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

4 Exploration

5 Linear Regression

```
lm.emp_attrition <- lm(as.numeric(Attrition) ~ . , data = emp_attrition)

lm.summary <- summary(lm.emp_attrition)
lm.drop1 <- drop1(lm.emp_attrition, test = "F")
```

Table 3: Summary Significant Variables

Drop1	Pr(>|t|)	Summary
Age	0.0069490	Age
BusinessTravel	0.0000017	BusinessTravel
DistanceFromHome	0.0001102	DistanceFromHome
EducationField	0.0076906	EnvironmentSatisfaction
EnvironmentSatisfaction	0.0000003	JobSatisfaction
JobInvolvement	0.0000009	NumCompaniesWorked
JobLevel	0.0000017	OverTime
JobRole	0.0112227	RelationshipSatisfaction
JobSatisfaction	0.0000144	StockOptionLevel
NumCompaniesWorked	0.0000011	WorkLifeBalance
OverTime	0.0000000	YearsInCurrentRole
RelationshipSatisfaction	0.0038475	YearsSinceLastPromotion
StockOptionLevel	0.0020909	YearsWithCurrManager
TrainingTimesLastYear	0.0583434	NA
WorkLifeBalance	0.0002238	NA
YearsAtCompany	0.0688877	NA
YearsInCurrentRole	0.0533036	NA
YearsSinceLastPromotion	0.0065725	NA
YearsWithCurrManager	0.0409552	NA

These are the dependent variables which are significant or weakly significant in Drop1, but not in the Summary statistics.

```
significance.table$Drop1[!(significance.table$Drop1 %in% significance.table$Summary)]
```

```
## [1] "EducationField"        "JobInvolvement"        "JobLevel"
## [4] "JobRole"               "TrainingTimesLastYear" "YearsAtCompany"
```

TODO: Check this?

```
lm.attrition <- lm(as.numeric(Attrition) ~ Age + BusinessTravel + DistanceFromHome +
                     EducationField + EnvironmentSatisfaction + JobInvolvement +
                     JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked +
                     OverTime + RelationshipSatisfaction + StockOptionLevel +
                     TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany +
                     YearsInCurrentRole + YearsSinceLastPromotion +
                     YearsWithCurrManager, data = emp_attrition)
```

A first simple linear regression model with and R^2 around 30%. However, it doesn't really make sense to use this model with Attrition being True or False.

5.1 Model Development (Interactions)

```
lm.attrition.interactions <- lm(as.numeric(Attrition) ~ Age + BusinessTravel +
    DistanceFromHome + EducationField + EnvironmentSatisfaction + JobInvolvement +
    JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked + OverTime +
    RelationshipSatisfaction + StockOptionLevel + TrainingTimesLastYear +
    WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager +
    Age:JobSatisfaction + Age:StockOptionLevel + BusinessTravel:YearsAtCompany +
    DistanceFromHome:OverTime + EnvironmentSatisfaction:JobInvolvement +
    EnvironmentSatisfaction:WorkLifeBalance + JobLevel:OverTime +
    JobRole:OverTime + JobRole:WorkLifeBalance + NumCompaniesWorked:OverTime +
    NumCompaniesWorked:StockOptionLevel + OverTime:RelationshipSatisfaction +
    OverTime:StockOptionLevel + WorkLifeBalance:YearsSinceLastPromotion,
    data = emp_attrition)

drop1(lm.attrition.interactions, test="F")
```

