HR Case Study

Linh Nguyen

2023-03-25

```
hr.data <- read.csv("~/Downloads/hr-data-v2.csv")
```

```
logit.model <- glm(attrition ~ ., data = train.data, family = "binomial")
summary(logit.model)
```

```
##
## Call:
## glm(formula = attrition ~ ., family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0489  -0.5754  -0.3640  -0.1713   3.5204
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.020e+00  8.519e-01   1.197  0.23125
## emp.id                           -2.618e-05  4.332e-05  -0.604  0.54566
## environ.satisfaction             -3.912e-01  5.053e-02  -7.743 9.74e-15 ***
## job.satisfaction                 -3.796e-01  5.162e-02  -7.354 1.92e-13 ***
## work.life.balance                -3.447e-01  7.625e-02  -4.521 6.15e-06 ***
## perf.rating                      -1.916e-01  2.394e-01  -0.801  0.42342
## age                              -3.960e-02  8.890e-03  -4.454 8.42e-06 ***
## business.travelTravel_Frequently  1.332e+00  2.355e-01   5.654 1.56e-08 ***
## business.travelTravel_Rarely      5.160e-01  2.187e-01   2.359  0.01830 *
## departmentResearch & Development -5.360e-01  3.575e-01  -1.499  0.13376
## departmentSales                  -4.672e-01  3.716e-01  -1.257  0.20867
## dist.from.home                   -8.713e-03  7.095e-03  -1.228  0.21943
## education                        -1.908e-02  5.559e-02  -0.343  0.73146
## education.fieldLife Sciences     -7.461e-01  4.691e-01  -1.591  0.11172
## education.fieldMarketing         -1.181e+00  5.122e-01  -2.306  0.02112 *
## education.fieldMedical           -9.553e-01  4.703e-01  -2.031  0.04226 *
## education.fieldOther             -1.247e+00  5.202e-01  -2.398  0.01650 *
## education.fieldTechnical Degree  -1.065e+00  5.067e-01  -2.102  0.03558 *
## genderNon-binary                 -7.770e-01  1.203e+00  -0.646  0.51848
## genderOther                      -1.161e+01  4.351e+02  -0.027  0.97871
## genderPrefer not to answer       -1.305e+01  3.031e+02  -0.043  0.96567
## genderWoman                      -2.273e-01  1.169e-01  -1.945  0.05179 .
## job.level                        -6.581e-02  5.017e-02  -1.312  0.18960
## job.roleHuman Resources           3.094e-01  3.614e-01   0.856  0.39195
## job.roleLaboratory Technician     3.433e-01  2.355e-01   1.458  0.14489
## job.roleManager                  -1.163e-01  3.024e-01  -0.385  0.70051
## job.roleManufacturing Director   -5.194e-01  2.844e-01  -1.827  0.06777 .
## job.roleResearch Director         8.869e-01  2.836e-01   3.127  0.00177 **
```

```
## job.roleResearch Scientist        3.148e-01  2.288e-01   1.376  0.16884
## job.roleSales Executive           5.407e-01  2.265e-01   2.387  0.01699 *
## job.roleSales Representative     -1.249e-02  3.075e-01  -0.041  0.96760
## annual.income                    -1.266e-06  1.196e-06  -1.059  0.28957
## num.companies.worked              1.302e-01  2.388e-02   5.452 4.98e-08 ***
## pct.salary.raise                  3.706e-02  2.380e-02   1.557  0.11938
## total.working.yrs                -8.005e-02  1.596e-02  -5.016 5.29e-07 ***
## trainings.last.year              -1.865e-01  4.524e-02  -4.123 3.74e-05 ***
## years.at.company                  3.701e-02  2.289e-02   1.617  0.10590
## years.since.promotion             1.415e-01  2.551e-02   5.546 2.92e-08 ***
## years.current.mgr                -1.597e-01  2.930e-02  -5.450 5.03e-08 ***
## avg.hours.worked                  4.667e-01  4.024e-02  11.596  < 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: 2711.3  on 3012  degrees of freedom
## Residual deviance: 2134.7  on 2973  degrees of freedom
## AIC: 2214.7
##
## Number of Fisher Scoring iterations: 13
```

