Predicting and Preventing Employee Turnover through HR Analytics

Guillermo Gallardo

2024-06-07

# 1. Project Overview

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. Project Question

We have two main questions that we want to answer during this project.

* What are the primary factors that contribute to employee attrition?
* What strategies can be implement to enhance employee retention and improve the overall employee experience?

# 3. 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.

## 3.1 **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.

## 3.2 **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': 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\_Frequently" ...  
 $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
 $ Department : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...  
 $ 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" "Research Scientist" ...  
 $ 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 ...

## 3.3 **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)

Female Male   
 588 882

## 3.4 **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)

1 2 3 4 5   
170 282 572 398 48

# 4. Analysis Methods

## 4.1 Data Cleaning Process

In the data cleaning process, I will use the data dictionary to enhance data interpretation. For example, I will address issues like the Education field being labeled from 1 to 5 instead of using descriptive education level labels. I will also examine variables to detect any inaccuracies. An example of that could be to exclude the compensation fields (hourly, daily, and monthly rates) and use the Monthly Income field to ensure consistency and clarity.

## 4.2 Descriptive Analysis

I will begin my project descriptive analysis, with the goal of summarizing the data. This first step will provide insights into any outliers or patterns within the dataset. This information will also help us with creating visualization that will highlight the patterns in the data so we can focus on specific areas.

## 4.3 Correlation Analysis

As we explore various analysis methods in our class, my focus is on finding potential correlation between variables and employee attrition. I am interested in using the demographic details, job attributes and survey data to uncover pattern (if any) that highlight this issue .

## 4.4 Predictive Modeling

My main goal is to predict when an employee might leave the company. I am interested in using predictive modeling techniques (also part of my other summer class) to see if we can predict attrition. Once I’ve identified potential departures, we can use this findings to create strategies to prevent or reduce attrition.