone as a leaver, it's only 34% right Accuracy (75%) is driven by the fact that people stay. F1 score (44%) shows that model needs balancing between catching leavers and avoiding false alarms. Generally, model is better at predicting who stays rather than leaves.