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(logit.model, type="text")
```

```
##
## ===============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                          attrition
## ---------------------------------------------------------------
## emp.id                                   -0.00003
##                                          (0.00004)
##
## environ.satisfaction                     -0.391***
##                                           (0.051)
##
## job.satisfaction                         -0.380***
##                                           (0.052)
##
## work.life.balance                        -0.345***
##                                           (0.076)
##
## perf.rating                               -0.192
##                                           (0.239)
##
## age                                      -0.040***
```

```
##                                          (0.009)
##
## business.travelTravel_Frequently         1.332***
##                                          (0.236)
##
## business.travelTravel_Rarely              0.516**
##                                          (0.219)
##
## departmentResearch                      Development
##                                          (0.357)
##
## departmentSales                          -0.467
##                                          (0.372)
##
## dist.from.home                           -0.009
##                                          (0.007)
##
## education                                -0.019
##                                          (0.056)
##
## education.fieldLife Sciences             -0.746
##                                          (0.469)
##
## education.fieldMarketing                 -1.181**
##                                          (0.512)
##
## education.fieldMedical                   -0.955**
##                                          (0.470)
##
## education.fieldOther                     -1.247**
##                                          (0.520)
##
## education.fieldTechnical Degree          -1.065**
##                                          (0.507)
##
## genderNon-binary                         -0.777
##                                          (1.203)
##
## genderOther                              -11.612
##                                         (435.134)
##
## genderPrefer not to answer               -13.046
##                                         (303.122)
##
## genderWoman                              -0.227*
##                                          (0.117)
##
## job.level                                -0.066
##                                          (0.050)
##
## job.roleHuman Resources                   0.309
##                                          (0.361)
##
## job.roleLaboratory Technician             0.343
```

```
##                                          (0.235)
##
## job.roleManager                          -0.116
##                                          (0.302)
##
## job.roleManufacturing Director           -0.519*
##                                          (0.284)
##
## job.roleResearch Director                0.887***
##                                          (0.284)
##
## job.roleResearch Scientist                0.315
##                                          (0.229)
##
## job.roleSales Executive                  0.541**
##                                          (0.227)
##
## job.roleSales Representative             -0.012
##                                          (0.307)
##
## annual.income                            -0.00000
##                                          (0.00000)
##
## num.companies.worked                     0.130***
##                                          (0.024)
##
## pct.salary.raise                          0.037
##                                          (0.024)
##
## total.working.yrs                        -0.080***
##                                          (0.016)
##
## trainings.last.year                      -0.187***
##                                          (0.045)
##
## years.at.company                          0.037
##                                          (0.023)
##
## years.since.promotion                    0.141***
##                                          (0.026)
##
## years.current.mgr                        -0.160***
##                                          (0.029)
##
## avg.hours.worked                         0.467***
##                                          (0.040)
##
## Constant                                  1.020
##                                          (0.852)
##
## ---------------------------------------------------------
## Observations                              3,013
## Log Likelihood                          -1,067.337
## Akaike Inf. Crit.                        2,214.674
```

```
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

The variables in the model that are statistically significant at $p < 0.01$ are: environ.satisfaction, job.satisfaction, work.life.balance, age, business.travelTravel_Frequently, job.roleResearch Director, num.companies.worked, total.working.yrs, trainings.last.year, years.since.promotion, years.current.mgr, avg.hours.worked. Thus, these are the most important variables that are likely to cause an employee's attrition.

environ.satisfaction, job.satisfaction, work.life.balance, age, total.working.yrs, trainings.last.year, years.current.mgr variables have negative coefficients in the model result. This suggests that these variables are important factors that positively affect employees' experiences in the company since, on average, when the values of these variables increase, it is less likely that the employees will leave the company (negative coefficients).

If employees are very dissatisfied with the company environment, the job, and the work-life balance situation of the company, they are likely to leave the company. If employees are not in their young ages, have many years of working, have a lot of trainings last year, and have a lot of years working with the current managers, they are less likely to leave the company.

