Predicting and Preventing Employee Turnover through HR Analytics

Guillermo Gallardo

2024-07-29

# 1. Project Overview

The primary objective of this project is to identify the key factors contributing to employee turnover by developing a model that can provide the predictors that influence employees to leave the organization. By identifying this factors, 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:

## 4.1 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

## 4.2 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

## 4.3 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)

## 4.4 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

## 4.5 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 |

### 5.1.3 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 |

### 5.1.4 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 |

### 5.1.5 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 |

### 5.1.6 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 Statistical Analysis Analysis

### 5.2.1 Logistic Regression with Age

I conducted logistic regressions to explore predictors of Attrition among employees. Age emerged as a significant factor, with each additional year decreasing Attrition odds by approximately 0.052 (p < 0.001), underscoring its impact on turnover.

These findings emphasize the need for age-sensitive strategies, such as mentorship programs for knowledge transfer, flexible work arrangements, and tailored career development opportunities. Leveraging experience can foster commitment and reduce turnover among senior employees. However, limitations include the model’s inability to capture all turnover influencers, like career aspirations or economic factors.

Table 2: Logistic Regression - Age

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.2061990 | 0.3059846 | 0.6738867 | 5.00383e-01 |
| Age | -0.0522501 | 0.0087004 | -6.0055212 | 1.90718e-09 |

### 5.2.2 Logistic Regression with Income

Money plays a crucial role in life decisions, including work-related choices. I examined Monthly Income to understand its impact on Attrition among employees. The results show that higher Monthly Income is associated with lower odds of Attrition. Specifically, for each unit increase in Monthly Income, the odds of Attrition decrease by about 0.0001271 (p < 0.001), highlighting its significant influence on turnover.

To leverage these findings, organizations can consider implementing competitive salary structures, performance-based incentives, and financial wellness programs. These initiatives can enhance employee satisfaction and retention, particularly among those sensitive to income-related factors.

Table 3: Logistic Regression - Income

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -0.9291087 | 0.1292021 | -7.191125 | 6.42596e-13 |
| MonthlyIncome | -0.0001271 | 0.0000216 | -5.879336 | 4.11915e-09 |

### 5.2.3 Logistic Regresion with Age, Income, and Job Satisfaction

I used a logistic regression model to study how Age, Monthly Income, and Job Satisfaction collectively affect Attrition among employees. This approach allowed me to explore their combined impact on turnover. The analysis revealed that older employees, those with higher incomes, and greater job satisfaction are significantly less likely to leave. Specifically, each additional year of Age reduces Attrition likelihood by about 0.0328 (p < 0.001), with similar significant impacts observed for Monthly Income and Job Satisfaction.

Table 4: GLM - Age, Income, and Job Satisfaction

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.7450702 | 0.3547161 | 2.100469 | 3.56876e-02 |
| Age | -0.0328493 | 0.0096477 | -3.404877 | 6.61938e-04 |
| MonthlyIncome | -0.0000951 | 0.0000237 | -4.018468 | 5.85778e-05 |
| JobSatisfaction | -0.2628956 | 0.0650454 | -4.041727 | 5.30591e-05 |

### 5.2.4 Random Forest Model

For our last model, I choose a Random Forest due to its capability to manage intricate, nonlinear relationships among numerous variables without overfitting. This model excels in accuracy and offers valuable insights into predictor importance, helping us identify the most influential factors affecting attrition.

For this model, I did the following:

* Split the dataset into 80% training and 20% testing to evaluate model performance. This suggestion was taken from my predictive modeling class.
* We then balanced the training data to address class imbalanced which can affect the model accuracy. There were many more examples of non attrition than attrition so we needed to address them for the model.

The call below specifies the model’s formula, data, and parameters used, while noting it’s a classification Random Forest predicting Attrition **(Yes 1 /No 0)** outcomes with 1000 trees. Each tree uses 5 variables per split to enhance stability and accuracy, balancing computational resources against model variance.

Table 5: Random Forest Model

| x |
| --- |
|  |
| Call: |
| randomForest(formula = Attrition ~ ., data = train\_data\_balanced, importance = TRUE, ntree = 1000) |
| Type of random forest: classification |
| Number of trees: 1000 |
| No. of variables tried at each split: 5 |
|  |
| OOB estimate of error rate: 3.95% |
| Confusion matrix: |
| 0 1 class.error |
| 0 973 14 0.0141844 |
| 1 40 340 0.1052632 |

#### 5.2.4.1 Random Forest Model Results

Performance: The Random Forest model demonstrated excellent predictive capabilities, with an OOB error rate of just 3.95%, indicating high accuracy. The confusion matrix confirmed strong performance, particularly in identifying true positives and negatives, with low false positive and false negative rates.

