IST 707 Data Mining

Team Project

Bomin Powers

Karen Kim

**HR attrition prediction**

Table of Contents

[Introduction 2](#_Toc90931580)

[About the data 2](#_Toc90931581)

[Preprocessing of the Dataset 3](#_Toc90931582)

[Descriptive Analysis 3](#_Toc90931583)

[Analysis and Modeling 8](#_Toc90931584)

[1. K-means Clustering 8](#_Toc90931585)

[1-1. The elbow method 8](#_Toc90931586)

[1-2. The Silhouette method 8](#_Toc90931587)

[1-3. Generate the cluster model 9](#_Toc90931588)

[1-4. Prediction results 9](#_Toc90931589)

[2. Decision Tree Algorithms 10](#_Toc90931590)

[2-1. Generate decision tree models 10](#_Toc90931591)

[2-2. Prediction results 11](#_Toc90931592)

[2-3. Naïve Bayes Model Prediction 12](#_Toc90931593)

[3. Association Rule Mining 12](#_Toc90931594)

[3-1. Data preparation for the model 12](#_Toc90931595)

[3-2. Generate the Association rules with rhs = “Attrition=No”. 13](#_Toc90931596)

[Results 16](#_Toc90931597)

[Conclusions 17](#_Toc90931598)

# Introduction

The year 2020 has been known as 'The great Resignation'. The past April, the number of workers who quit their job in a single month broke an all-time U.S. record. As the months went by, there was an increasing number of people quit their jobs. Unlike in the past, the younger generation of workers constantly longs for their own growth potential, happiness, and value in what they do. According to Indeed research (Feb 25, 2021), common reasons employees leave their job are include: 1. needing more of a challenge, 2. looking for a higher salary, 3. feeling uninspired, 4. wanting to feel valued, and so on. If there is a way to find out why employees leave their current job in advance, companies can decrease its employee’s attrition rate significantly.

It is important to know why employees leave their jobs because a high turnover rate can lead to lower job satisfaction. Hiring new employees to fill this gap takes time and effort, which is why it is more efficient to find a way to keep the attrition rate low and keep employees happy. After finding out why employees are leaving, employers can directly solve problems that recur and create a more pleasant working environment for everyone.

It would be helpful if there is a way to predict which employees are likely to quit in the future. More efficient HR tasks will be possible if employees who are likely to be attrite are selected and take appropriate actions on those who are likely to leave. Here, the actual data set regarding employee’s attrition is presented, including age, employment date, employment source, and monthly income, etc. With this data set, a few of the data mining analysis methods will be used to predict whether the employee will leave or not.

## About the data

The origin data set was acquired from the Kaggle, [Kaggle - HR Attrition Dataset](https://www.kaggle.com/singhnproud77/hr-attrition-dataset). The dataset was created based on the IBM attrition dataset to predict employee attrition. The data set includes 1,470 observations across 32 variables. The description of the variables is the following:

1. "Age" = The age of the employee
2. "Attrition" = Whether the employee has attired or not
3. "BusinessTravel" = Whether the employee used to travel for business or not
4. "Department" = Which department the employee was employed under
5. "DistanceFromHome" = The distance the employee travels to reach for job on a day-to-day basis
6. "Gender" = Gender of the employee
7. "JobInvolvement" = The involvement rating of an employee over the job handled
8. "JobLevel" = Level at which the employee is working
9. "JobRole" = The roles and responsibilities of the employee
10. "JobSatisfaction" = Satisfaction rating of the employee for the job
11. "MaritalStatus" = Marital status of the employee
12. "MonthlyIncome" = Monthly income of the employees
13. "NumCompaniesWorked" = Number of companies the employees have worked for
14. "OverTime" = Whether working Overtime or not
15. "PercentSalaryHike" = Percentage salary hike since their appointment in the company
16. "PerformanceRating" = Performance rating
17. "StockOptionLevel" = Level of opted for sharing the stock
18. "TotalWorkingYears" = Total years worked by the employees
19. "TrainingTimesLastYear" = How many trainings the employee has undergone
20. "YearsAtCompany" = Years spent at the present organization
21. "YearsSinceLastPromotion" = Time gone in years since last promotion
22. "YearsWithCurrManager" = Years working under current manager
23. "Higher\_Education" = Higher education level of the employee
24. "Date\_of\_Hire" = Date of hire of the employee in the current organization
25. "Date\_of\_termination" = Date of termination from the organization
26. "Status\_of\_leaving" = Reason for leaving the organization
27. "Mode\_of\_work" = WFH or WFO
28. "Leaves" = Total permitted leaves taken by the employee
29. "Absenteeism" = Total days absent for the employee
30. "Work\_accident" = Work accident if any
31. "Source\_of\_hire" = Source of hire
32. "Job\_Mode" = Working full time/ part or contractual

## Preprocessing of the Dataset

In order to get to know better of the data set, the following steps were performed:

* Removed Date\_of\_termination because of NA values
* Convert character data type to nominal types to conduct correlation analysis
* For variables with inconsistent distribution of values, they were distributed so that similar numbers were included per group (e.g., in the age group, the most distributed was between 29 and 36 years old among 18 and 60 years old. Thus, the first group is between 18 and 28 years old and, the second group is between 29 and 32 years old)

The additional data preparation process is applied to meet the criteria required by each model.