On the other hand, business.travelTravel_Frequently, job.roleResearch Director, num.companies.worked, years.since.promotion, avg.hours.worked variables have positive coefficients in the model result. This indicates high values of these variables affect the employees' in a negative way, since on average, when the values of these variables increase, it is more likely that the employees will leave the company (positive coefficients).

If the job requires them to travel frequently, their job roles are Research Director, they have worked for a lot of companies, there have been many years since their last job promotion, and the number of hours they have to work is high, they are more likely to leave the company.

```
test.probs <- predict(logit.model, newdata = test.data, type = "response")
test.pred <- ifelse(test.probs > 0.5, 1, 0)
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
conf.mat <- confusionMatrix(factor(test.pred, levels = c(0,1)), factor(test.data$attrition, levels = c(
accuracy <- conf.mat$overall[1]
sensitivity <- conf.mat$byClass[1]
precision <- conf.mat$byClass[5]
accuracy
```

```
##  Accuracy
## 0.8578089
```

```
sensitivity
```

```
## Sensitivity
##    0.978957
```

```
precision
```

```
## Precision
## 0.8699187
```

The logistic regression model has an accuracy rate of 85.78%, sensitivity rate of 97.9%, and precision rate of 87%. Overall, the model is likely to do well in predicting which employees will leave the company.

85.78% of the predictions made by the model are correct.

The model is really good at capturing as many employees leaving the company as possible, as 97.9% of the number of employees predicted to leave the company actually left the company. This model is good when we do not care too much about sometimes we pull in some employees who will not leave the company.

The ratio of the number of employees who left the company to the number of employees predicted to leave the company is 87%. The model is not as precise as it is sensitive, but overall it is still a good model.

```
library(ranger)
rf.model <- ranger(attrition ~ ., data = train.data, importance = "permutation", probability = TRUE)
summary(rf.model)
```

```
##                             Length Class         Mode
## predictions                 6026   -none-        numeric
## num.trees                      1   -none-        numeric
## num.independent.variables      1   -none-        numeric
## mtry                           1   -none-        numeric
## min.node.size                  1   -none-        numeric
## variable.importance           23   -none-        numeric
## prediction.error               1   -none-        numeric
## forest                         9   ranger.forest list
## splitrule                      1   -none-        character
## treetype                       1   -none-        character
## call                           5   -none-        call
## importance.mode                1   -none-        character
## num.samples                    1   -none-        numeric
## replace                        1   -none-        logical
```

```
test.probs <- predict(rf.model, data = test.data)$predictions[,2]
test.pred <- ifelse(test.probs > 0.5, 1, 0)
```

```
conf.mat <- confusionMatrix(factor(test.pred, levels = c(0,1)), factor(test.data$attrition, levels = c(

# Compute the accuracy, sensitivity, and precision
accuracy <- conf.mat$overall[1]
sensitivity <- conf.mat$byClass[1]
precision <- conf.mat$byClass[5]

accuracy
```

```
## Accuracy
## 0.955711
```

```
sensitivity
```

```
## Sensitivity
##           1
```

```
precision
```

```
## Precision
## 0.9504348
```

This model performs better than the logistic regression model in this case, as all the accuracy, sensitivity, and precision rate by the random forest model are higher than those of the logistic regression model. Overall, 95.57% of the random forest's predictions are correct. Sensitivity is 1, which means the random forest model picked up everyone who actually left the company. Everyone who actually left the company was predicted as leaving the company by the model. It has 95.04% precision, meaning less than 5% of the employees predicted as leaving by the model actually stayed at the company. However, if we care more about not missing anyone

who will leave the company and tolerate that sometimes we may pick up people who actually will stay, then this model is perform very well because sensitivity is 1. Accuracy and precision are also high, indicating a good model.

```
write.csv(test.data, file = 'hr-predictions.csv', row.names = FALSE)
```

