Predicting Employee Attrition

Megan Tan

October 2, 2021

Contents

[Overview 1](#_Toc84693110)

[Data Description 2](#_Toc84693111)

[Data Exploration 3](#_Toc84693112)

[Age 3](#_Toc84693113)

[Travel and commute 3](#_Toc84693114)

[Percent salary increase 4](#_Toc84693115)

[Multicolinearity 4](#_Toc84693116)

[Years with company versus time since last promotion 5](#_Toc84693117)

[Data Modeling 6](#_Toc84693118)

[Data preparation procedure 6](#_Toc84693119)

[Initial load of dataset 6](#_Toc84693120)

[Convert binary and hierarchical categorical data to numeric 6](#_Toc84693121)

[Separation of numeric, categorical, and label variables 7](#_Toc84693122)

[Convert categorical values into dummy code and drop first category 7](#_Toc84693123)

[Create test and train datasets 7](#_Toc84693124)

[Multiple Linear Regression 7](#_Toc84693125)

[Multilayer Perceptron 8](#_Toc84693126)

[Gaussian Naïve Bayes 8](#_Toc84693127)

[Conclusion 9](#_Toc84693128)

[References 10](#_Toc84693129)

# Overview

Growing up playing basketball, there was always a balancing act between putting out your best players and developing your future stars. How much do you invest is training a player versus their return on investment in the following years? In college, players have a maximum of four years to join a program, learn plays and defense, integrate themselves into the team strategy, and contribute to the team. The graduation of a key player or even worse, players, could cripple a team for years as they work to rebuild.

The balancing act faced by these teams represents a small scale of what HR departments across the globe are dealing with. While players may join a program for four years, learn a binder of plays, and compete in around 30 games a year, the stakes faced by employers is much higher. Employees may have the opportunity to contribute to an employer for 40+ years. Over the course of their employment, the information contained in the trainings they attend, the procedures they implement, and experiential data they collect would easily fill numerous binders. Each year, these employees spend hundreds of days making decisions that can be beneficial, detrimental, or maintain the status quo for employers.

So how can employers grow the best teams, maintain a consistent level of performance, and increase their bottom line? One of the most important methods is through the study of employee attrition followed by strategic actions geared toward increasing the retention of valuable individuals.

This research looks at employee features that may be associated with attrition. It uses the Employee Attrition dataset from IBM to gain insight on factors that may increase an employee’s likelihood of leaving the company. It then looks at predicting employee attrition using Linear Regression, MLP Neural Network, and Naïve Bayes.

The purpose of this research is to explore possible factors that impact employee attrition to gain insight into future areas of research. Future studies should include analysis of monetary value associated with hiring and retaining employees. By coupling this information with predictive algorithms, an employer can determine ways to maximize their average ROI per employee while increasing the stability, satisfaction, and value that their employees receive from their work.

# Data Description

The *employee\_attrition\_train.csv* dataset includes multiple features on almost 1500 employees. Categories ranged from demographic information, work experience, and compensation. The predicted label was *Attrition* for which 17.1% of the employees had a value *True* (meaning they had left the company).

