Predicting and Preventing Employee Turnover through HR Analytics

Guillermo Gallardo

2024-06-21

# 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 we 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 observations 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. You can update this code to point to a specific folder path if needed but this library will make it easier to reproduce the file with the existing folders.

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

**Removing Columns**

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) # the (-) will remove these variables from the dataset  
  
head(d1) #priting the first 6 observations

**Adding Descriptive 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)

**Tenure Categorization**

To enhance the visualization of tenure distribution, I opted to introduce a new variable categorizing employee tenure. This approach allows us to identify the predominant tenure group, offering insights into potential turnover risks.

tenure\_breaks <- seq(0, 40, by = 5) #defining breaks for categories  
tenure\_labels <- paste0(tenure\_breaks[-length(tenure\_breaks)], "-", tenure\_breaks[-1], " years")  
  
d3 = d2 %>%  
 mutate(TenureCategory = cut(YearsAtCompany, breaks = tenure\_breaks, labels = tenure\_labels, include.lowest = TRUE)) #creating a new column

**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 #creating final object  
  
save\_data\_location = here::here("data","processed-data","HRprocesseddata.rds") #using the here library for the location of new file  
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. The variables shown are derived from the dataset that we cleaned in the previous step. This includes new columns, updated values, and excludes columns that we removed.

### 5.1.1 Data Variables

This dataset contains 33 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 | TenureCategory |

### 5.1.2 **Employee Tenure (in years)**

I examined the distribution of employee tenure and identified two distinct peaks: one at 5 years and another at 1 year. This provide quick insights into the typical duration of employees within the organization, and help us identify patterns that may influence employees to resign. We could concentrate on these groups to identify commonalities and understand the reasons behind their departure from the organization.

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

**Tenure category**

More than half of the employees have tenure of less than 5 years, suggesting a relatively young workforce. Shorter tenure could potentially indicate a higher turnover risk, although a comprehensive analysis incorporating other variables is necessary to fully understand the situation.

|  |
| --- |
| Figure 2: Tenure Category |

### 5.1.3 **Gender**

The dataset consists of 1470 employees, with a gender distribution showing that males constitute the majority at 60%, while females make up 40%. This also highlights the gender disparity, and it would be interesting to explore this gap by department and examine attrition rates between the two groups.

**Distribution**

Table 1: Employee Gender

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

### 5.1.4 **Monthly Income and Education**

I’m interested in examining which educational group experiences higher attrition rates. It’s evident from the data that median salaries increase with higher levels of education, but there are also numerous outliers present. This visualization serves to highlight the relationship between monthly income and educational background, which could provide insights into how these factors influence employee turnover.

|  |
| --- |
| Figure 3: Monthly Income by Education Level |

### 5.1.5 **Age vs Attrition**

In the bar chart, it is evident that attrition peaks among employees aged between 28 and 32. This trend suggests that individuals in this age group may be more likely to leave the company. Possible reasons for increased turnover in this demographic could include career advancement opportunities elsewhere, desire for higher compensation, or life changes such as starting a family.

|  |
| --- |
| Figure 4: Attrition by Employee Age |

Some of these variables may explain why employees are leaving. Understanding the employee profile by examining age, gender, department, and job satisfaction can reveal the reasons behind their departure. By creating profiles based on these insights, we can develop targeted retention strategies to address the specific needs and concerns of these employees.

## 5.2 Additional Analysis