Employee Attrition Model

Authors: Levin Reichmuth, Jorit Studer and Taejun Moon

Module: Machine Learning I

Submitted on June 10th, 2022

SUPERVISOR: DR. MATTEO TANADINI, DANIEL MEISTER AND DR. ALBERTO PAGANINI

Lucerne University of Applied Sciences and Arts

Table of Contents

1 Introduction

Employees, according to Swaminathan & Hagarty (2020), are the foundation of any business. Its success is largely determined by the quality of its employees and their ability to stay with the company. Organizations confront a number of issues as a result of staff attrition:

1. Training new personnel is costly in terms of both money and time.
2. Potential to lose experienced employees
3. Productivity impact
4. Profitability impact

Therefore, IBM data scientists created a fictitious data set as a challenge for data scientists. Among the data types are metrics such as education level, job satisfaction, and commute distance. The dataset can be found on the company's GitHub account (IBM, 2019).

2 Methodology

The following topics are layed out through out this paper:

1. Linear Models
2. Extending the Linear Model: Non-linearity (GAM)
3. Extending the Linear Model: Generalised Linear Models (GLM)

- poisson as well as binominal

4. Support Vector Machines
5. Neural Networks
6. Optimisation

3 Data preparation

3.1 Data Transformation and Sanity Check

The code for this part is left out from the PDF due to its length…

3.2 Data Cleaning

```
emp_attrition <- emp_attrition %>% dplyr::select(-c(EmployeeCount, StandardHours, Over18))
```

- EmployeeCount (represents the head count which is 1 for all employee, hence drop this)
- StandardHours (StandardHours for all employee's is 80, therefore this data has a 9/80 work schedule. Hence, employees work 80 hours in 9 days. So not a standard as 5/42 as in switzerland, we drop this)
- Over18 (all employee's are 18 or above and it's capured in age, hence drop this variable)

3.3 Missing Value Check

```
# Do we have any missing values?
sapply(emp_attrition, function(x) all(is.na(x) | x == '' ))
```

There are no missing values in this dataset.

3.4 Overview of Dataset

Table 1: Summary Numeric Variables

	N	Mean	SD	Min	Q1	Median	Q3	Max
Age	1470	36.92	9.14	18	30	36.0	43	60
DailyRate	1470	802.49	403.51	102	465	802.0	1157	1499
DistanceFromHome	1470	9.19	8.11	1	2	7.0	14	29
EmployeeNumber	1470	1024.87	602.02	1	491	1020.5	1556	2068
HourlyRate	1470	65.89	20.33	30	48	66.0	84	100
MonthlyIncome	1470	6502.93	4707.96	1009	2911	4919.0	8380	19999
MonthlyRate	1470	14313.10	7117.79	2094	8045	14235.5	20462	26999
NumCompaniesWorked	1470	2.69	2.50	0	1	2.0	4	9
PercentSalaryHike	1470	15.21	3.66	11	12	14.0	18	25
TotalWorkingYears	1470	11.28	7.78	0	6	10.0	15	40
TrainingTimesLastYear	1470	2.80	1.29	0	2	3.0	3	6
YearsAtCompany	1470	7.01	6.13	0	3	5.0	9	40
YearsInCurrentRole	1470	4.23	3.62	0	2	3.0	7	18
YearsSinceLastPromotion	1470	2.19	3.22	0	0	1.0	3	15
YearsWithCurrManager	1470	4.12	3.57	0	2	3.0	7	17