# Descriptive Analysis

This dataset contains 32 variables, so you need to make sure which variables are more meaningful in predicting attrition than others. Through the correlation analysis, let's see which variables most correlated to the Attrition attribute.

hr\_cor <- cor(hrdata[ ,colnames(hrdata) != "Attrition"], hrdata$Attrition)  
hr\_cor

## [,1]  
## BusinessTravel 7.377695e-05  
## Department 6.399060e-02  
## Gender 2.945325e-02  
## JobInvolvement -1.300160e-01  
## JobLevel -1.691048e-01  
## JobRole 6.715150e-02  
## JobSatisfaction -1.034811e-01  
## MaritalStatus 1.620702e-01  
## NumCompaniesWorked 4.349374e-02  
## OverTime 2.461180e-01 # most correlated  
## PerformanceRating 2.888752e-03  
## StockOptionLevel -1.371449e-01  
## TrainingTimesLastYear -5.947780e-02  
## Higher\_Education 3.641604e-03  
## Status\_of\_leaving 2.075012e-02  
## Mode\_of\_work 6.741883e-03  
## Leaves -4.181966e-02  
## Absenteeism -3.786683e-02  
## Work\_accident 9.845918e-03  
## Source\_of\_Hire 4.462131e-03  
## Job\_mode -5.566298e-02  
## ageG -1.579734e-01  
## DistanceFromHomeG 7.894931e-02  
## MonthlyIncomeG -1.965973e-01  
## PercentSalaryHikeG -2.062394e-02  
## TotalWorkingYearsG -1.829648e-01  
## YearsAtCompanyG -1.641019e-01  
## YearsSinceLastPromotionG -5.334780e-02  
## YearsWithCurrManagerG -1.624467e-01  
## Month\_of\_Hire -1.184360e-02  
## Year\_of\_Hire\_G 1.871048e-01

The correlation of each attribute found is quite small. The most correlated attribute is OverTime with a 0.25 correlation. There are 8 more attributes that had correlation between 0.16 and 0.2 in absolute value: JobLevel(-0.17), MaritalStatus (0.16), AgeGroup(-0.16), MonthlyIncomeGroup(-0.20), TotalWorkingYearsGroup(-0.18), YearsAtCompanyGroup(-0.16), YearsWithCurrManagerGroup(-0.16), and Year\_of\_HireGroup(0.19). Because OverTime had the largest correlation coefficient, this variable will be investigated. It is a binary variable which is yes and no for Overtime work had existed. (Figure 1) It seems it is biased toward one value (overtime=no). This means Over time variable could make a generalized prediction

Chart, histogram

Description automatically generated

Figure 1. Histogram of Overtime

The variable, MonthlyIncome shall be examined next. The original variable has a non-evenly distributed pattern (Figure 2). It seems heavily distributed around $5,000. After splitting those values into 6 equal-sized groups, the details of each group are as follows: Group1 (1009: 2561), Group2 (2561: 3633), Group3 (3633: 4930), Group4 (4930: 6538), Group5 (6538: 10552), Group6 (10552: 19999). Looking at the correlation amongst the distributed groups of 1 to 6 and their attrition (Figure 3), it was observed that the attrition decreases from Group 1 to 6, that is, the higher the monthly income, the closer the attrition rate is to 0.

Chart, histogram

Description automatically generated

Figure 2. Histogram of Monthly Income

Line chart

Description automatically generated with low confidence

Figure 3. Correlation scatterplot between Attrition and Monthly Income Group variables

The next variable is Year\_of\_Hire. The original variable has a non-evenly distributed pattern (Figure 4).

Chart, histogram

Description automatically generated

Figure 4. Histogram of Year\_of\_Hire

Same as monthly income variable, equal-sized group dividing process had done as following codes.

# split Year\_of\_Hire variable into somewhat equal-sized groups  
hrdata$Year\_of\_Hire\_G <- as.factor(cut2(hrdata$Year\_of\_Hire, g=5))  
# check how each group range look like  
summary(hrdata$Year\_of\_Hire\_G)

## [1969,2012) [2012,2015) [2015,2017) [2017,2020) [2020,2021]   
## 366 252 272 365 215

Chart, line chart

Description automatically generated

Figure 5. Correlation scatterplot between Attrition and Year of Hire Group variables

As can see above scatterplot, there is a significant correlation between Attrition and Year of Hire Group variables *(Figure 5).* The group 1 to 5, which means, the closer it is to Attrition=1 as it is recently employed. Graphical user interface, chart, line chart

Description automatically generated

Figure 6. Correlation scatterplot between Attrition and 4 other variables

For the remaining variables that showed relatively high correlation, a scatterplot with Attrition was formed. (Figure 6) Individual correlations are small however combination of these attributes would make a good prediction model for Attrition.

