Predicting and Preventing Employee Turnover through HR Analytics

Guillermo Gallardo

2024-06-07

# 1. Introduction

## 1.1 General Background Information

The primary objective of this project is to identify the key factors contributing to employee turnover and to develop a model that can predict potential departures within a company. By anticipating these departures, we can create strategies for employee retention and enhance the overall employee experience.

I find this dataset interesting, as it aligns with my previous experience in HR Analytics. Although I’ve worked in this field before, I never had the opportunity to work on project where we could predict attrition. This project allows me to deep dive into the data I am interested about, while also enhancing my skills in modeling using R.

# 2. Dataset

For this project, I was looking for a dataset with employee data that included a wide range of variables to explore potential relationships with attrition. Below, I will provide an overview of the dataset’s origin, author, the types of variables it includes, and other relevant details.

## 2.1 **IBM HR Analytics Employee Attrition**

The dataset for this project was sourced from [Kaggle.com](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset) and was created by IBM Data Scientists as a fictional representation of HR data. It consists of 1,470 rows and 35 variables that cover demographic details, job characteristics, compensation, performance metrics, and employee satisfaction data. This data serves as the foundation for analyzing factors contributing to employee turnover and developing a predictive model.

## 2.2 **Variables**

This file includes 35 variables, detailed below, which includes a combination of character and integer types, with some of the character variables being categorical.

str(hr\_data)

'data.frame': 35 obs. of 2 variables:  
 $ Name : chr "AGE" "ATTRITION" "BUSINESS TRAVEL" "DAILY RATE" ...  
 $ Description: chr "Numerical Value" "Employee leaving the company (0=no, 1=yes)" "(1=No Travel, 2=Travel Frequently, 3=Tavel Rarely)" "Numerical Value - Salary Level" ...

## 2.3 **Gender**

Out of the 1,470 observation in the dataset, 588 are female and 882 are male. This provides a quick snapshot of the gender distribution within the data.

table(hr\_data$Gender)

< table of extent 0 >

## 2.4 **Education**

During data cleaning, we’ll use the data dictionary to decode each field. For instance, the education field is labeled from 1 to 5, representing different educational levels: 1 = “Below College,” 2 = “College,” 3 = “Bachelor’s,” 4 = “Master’s,” and 5 = “Doctorate.” Notably, about 40% of the dataset holds a Bachelor’s degree.

table(hr\_data$Education)

< table of extent 0 >

## 2.5 Data Dictionary

The data

*DELETE —– Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

# 3. Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020).

# 4. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 4.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 4.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 4.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 4.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 5. Results

## 5.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

Table 1: Data summary table.

| skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA |
| numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ |
| numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ |

## 5.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 5.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

Table 2: Linear model fit table.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 |
| Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 |
| GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 |
| GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 |

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>