Table 2: Summary Factor Variables

	Level	N	%		Level	N	%
Attrition	No	1233	83.9		5	69	4.7
	Yes	237	16.1	JobRole	Healthcare Representative	131	8.9
BusinessTravel	None	1043	71.0		Human Resources	52	3.5
	Rarely	150	10.2		Laboratory Technician	259	17.6
	Frequently	277	18.8		Manager	102	6.9
Department	Sales	63	4.3		Manufacturing Director	145	9.9
	R&D	961	65.4		Research Director	80	5.4
	HR	446	30.3		Research Scientist	292	19.9
Education	Below College	170	11.6		Sales Executive	326	22.2
	College	282	19.2		Sales Representative	83	5.6
	Bachelor	572	38.9	JobSatisfaction	Low	289	19.7
	Master	398	27.1		Medium	280	19.0
	Doctor	48	3.3		High	442	30.1
EducationField	Human Resources	27	1.8		Very High	459	31.2
	Life Sciences	606	41.2	MaritalStatus	Divorced	327	22.2
	Marketing	159	10.8		Married	673	45.8
	Medical	464	31.6		Single	470	32.0
	Other	82	5.6	OverTime	No	1054	71.7
	Technical Degree	132	9.0		Yes	416	28.3
EnvironmentSatisfaction	Low	284	19.3	PerformanceRating	Excellent	1244	84.6
	Medium	287	19.5		Outstanding	226	15.4
	High	453	30.8	RelationshipSatisfaction	Low	276	18.8
	Very High	446	30.3		Medium	303	20.6
Gender	Female	588	40.0		High	459	31.2
	Male	882	60.0		Very High	432	29.4
JobInvolvement	Low	83	5.6	StockOptionLevel	0	631	42.9
	Medium	375	25.5		1	596	40.5
	High	868	59.0		2	158	10.7
	Very High	144	9.8		3	85	5.8
JobLevel	1	543	36.9	WorkLifeBalance	Bad	80	5.4
	2	534	36.3		Good	344	23.4
	3	218	14.8		Better	893	60.7
	4	106	7.2		Best	153	10.4

3.5 Splitting Dataset (Stratification)

There is significant class imbalance in the variable Attrition (84 / 16) as can be seen in Table 2, which makes sense since most employees want to work at the company. However it is important to stratify the train and test split so that we receive a more realist estimate on how our model is going to perform.

```
set.seed(111)
emp_attrition_split <- initial_split(emp_attrition, prop = 0.80, strata = "Attrition")
emp_attrition_train <- training(emp_attrition_split) # i.e. 986 / 189 = 5.22
emp_attrition_test  <- testing(emp_attrition_split) # i.e. 248 / 48 = 5.17
```

In this analysis we are splitting the data set with 80% training and 20% test / validation. As can be observed above we have an almost equal ratio of yes to no in the training as well as the testing data set due to stratification.

4 Exploration

5 Linear Regression

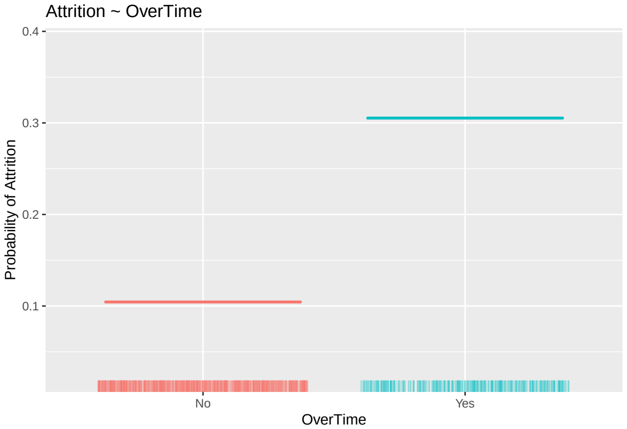
6 Generalised Linear Models

6.1 Simple Logistic Regression

```
attrition_model1 <- glm(Attrition ~ OverTime, family = "binomial", data = emp_attrition)
summary(attrition_model1)$coefficients
```

```
##              Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -2.149646  0.1007431 -21.337888 5.052603e-101
## OverTimeYes  1.327406  0.1465721   9.056333  1.349094e-19
```

Not unexpectedly overtime seems to play a relevant role. Indeed, its p-value is highly significant.

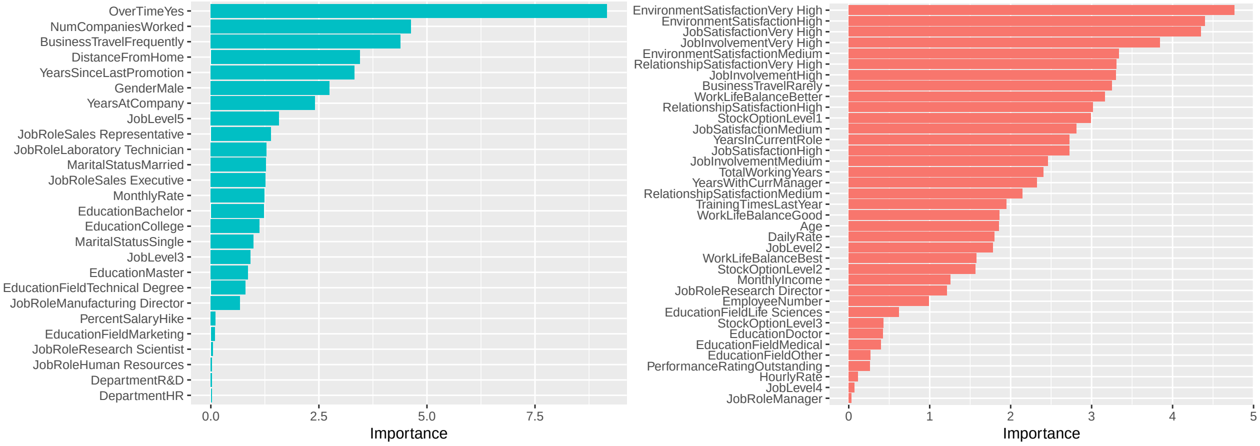


```
exp(coef(attrition_model1)["OverTimeYes"])
```

```
## OverTimeYes
##    3.771249
```