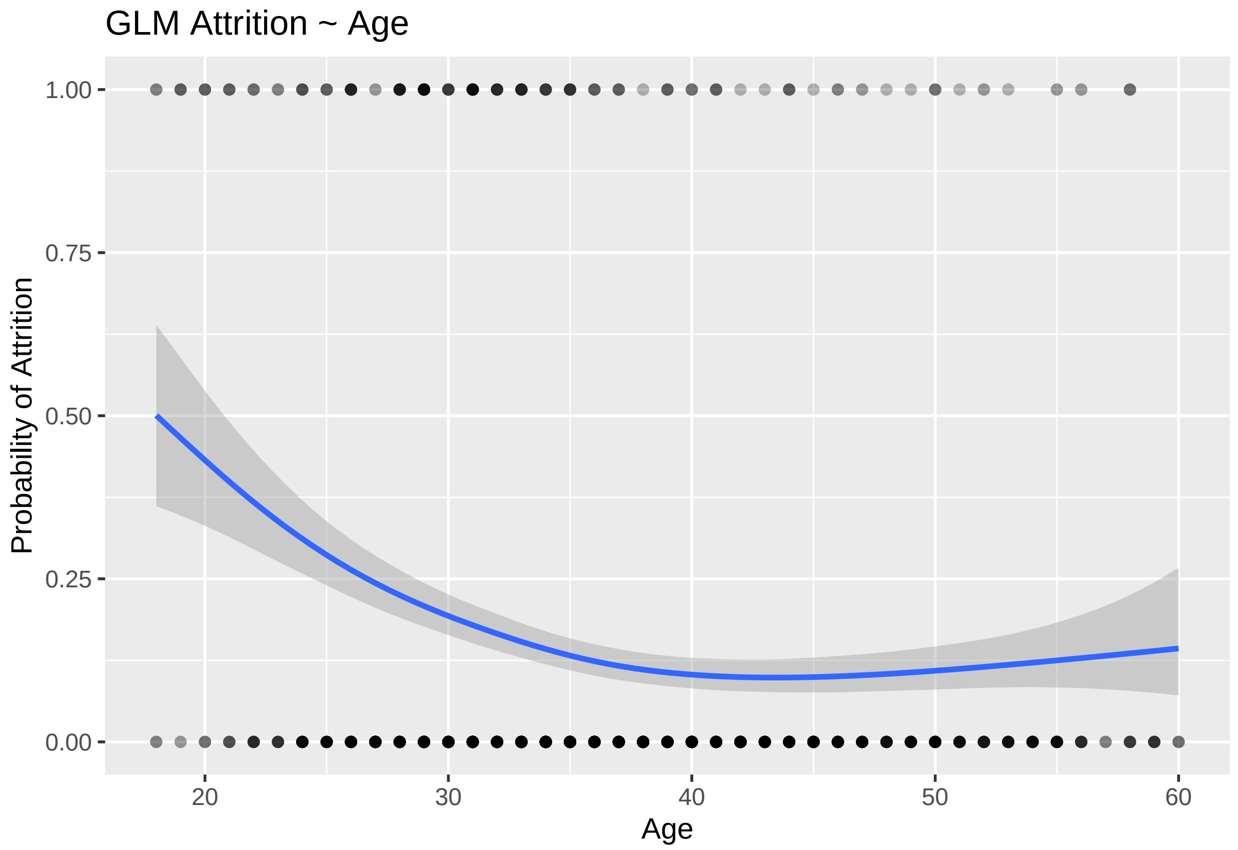
6 Extending the Linear Model

6.1 Generalised Linear Models

```
#glm.attrition <- glm(Attrition ~ Age + BusinessTravel, data = emp_attrition, family = binomial)
ggplot(emp_attrition, aes(Age, ifelse(Attrition == "Yes", 1, 0))) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "gam", method.args = list(family = "binomial")) +
  ggtitle("GLM Attrition ~ Age") +
  xlab("Age") +
  ylab("Probability of Attrition")
```

```
## Don't know how to automatically pick scale for object of type lv/lv/integer. Defaulting to continuou

## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```


GLM Attrition ~ Age
Probability of Attrition
1.00
0.75
0.50
0.25
0.00
20
30
40
50
60
Age

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
## 
## 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] mgcv_1.8-38      nlme_3.1-155     kableExtra_1.3.4 papeR_1.0-5     
##  [5] xtable_1.8-4     car_3.0-12       carData_3.0-5    ggplot2_3.3.5   
##  [9] dplyr_1.0.8      bookdown_0.24   
## 
## loaded via a namespace (and not attached):
##  [1] gtools_3.9.2      tidyselect_1.1.1  xfun_0.29         purrr_0.3.4      
##  [5] splines_4.1.2     lattice_0.20-45   colorspace_2.0-2  vctrs_0.3.8      
##  [9] generics_0.1.2    htmltools_0.5.2   viridisLite_0.4.0 yaml_2.3.4       
## [13] utf8_1.2.2        rlang_1.0.1       pillar_1.7.0      glue_1.6.1       
## [17] withr_2.4.3       DBI_1.1.2         lifecycle_1.0.1   stringr_1.4.0    
## [21] munsell_0.5.0     gtable_0.3.0      rvest_1.0.2       evaluate_0.14    
## [25] labeling_0.4.2    knitr_1.37        fastmap_1.1.0     fansi_1.0.2      
## [29] highr_0.9         scales_1.1.1      gdata_2.18.0      webshot_0.5.2    
## [33] abind_1.4-5       farver_2.1.0      systemfonts_1.0.3 digest_0.6.29    
## [37] stringi_1.7.6     gmodels_2.18.1    grid_4.1.2        cli_3.2.0        
## [41] tools_4.1.2       magrittr_2.0.2    tibble_3.1.6      crayon_1.5.0     
## [45] pkgconfig_2.0.3   Matrix_1.3-4      ellipsis_0.3.2    MASS_7.3-54      
## [49] xml2_1.3.3        assertthat_0.2.1  rmarkdown_2.11    svglite_2.1.0    
## [53] httr_1.4.2        rstudioapi_0.13   R6_2.5.1          compiler_4.1.2
```

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/