|  |  |
| --- | --- |
| COLUMN NAME | DESCRIPTION |
| AGE | Numerical Value |
| ATTRITION | Employee leaving the company (0=no, 1=yes) |
| BUSINESS TRAVEL | (1=No Travel, 2=Travel Frequently, 3=Travel Rarely) |
| DAILY RATE | Numerical Value - Salary Level |
| DEPARTMENT | (1=HR, 2=R&D, 3=Sales) |
| DISTANCE FROM HOME | Numerical Value - THE DISTANCE FROM WORK TO HOME |
| EDUCATION | Numerical Value. (1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor') |
| EDUCATION FIELD | (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 8= TECHNICAL) |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Numerical Value - EMPLOYEE ID |
| ENVIRONMENT SATISFACTION | Numerical Value - SATISFACTION WITH THE ENVIRONMENT (1 'Low' 2 'Medium' 3 'High' 4 'Very High') |
| GENDER | (1=FEMALE, 2=MALE) |
| HOURLY RATE | Numerical Value - HOURLY SALARY |
| JOB INVOLVEMENT | Numerical Value - JOB INVOLVEMENT (1 'Low' 2 'Medium' 3 'High' 4 'Very High') |
| JOB LEVEL | Numerical Value - LEVEL OF JOB |
| JOB ROLE | (1=HR REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 8= RESEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIVE, 9= SALES REPRESENTATIVE) |
| JOB SATISFACTION | Numerical Value - SATISFACTION WITH THE JOB (1 'Low' 2 'Medium' 3 'High' 4 'Very High') |
| MARITAL STATUS | (1=DIVORCED, 2=MARRIED, 3=SINGLE) |
| MONTHLY INCOME | Numerical Value - MONTHLY SALARY |
| MONTHLY RATE | Numerical Value - MONTHLY RATE |
| NUMCOMPANIES WORKED | Numerical Value - NO. OF COMPANIES WORKED AT |
| OVER 18 | (1=YES, 2=NO) |
| OVERTIME | (1=NO, 2=YES) |
| PERCENT SALARY HIKE | Numerical Value - PERCENTAGE INCREASE IN SALARY |
| PERFORMANCE RATING | Numerical Value - PERFORMANCE RATING |
| RELATIONS SATISFACTION | Numerical Value - RELATIONS SATISFACTION |
| STANDARD HOURS | Numerical Value - STANDARD HOURS |
| STOCK OPTIONS LEVEL | Numerical Value - STOCK OPTIONS (Higher the number, the more stock option an employee has) |
| TOTAL WORKING YEARS | Numerical Value - TOTAL YEARS WORKED |
| TRAINING TIMES LAST YEAR | Numerical Value - HOURS SPENT TRAINING |
| WORK LIFE BALANCE | Numerical Value - TIME SPENT BETWEEN WORK AND OUTSIDE |
| YEARS AT COMPANY | Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPANY |
| YEARS IN CURRENT ROLE | Numerical Value -YEARS IN CURRENT ROLE |
| YEARS SINCE LAST PROMOTION | Numerical Value - LAST PROMOTION |
| YEARS WITH CURRENT MANAGER | Numerical Value - YEARS SPENT WITH CURRENT MANAGER |

# Data Exploration

Initial visualizations of the data were performed using Power BI. This process provides a look into possible features that may be related to employee attrition and a launching pad for future explorations on this topic.

## Age

The comparison of employee age versus rate of attrition strongly suggests that younger employees have a higher rate of attrition. However, the relationship is not linear as the rate of attrition begins to increase starting at age 50. This may be related to early retirement. The 60+ group only had 5 participants. More data should be collected for this age group to determine if this increase in attrition rate continues.

A possible direction for further research would be looking into similar features that may present in age groups and how it relates to their probability for leaving the company. Perhaps there is high correlation between employees leaving the company that are between the age of 25-29, with bachelor’s or master’s degrees, and have not received a promotion or significant wage increase in the past two years? By determining these types of correlations, a company can design policies targeted to maintaining these individuals.

Chart, bar chart

Description automatically generatedIt may be that many young professionals find themselves leaving an academic world, where every couple of years they are celebrated for achieving significant milestones, only to find themselves in a world where the timelines to future milestones are nebulous at best. The best and most ambitious of these employees may find waiting on such uncertainties as senior employees retiring unacceptable and take fate into their own hands as they seek out new and better opportunities.

Figure Shows the percentage of employees in a particular age group that leave the company.

## Travel and commute

Both business travel and daily commutes are strongly linked to employee attrition. Almost 25% of Employees who travel frequently leave the company while employees that live more than 11 miles from work are 1.53 times more likely to leave the company.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedArmed with this information, the employer could perform periodic checks with employees who are traveling frequently to gauge their satisfaction and provide options to travel less. To combat long commutes, employers can investigate the effects of options such as flexible work weeks, hybrid, or remote work.