The odds of someone leaving the company with overtime are about ~3.8 times higher than the odds for no overtime in this simple model.

6.2 Multiple logistic regression

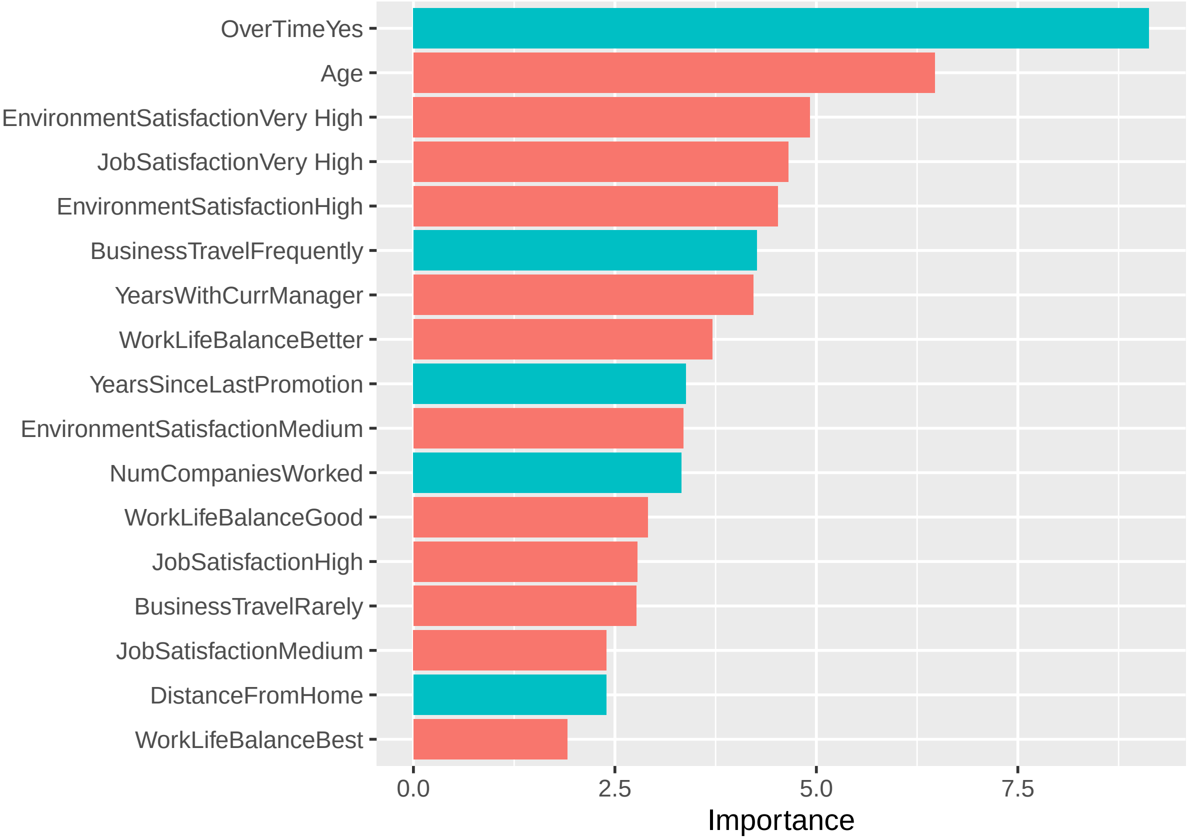


In the above plots we can observe the most important variables (Variable Importance) to predict employment attrition according to the absolute value of the z-statistic for each coefficient in the dataset. Moreover the importance of independent variables are colored to indicate increasing (blue) or decreasing (red) risk of employee attrition. We again observe that OverTime seems to be highly correlated with employee attrition in this data set. Moreover, EnviromentSatisfaction and JobSatisfaction seem to be also be critical, which would make sense since we are talking about employment attrition.