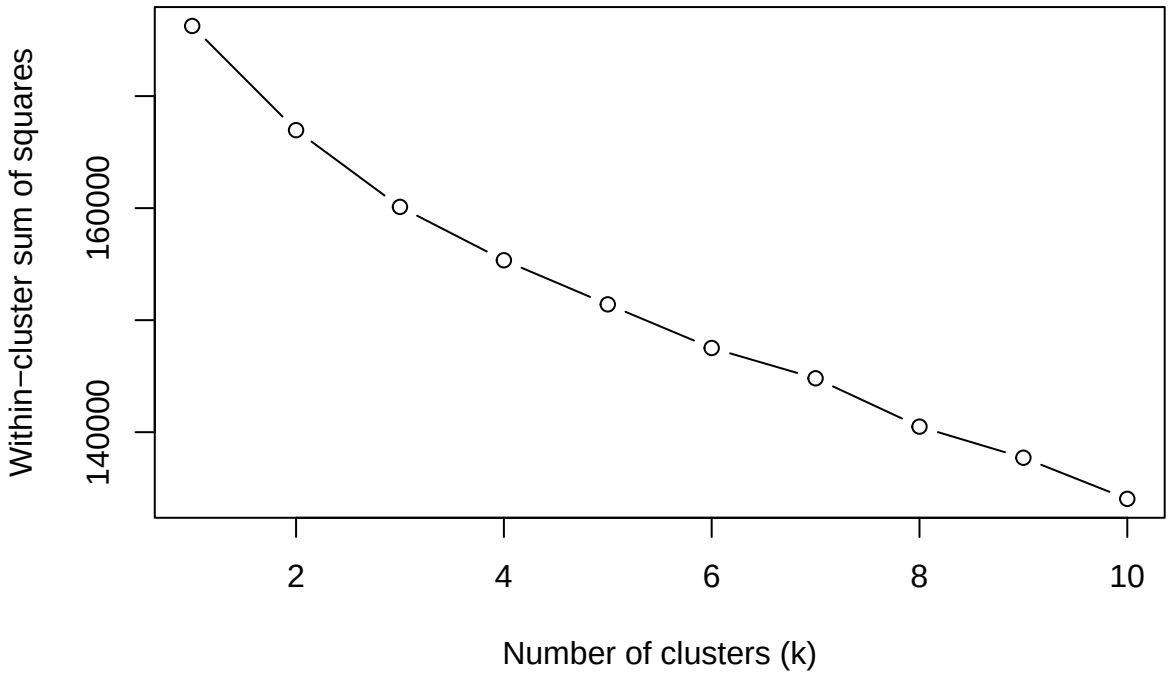
```
library(cluster)
library(dplyr)
```

```
##
## 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
```

```
cat_vars <- select(hr.data, business.travel, department, education.field, gender, job.role)
cat_matrix <- model.matrix(~.-1, data = cat_vars)
hr.data_new <- cbind(hr.data[, !(names(hr.data) %in% c('business.travel', 'department', 'education.field
kmeans.data <- scale(hr.data_new)
wss <- c()
for (i in 1:10) {
  kmeans.model <- kmeans(kmeans.data, centers = i, nstart = 10)
  wss[i] <- kmeans.model$tot.withinss
}
plot(1:10, wss, type = "b", xlab = "Number of clusters (k)", ylab = "Within-cluster sum of squares")
```



The elbow point of the plot when the within-cluster sum of squares starts to decrease more slowly is 4. This point represents the point of diminishing returns in terms of improving the clustering solution, and additional clusters do not offer substantial improvements. Thus, 4 is the optimal number of clusters to use for the K-means clustering method in this case.

```
kmeans.model <- kmeans(kmeans.data, centers = 4, nstart = 10)
kmeans.labels <- kmeans.model$cluster
hr.data$cluster <- kmeans.labels
```

```
table(kmeans.labels)
```

```
## kmeans.labels
##    1    2    3    4
##  777 1902 1048  573
```

```
library(dplyr)
kmeans.data_df <- as.data.frame(kmeans.data)
kmeans.data_df$cluster <- kmeans.labels
cluster_means <- kmeans.data_df %>%
  group_by(cluster) %>%
  summarize_all(mean)
```

Looking at the average values of the attributes in the different clusters of employees, we can see some information about different groups of employees. Cluster 1 has 777 employees. These employees generally have high working environment satisfaction, good work-life balance, high age, low job levels, high total years of working, high total years at the company, low number of companies they have worked at, long time since their last promotion, long time with their current manager. Employees in this cluster are most likely to stay at the company, as the average value of attrition for this cluster is the lowest among the clusters. These employees seem to be loyal employees who are older and have worked with the companies and their managers for a long time. They generally do not have high-leveled jobs, but they are satisfied with their position and the company's working environment.

