Predicting and Preventing Employee Turnover through HR Analytics

Guillermo Gallardo

2024-06-13

# 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.

# 4. Data Cleaning Process

Data cleaning is an important step in data analysis process and focuses on detecting and correcting errors and inconsistencies in the dataset. The goals is to ensure that the data is accurate, complete and reliable. Overall, this dataset didn’t have any anomalies and there were only a couple of steps I took to prepare the data. The steps we took are the following:

**Loading Data for Cleaning Process**

In order to start the cleaning process, we need make to sure we load the data. In the code below, we use the *library(here)* function to set the file path dynamically, so our script can find and access the data files regardless of the current working directory.

hr\_data = here::here("data","raw-data","IBM HR Analytics Data.xlsx")  
rawdata = readxl::read\_excel(hr\_data)

**Removing Column**

In the next step of my data cleaning process, I removed three columns from the dataset. The columns we removed are *DailyRate*, *HourlyRate* and *MonthlyRate.* These columns were removed because the numbers they contained did not align logically or meaningfully. Upon further investigation, it became evident that there was no clear resolution to these discrepancies in Kaggle, and many suggested focusing instead on the Monthly Income field to maintain data consistency and reliability.

d1 = rawdata %>%   
 select(-DailyRate, -HourlyRate, -MonthlyRate)  
  
head(d1)

**Adding Labels**

Next, we addressed the *Education* field as the third step in our process. Initially numeric, it contained values ranging from 1 to 5. Using our data dictionary, we updated these numeric values to their corresponding labels. This adjustment enhances our understanding of employee education levels in the dataset and ensures consistency in the data.

d2 = d1 %>%   
 mutate(Education = case\_when(  
 Education == 1 ~ "Below College",  
 Education == 2 ~ "College",  
 Education == 3 ~ "Bachelor's",  
 Education == 4 ~ "Master's",  
 Education == 5 ~ "Doctorate",  
 TRUE ~ as.character(Education)  
 ))  
  
head(d2)

**Saving File**

The final step in this process involved saving the file as an RDS file. This ensures that we can use it for our exploratory data analysis and maintain an organized project that is reproducible.

HRprocesseddata = d2  
  
save\_data\_location = here::here("data","processed-data","HRprocesseddata.rds")  
saveRDS(HRprocesseddata, file = save\_data\_location)

# 5. Analysis Methods

## 5.1 Exploratory Data Analysis

I am starting my exploratory data analysis project with the objective of summarizing the dataset. I will begin by providing a concise overview of the variables included. Additionally, I will generate plots and tables to identify any outliers or patterns within the dataset. This analysis will help us understand the data available to us and guide our strategy to predict employee attrition within the organization.

### 5.1.1 Data Variables

This dataset contains 35 variables, consisting of both character and integer types, with several character variables categorized as categorical. It provides valuable insights into employees, including demographic information such as age, gender, and marital status. Additionally, it offers details about their employment, including monthly income, department, and tenure, which are crucial for understanding their professional profiles and organizational dynamics.

HR Data Variables

| V1 | V2 | V3 |
| --- | --- | --- |
| Age | Attrition | BusinessTravel |
| Department | DistanceFromHome | Education |
| EducationField | EmployeeCount | EmployeeNumber |
| EnvironmentSatisfaction | Gender | JobInvolvement |
| JobLevel | JobRole | JobSatisfaction |
| MaritalStatus | MonthlyIncome | NumCompaniesWorked |
| Over18 | OverTime | PercentSalaryHike |
| PerformanceRating | RelationshipSatisfaction | StandardHours |
| StockOptionLevel | TotalWorkingYears | TrainingTimesLastYear |
| WorkLifeBalance | YearsAtCompany | YearsInCurrentRole |
| YearsSinceLastPromotion | YearsWithCurrManager | Age |

### 5.1.2 **Gender**

The dataset consists of 1470 employees, with a gender distribution showing that males constitute the majority at 60%, while females make up 40%.

**Distribution**

Employee Gender

| Gender | Total Count | % of Total |
| --- | --- | --- |
| Female | 588 | 40 |
| Male | 882 | 60 |

### 5.1.3 **Tenure**

|  |
| --- |
| Figure 1: Employee Tenure |

### 5.1.4 **Education** MOVE TO EDA

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.

## 5.2 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 .

## 5.3 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.