Based on the exploratory data analysis, the previous section as well as the above variable importance scores we are trying to fit a better multiple logistic regression model.

```
glm.model.2 <- glm(Attrition ~ OverTime + EnvironmentSatisfaction + NumCompaniesWorked +
                       JobSatisfaction + BusinessTravel + DistanceFromHome +
                       WorkLifeBalance + Age + YearsWithCurrManager + YearsSinceLastPromotion,
                   family = "binomial", data = emp_attrition_train)
```

The new model seems to have quite a good fit with all independent variables having significant p-values.



The new variable importance plot seems to have ranked Age a lot higher than before while overtime still seems to remain a large main effect. Thus, there might be interactions in these variable and we should start developing the model.

6.3 Model Development

The code for the model development is hidden since it is to large for this paper.

By using the drop1 function we have added and remove significant interactions. Finally we end up with significant interaction inclusion of JobSatifaction with Age, Number of Companies Worked for, Work Life Balance and Business Travel Frequency.

```
final.glm <- glm(Attrition ~ EnvironmentSatisfaction + NumCompaniesWorked + JobSatisfaction +
    BusinessTravel + DistanceFromHome + OverTime + YearsAtCompany +
    PercentSalaryHike + WorkLifeBalance + Age + YearsWithCurrManager +
    YearsSinceLastPromotion +
    JobSatisfaction:Age + JobSatisfaction:NumCompaniesWorked +
    JobSatisfaction:WorkLifeBalance + JobSatisfaction:BusinessTravel, family = "binomial", data = emp_a
# exp(coef(final.glm))
```

Interpretation of the Logistic Regression

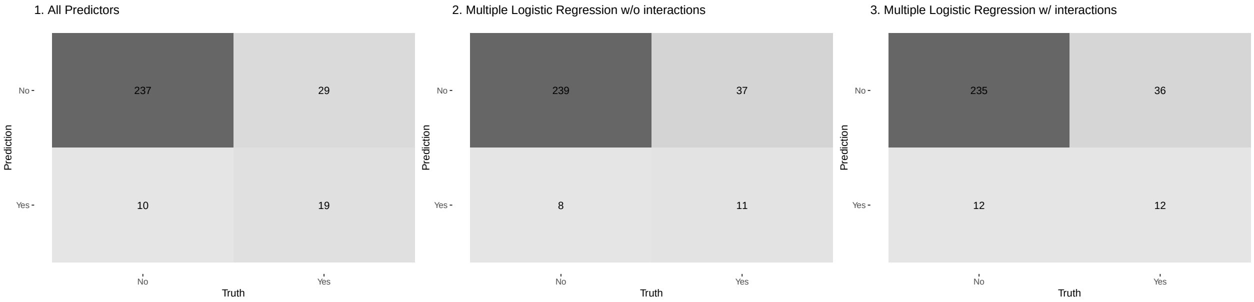
The odds of someone leaving the company

- with overtime are about ~5.6 times higher than the odds for no overtime
- with a ‘very high’ environment satisfaction are lower than the ones of low environment satisfaction by ~0.31 times.
- with a ‘high’ environment satisfaction are lower than the ones of low environment satisfaction by ~0.37 times.
- with a ‘medium’ environment satisfaction are lower than the ones of low environment satisfaction by ~0.38 times.
- are multiplied by 1.19 i.e increasing for each additional company the employee has worked for
- with a ‘very high’ job satisfaction are lower than the ones of low job satisfaction by ~0.009 times.
- with a ‘high’ job satisfaction are lower than the ones of low job satisfaction by ~0.12 times.
- with a ‘medium’ job satisfaction are lower than the ones of low job satisfaction by ~0.04 times.
- having to travel for work ‘rarely’ increases the risk by ~1.38 times
- having to travel for work ‘frequently’ increases the risk by ~2.04 times
- are multiplied by 1.03 i.e increasing for each unit of distance between work and home
- are multiplied by 0.97 i.e slightly decreasing for each additional year the employee has worked at the company
- are multiplied by 0.99 i.e slightly decreasing for each percentage in salary hike an employee has received
- with a ‘good’ work life balance score are lower than the ones of ‘bad’ work life balance score by ~0.11 times.
- with a ‘better’ work life balance score are lower than the ones of ‘bad’ work life balance score by ~0.11 times.
- with a ‘best’ work life balance score are lower than the ones of ‘bad’ work life balance score by ~0.12 times.
- are multiplied by 0.9 i.e decreasing for each additional year of age an employee has
- are multiplied by 0.86 i.e decreasing for each additional year an employee has worked for the same manager
- are multiplied by 1.17 i.e increasing for each year an employee has not received a promotion