Cluster 2 has 1902 employees. These employees generally are the newest employees in the company with the least years at the company, the least years with current managers, young age, a short distance from home, low company environment satisfaction, low job satisfaction, low work-life balance, high level of education, considerably high job level, average number of companies they have worked at, low performance ratings, low total working years, received the most training in the last year, are newly promoted. The employees in this cluster are the second most likely group of employees to leave the company, as their average attrition value is the second highest. These employees seem to be young employees who are newly hired by the company. They have worked at other companies, and they do not mind job hopping or changing companies. Their jobs are considerably high-leveled, and they have a high level of education, but they are not fully satisfied with their jobs and the companies. They may want to leave the company for a better job or working environment.

Cluster 3 has 1048 employees. These employees generally have a young age, high job satisfaction, good work-life balance, high job levels, and high education levels. However, the working environment satisfaction is low, the annual income is low, the number of companies they have worked at is low, the total number of working years is low, and their number of working hours is low. These also seem to be the young employees who are newly hired by the company, but they are more satisfied with their current jobs and they haven't had many worked experiences at other companies or anywhere. They think their jobs are high-leveled, and they don't need to work many hours and have good work-life balance. They don't seem to be frequent job-hoppers as they did not work with many companies before. These employees are the second most likely type of employees to stay at the company.

Cluster 4 has 573 employees. These employees generally have high job satisfaction, high performance rating, high annual income, and a high percentage of salary rise. However, they have low work-life balance, low environment satisfaction, high distance from job to home, a high number of companies they have worked for, low training received last year, and a high number of working hours. They have average age, the average total number of working years, average job levels, and average years at the company. These employees seem to be employees who are several years in their careers and are having stable careers. However, they may feel burned out or bored at their jobs, and they have to work considerably a lot of hours and have a low work-life balance. They do not stay at the company for too long to have a very loyal bond with the company, and they

seem to not learn a lot of new things from their jobs anymore given the lower training received. Since they have decent working experiences and job levels, they will likely want to switch to companies where they can learn new things and have a better work-life balance. These are the employee group that are most likely to leave the company.

Overall, the company can use both the logistic regression model and the random forest model to predict which employees will leave the company. The random forest model does better at predicting which employees leave the company as the accuracy, precision, and sensitivity rates are all higher than those of the logistic regression model. The sensitivity rate of the random forest model is particularly high, so it is not likely to miss anyone who actually leaves the company when predicting.

The company should also consider the results of the logistic regression and K-means clustering models to understand more about their employees and which factors lead to employee attrition. Factors such as environment satisfaction, job satisfaction, work-life balance, and the amount of training they received last year are factors that the company needs to work on to increase and improve as employees tend to consider these factors the most when making decisions about whether to leave the company. High values of these factors will likely encourage employees to stay at the companies.

The company should also pay attention to the employees' number of hours worked, the amount of travel required for the job, and the number of years since their last promotion and work on and try to decrease the values of those factors as they are important factors that contribute to employees' attrition too. Working many hours a day, traveling too frequently for the jobs, and waiting many years for their promotion will likely discourage the employees, making them want to leave the company.

The company should also try to improve the employees' loyalty to the company as their total years of working and their total years with the managers are also important factors that differentiate who may or may not leave the company. If they work long at the company and with their current managers, they are more likely to leave the company. Likewise, the number of companies they worked at is also a statistically significant variable, and the more companies they have worked at, the more likely they will leave the companies. Thus, besides working on factors to improve employee satisfaction, developing strong bonds with the employees is also essential to prevent employee attrition.