Figure Shows the percentage of employees that leave the company compared to their frequency of travel.

Figure Shows the percentage of attrition for employees whose commute is less than or equal to 11 miles and those with greater commutes.

## Percent salary increase

Chart, bar chart

Description automatically generatedIntuitively, one would predict that higher percent salary increases would result in lower attrition rates. However, there does not appear to be a trend between percentage in salary hike and attrition rate. Possible reason may be that employees are receiving higher salary increases because they are at risk of leaving. Another possibility is that employees that received a high increase have been undervalued over the course of their employment and the recent increase is still not enough to convince them into continuing their employment.

Figure Shows the percentage of employees that leave the company based on the percentage of their last salary increase.

Understanding the necessary pay requirements to keep valuable employees as well as the cost of hiring and training new employees is an integral part of growing a productive team. When and to what level pay increases are given can influence employees staying in the company.

## 

## Multicollinearity

The dataset contains many correlated variables such as total years working, years with the company, and years since last promotion. To avoid multicollinearity, some of these variables were eliminated during the modeling process.

Chart, scatter chart

Description automatically generated

Figure Shows a correlation matrix for numeric variables.

## Years with company versus time since last promotion

Chart, scatter chart

Description automatically generatedIt appears that most employees leave the company within the first 10 years of employment. The figure suggests the ratio of time since an employee’s last promotion compared to their years at the company increases the likelihood of attrition.

Further research could look at how different employee features such as salary, age, and education relate to an employee’s likelihood to leave the company versus the time with the company and since their last promotion. Employees with higher levels of education, pay expectations, or ambition may require more frequent promotions to increase their rates of retention.

Figure Show Years at company plotted against years since the employee's last promotion. The size of the circle represents the number of employees in that category.

# Data Modeling

Data modeling was performed using python is VSCode.

## Data preparation procedure

### Initial load of dataset

A picture containing table

Description automatically generated

The data did not have any null or infinite values.

Columns were removed that contained index information, similar variables, or showed only one value.

* EmployeeNumber (Index)
* EmployeeCount (value = 1)
* StandardHours (Value = 80)
* Over18 (Value = yes)
* DailyRate (leave MonthlyIncome)
* HourlyRate (leave MonthlyIncome)
* MonthlyRate (leave MonthlyIncome)

### Convert binary and hierarchical categorical data to numeric

Text

Description automatically generated

### Separation of numeric, categorical, and label variables

Text

Description automatically generated

### Convert categorical values into dummy code and drop first category

Text

Description automatically generated

### Create test and train datasets



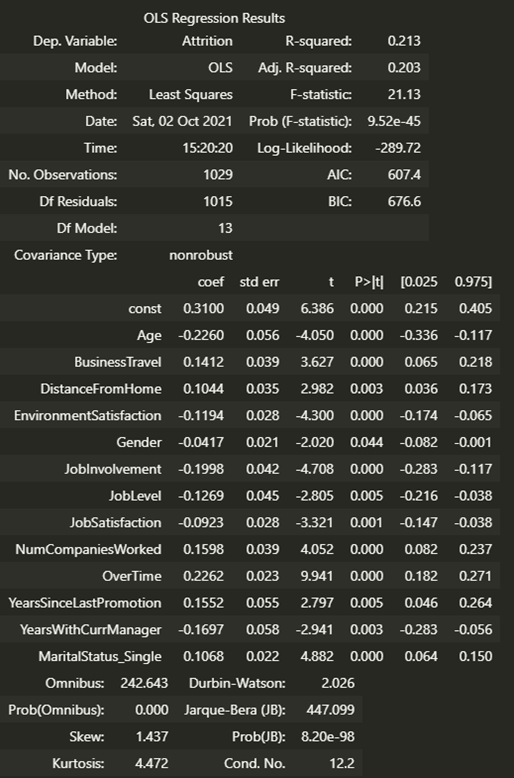
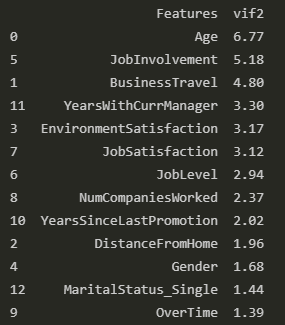
### Multiple Linear Regression