Summary

Overtime seems to increase the risk of attrition by almost six times as much and is by far the most highly critical attribute on whether an employee continues to stay at the company or not. It does not seem to matter to much how well the *job environment* score is as long as it is above 'low' as they almost equally decrease the risk of someone leaving the company. The amount of *business traveling* an employee has to do seems to play an important role as well as the risk of someone quitting the company increases by two times if said person has frequently travel for work. *Age* seems to play a role as well as young employee seem to leave the company more often than old employees. Employee seem to leave the company less often if they are not bound to re-organisations i.e. have the same *manager* for an extended period of time. For each year an employee has not received a *promotion* the risk of them leaving increases.

6.4 Confusion Matrix



1. When looking at the first plot using cross validation & including all predictors we get the following scores.

   Specificity = True Negatives / (True Negatives + False Positives)

   - i.e. 237 / (10 + 237) = 0.95
   - 95% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Sensitivity = True Positives / (True Positives + False Negatives)

   - i.e. 19 / (29 + 19) = 0.4
   - 40% of the people leaving the company were correctly identified by the Logistic Regression model.

2. Without interactions using a much simpler model we actually get a better specificity, but can only predict 23% of the people who are actually leaving the company. In other words the True Positive Rate which we are looking for is significantly worse.

   Specificity = 239 / (8 + 239) = 0.97

   - 97% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Sensitivity = 11 / (37 + 11) = 0.23

   - 23% of the people leaving the company were correctly identified by the Logistic Regression model.

3. When including the interactions we are able to predict the True Positive Rate a bit better however we sacrifice some of the specificity

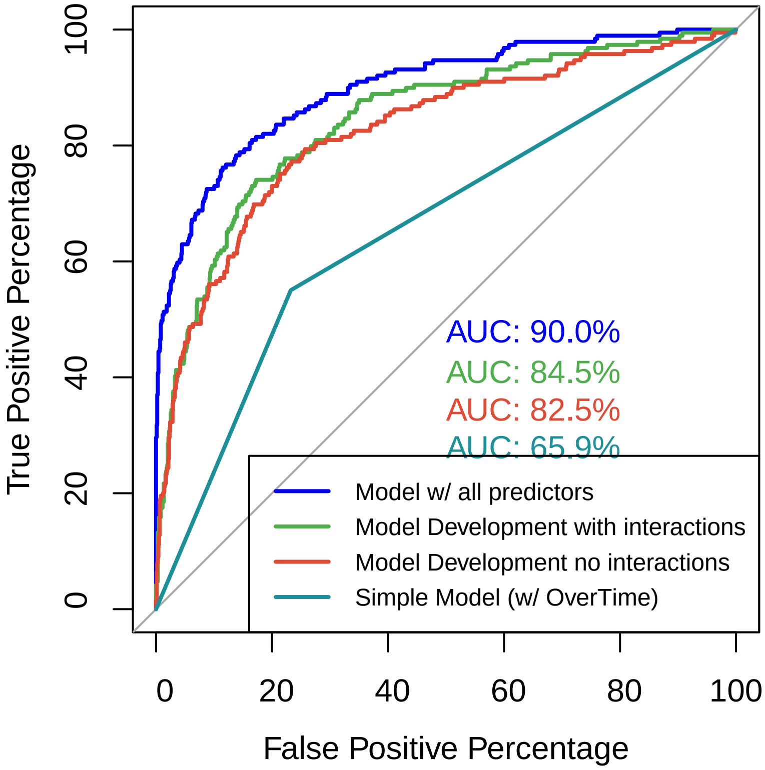
   Specificity = 235 / (12 + 235) = 0.95

   - 95% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Sensitivity = 12 / (36 + 12) = 0.25

   - 25% of the people leaving the company were correctly identified by the Logistic Regression model.