**Class Error:**

* 0.0141844 for ‘0’ (No Attrition): Very low error rate, indicating high accuracy for predicting non-attrition.
  + 73 True Negatives (TN): Non-attrition cases correctly identified.
  + 14 False Positives (FP): Non-attrition cases incorrectly predicted as attrition.
* 0.1052632 for ‘1’ (Attrition): Higher error rate, suggesting room for improvement in predicting actual attrition cases.
  + 40 False Negatives (FN): Attrition cases missed by the model.
  + 340 True Positives (TP): Attrition cases correctly identified.

### 5.2.5 Random Forest Top 5 Predictors

**Top 5 Predictors:**

1. Overtime: has the highest decrease in accuracy and Gini when excluded, indicating it’s a crucial predictor of attrition. High values suggest that employees working overtime are significantly more likely to leave.
2. Monthly Income: This variable strongly influences attrition, implying that employees with certain income levels may be more prone to leaving, possibly due to pay dissatisfaction or better opportunities elsewhere.
3. Age: impacts attrition decisions, with potential reasons including generational expectations, life stage career goals, and job fit.
4. Job Satisfaction: A critical predictor indicating that lower job satisfaction is strongly associated with higher attrition.
5. Distance from Home: Longer commutes are linked to higher attrition, suggesting that travel time affects employee satisfaction and retention.

Table 6: Random Forest Top Predictors

|  | 0 | 1 | MeanDecreaseAccuracy | MeanDecreaseGini | Variable |
| --- | --- | --- | --- | --- | --- |
| OverTime | 26.776596 | 60.09822 | 59.26224 | 34.19040 | OverTime |
| MonthlyIncome | 10.471677 | 53.43746 | 52.08189 | 40.41331 | MonthlyIncome |
| Age | 11.551364 | 53.32729 | 49.67431 | 37.39627 | Age |
| JobSatisfaction | 11.522873 | 52.90910 | 48.50190 | 23.95524 | JobSatisfaction |
| DistanceFromHome | 3.583941 | 54.10550 | 47.85929 | 27.65922 | DistanceFromHome |

# 6. Strategies to Prevent Attrition

Based on insights from the models, it’s evident that factors such as overtime, monthly income, age, job satisfaction, and distance from home significantly influence employees’ decisions to leave the organization. Armed with this information, we can create strategies focus at reducing employee turnover.

Proposed strategies include:

* OverTime Management: Implement strict overtime policies and ensure workload distribution is balanced. Encouraging time-off and providing support during peak times can prevent burnout.
* Compensation and Benefits: Regular market analysis to ensure competitive pay scales and comprehensive benefits tailored to meet the diverse needs of the workforce can help in retaining talent, particularly those influenced by financial incentives.
* Age-Adaptive HR Policies: Designing age-specific policies that cater to the unique needs and preferences of different generational cohorts within the workplace. For example, younger employees might value professional development and upskilling opportunities more than their older counterparts, who might prioritize job security and healthcare benefits.
* Enhancing Job Satisfaction: Focus on creating a positive work environment where employees feel valued and have clear career paths. This can include team-building activities, transparent communication from management, and continuous professional development.
* Flexible Working Arrangements: For employees burdened by long commutes, flexible working arrangements can be a game-changer. This could range from offering telecommuting options to establishing satellite offices closer to where clusters of employees live.

# 7. Conclusion

Building on these strategies, it’s clear that addressing key factors like overtime, monthly income, age, job satisfaction, and distance from home can mitigate attrition risks. However, retaining talent involves ongoing efforts beyond initial strategies. Establishing a robust feedback system for continuous improvement, fostering a culture of open communication and trust, and regularly reviewing and adapting policies based on employee feedback can further strengthen retention efforts. Investing in leadership development programs to empower managers with skills to support and engage their teams effectively can also play a crucial role. By integrating these approaches with a commitment to fostering a supportive and inclusive workplace culture, organizations can create an environment where employees are motivated to stay and thrive, ultimately reducing turnover and fostering long-term organizational success.