The initial model used ordinary least squares to look at multicollinearity. The correlation chart shows that many variables are highly related. Some of these will have to be removed if the variance inflation factor is over 7.0. Additionally, variables with p-values above 0.05 (suggesting randomness in their predictive powers) will be removed.

The initial linear regression shows multiple variables with VIF and p-values above the cutoff. These variables were pruned over multiple rounds modeling until a final feature set was determined.

The final linear regression model showed the variables that contributed most to employee attrition were:

* Having overtime
* Younger age
* Lesser job involvement
* Less time with current manager
* Higher number of companies worked at



### Multilayer Perceptron

Using the features determined from the linear regression model, attrition was predicted using a neural network. After multiple tests, the highest prediction accuracy was 83.45%.

Text

Description automatically generated

### Gaussian Naïve Bayes

The final model tested was Naïve Bayes. This prediction method had the highest accuracy at 84.58%.

Text

Description automatically generated

Chart

Description automatically generated

# Conclusion

This research identifies multiple avenues for continued research such as:

* Will further classification of groups of employees reveal unique features that increase the likelihood of employee attrition within an age category?
* Since travel and commute play a significant role in employee retention, can work from home options and programs that gradually transition employees to roles requiring less travel maintain productivity and reduce employee attrition?
* How can salary increases be used more effectively to increase employee retention?

In terms of modeling, linear regression revealed that some of the features with highest impact on employee attrition were overtime, age, job involvement, years with the current manager, and number of companies worked for. While the relationship of features does not appear to be linear, both the multilayer perceptron on naïve bayes were able to predict attrition to a high degree of accuracy (83.45% and 84.58% respectively).

If an employer can determine employees that are more likely to leave the company, they can use this information in powerful ways. From a more pessimistic standpoint, they look at ways to mitigate the effects of employee that leave. Perhaps there is a department where a significant number of employees have a high probability of leaving the company. Like a sports team, it may be advantageous to increase focus on training new employees and making sure they have experience working on higher difficulty tasks commonly assigned to other individuals.

From an optimistic view, employers can harness this research to create workspaces where employees are encouraged to stay. Like college athletes, new employees require additional time and training before they contribute substantially to a company. Unlike college athletes, employees are not limited to four years of contribution to a company. By targeted, data-driven policies, an employer can increase the retention of valuable employees.

This type of research is not new. For years, individuals and companies have research how to retain the best employees. The difference is now we have the technology to do something about it. We no longer require years of data collection and then slow statistical calculations to determine what factors were affecting employee attrition five years ago. We have forms and surveys that can reach thousands of employees, be processed automatically, and trends displayed to a dashboard on your mobile device.

The opportunity is there, the value is real, and the time is now. Each year, more and more companies are harnessing the power of analytics to maximize their workforce’s potential. Each year they retain more star players, increase in overall skill, while cutting cost on hiring and training. This is the future, and the benefits are being realized by both employer and employee. For companies not in this group, the question is, “why not?”

# References

*Employee Attrition*. (2020, August 6). Retrieved from kaggle.com: https://www.kaggle.com/colearninglounge/employee-attrition

*Watson Analytics Use Case for HR Retaining Valuable Employees*. (2021, October 1). Retrieved from ibm.com: https://www.ibm.com/communities/analytics/watson-analytics-blog/watson-analytics-use-case-for-hr-retaining-valuable-employees