6.5 Receiver Operating Characteristics



The above plot illustrates how the True Positive Rate (Sensitivity) behaves in relation with the False Positive Rate (1-Specificity). In this paper we want to maximize the amount of correct classifications of people leaving the company i.e. True Positive Rate. Thus we can take away from the above plot that including the interactions indeed makes sense, not solely to reach a higher Area Under the Curve (AUC), but also since a threshold of around ~90% on the True Positive Rate can reach a very similar False Positive Rate (around 25-30%) with a much simpler model (12 independent variables & 4 interactions vs 31 total independent variables).

7 Conclusion

Session Information

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.3.1
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] pROC_1.18.0      arm_1.12-2       lme4_1.1-27.1    Matrix_1.3-4    
##  [5] MASS_7.3-54      yardstick_0.0.9  forcats_0.5.1    stringr_1.4.0   
##  [9] purrr_0.3.4      readr_2.1.2      tidyr_1.2.0      tibble_3.1.6    
## [13] tidyverse_1.3.1  rsample_0.1.1    vip_0.3.2        caret_6.0-92    
## [17] lattice_0.20-45  mgcv_1.8-38      nlme_3.1-155     kableExtra_1.3.4
## [21] papeR_1.0-5      xtable_1.8-4     car_3.0-12       carData_3.0-5   
## [25] ggplot2_3.3.6    dplyr_1.0.9      bookdown_0.24   
## 
## loaded via a namespace (and not attached):
##  [1] minqa_1.2.4          colorspace_2.0-2     ellipsis_0.3.2      
##  [4] class_7.3-19         fs_1.5.2             proxy_0.4-26        
##  [7] rstudioapi_0.13      farver_2.1.0         listenv_0.8.0       
## [10] furrr_0.3.0          prodlim_2019.11.13   fansi_1.0.2         
## [13] lubridate_1.8.0      xml2_1.3.3           codetools_0.2-18    
## [16] splines_4.1.2        knitr_1.37           jsonlite_1.7.3      
## [19] nloptr_2.0.0         broom_0.7.12         dbplyr_2.1.1        
## [22] compiler_4.1.2       httr_1.4.2           backports_1.4.1     
## [25] assertthat_0.2.1     fastmap_1.1.0        cli_3.2.0           
## [28] htmltools_0.5.2      tools_4.1.2          coda_0.19-4         
## [31] gtable_0.3.0         glue_1.6.1           reshape2_1.4.4      
## [34] gmodels_2.18.1       Rcpp_1.0.8           cellranger_1.1.0    
## [37] vctrs_0.4.1          gdata_2.18.0         svglite_2.1.0       
## [40] iterators_1.0.13     timeDate_3043.102    gower_0.2.2         
## [43] xfun_0.29            globals_0.14.0       rvest_1.0.2         
## [46] lifecycle_1.0.1      gtools_3.9.2         future_1.25.0       
## [49] scales_1.1.1         ipred_0.9-12         hms_1.1.1           
## [52] parallel_4.1.2       yaml_2.3.4           gridExtra_2.3       
## [55] rpart_4.1-15         stringi_1.7.6        highr_0.9           
## [58] foreach_1.5.1        e1071_1.7-9          boot_1.3-28         
## [61] hardhat_0.2.0        lava_1.6.10          rlang_1.0.2         
## [64] pkgconfig_2.0.3      systemfonts_1.0.3    evaluate_0.14       
## [67] labeling_0.4.2       recipes_0.2.0        tidyselect_1.1.1    
## [70] parallelly_1.30.0    plyr_1.8.6           magrittr_2.0.2      
## [73] R6_2.5.1             generics_0.1.2       DBI_1.1.2           
## [76] pillar_1.7.0         haven_2.4.3          withr_2.4.3         
## [79] survival_3.2-13      abind_1.4-5          nnet_7.3-16
## [82] future.apply_1.8.1   modelr_0.1.8         crayon_1.5.0
## [85] utf8_1.2.2           tzdb_0.2.0           rmarkdown_2.11
## [88] grid_4.1.2           readxl_1.3.1         data.table_1.14.2
## [91] ModelMetrics_1.2.2.2 reprex_2.0.1         digest_0.6.29
## [94] webshot_0.5.2        stats4_4.1.2         munsell_0.5.0
## [97] viridisLite_0.4.0
```

References

IBM. (2019). *IBM HR analytics employee attrition & performance.* https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv

Swaminathan, S., & Hagarty, R. (2020). *IBM HR analytics employee attrition & performance* (2nd ed.). IBM. https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/