# Analysis and Modeling

Initially, the correlation between each variable and attrition was examined. Several variables showed relatively high correlation with Attrition. From here, it will be helpful to examine if the variables observed here will also help build an actual predictive model. First, using K-means Clustering (unsupervised learning), employees will be grouped in a few segments that share similar attributes (i.e., age, job level, distance from work, years at the company etc.), and this will be used to predict the attrition rate of the employees. Then, using Decision Tree Algorithms (supervised learning), a portion of the data set will be trained, and based on the trained data set to predict the attrition rate of the remaining data set. Lastly, using Association Rule Mining top 10 rules that have high support and lift measures will be examined to find out which attributes most likely will increase employees’ attrition rate.

## K-means Clustering

### 1-1. The elbow method

In order to come up with an optimal number of clusters, we used elbow method for cluster analysis. The kink seems to occur at k =3.

Chart, line chart

Description automatically generated

Figure 7. Elbow method for optimal number of clusters

### 1-2. The Silhouette method

Silhouette analysis was used to determine the number of clusters by examining how close each pint in one cluster is to points in the neighboring clusters to find a way to assess data points.

Chart, line chart

Description automatically generated

Figure 8. Silhouette method for optimal number of clusters

### 1-3. Generate the cluster model

Now, generate the cluster model based on the suggestion from Silhouette and Elbow method. (Figure 9). The clusters are clearly divided at k=2 and k-3.

Chart, scatter chart

Description automatically generated

Figure 9. Cluster models

### 1-4. Prediction results

From the cluster analysis results, it would be helpful to understand if there is a relationship between Attrition variable and the cluster group. Since there is a huge difference between Attrition=1 and 0, sampled 50 from each, confusion matrix will be generated to see if there is a relationship.

# summary of Attrition variable  
summary(as.factor(hrdata$Attrition)  
## 0 1   
## 1233 237

First, the 2 clusters result is as follows. Certainly, there is no difference in attrition between the two clusters. Even though the sample was repeated in the same way, the results were similar.

# confusion matrix to find the connection with attrition variable  
table(hrdata\_km2sam$Attrition,hrdata\_km2sam$hrdata\_km2.cluster)

##   
## 1 2  
## 0 24 26  
## 1 39 11

As the confusion matrix shown below, like clusters k=2, there is no difference in attrition between cluster groups. In addition, even though the sample was repeated in the same way, the results were the same.

# confusion matrix to find the connection with attrition variable  
table(hrdata\_km3sam$Attrition,hrdata\_km3sam$hrdata\_km3.cluster)

##   
## 1 2 3  
## 0 15 20 15  
## 1 16 24 10

## Decision Tree Algorithms

### 2-1. Generate decision tree models

For the correlation analysis, all variables were converted to a numeric type. Thus, no more data preparation is needed at this time. Let’s generate the first decision tree model.

Diagram

Description automatically generated

Figure 10. FancyPlot of decision tree

Decision Tree Model indicates that OverTime has an impact on the monthly income and year of hire. 85% of employees work overtime, and 91% of those employees’ year of hire was less than 4.5 years. 15% of the entire employee group do not work overtime, and 70% of those not working overtime makes greater than 1,500 a month.

Surprisingly, these key attributes (OverTime, Monthly Income, and Year of Hire) on the decision tree model, are the most correlated variables when we conducted correlation analysis.

### 2-2. Prediction results

From the decision tree model above, let’s implement the prediction. Prior to prediction, design the experiment which randomly select a train and test set for validation. Set the ratio of train / test is .6 Below is the result of the prediction. The accuracy of the model is 478/588 = ~.813. [Refer to Accuracy = TP /(TP+FP). TP is the number of true positives, and FP is the number of false positives.]

# confusion matrix and statistics results  
confusionMatrix(data = predicted1, as.factor(test$Attrition))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 449 75  
## 1 35 29  
##   
## Accuracy : 0.8129   
## 95% CI : (0.779, 0.8437)  
## No Information Rate : 0.8231   
## P-Value [Acc > NIR] : 0.7606539

### 2-3. Naïve Bayes Model Prediction

To further validate the classification model, Naïve Bayes model was conducted on the training and testing model. This provides a way to validate the assumption of independence of the variables, which indicates that the presence of any attributes in a class is unrelated to the presence of any other attributes. In this project, 60% of the data set was trained and predict the rest of the data set.

summary(train\_naibayes)

## Length Class Mode   
## apriori 2 table numeric   
## tables 31 -none- list   
## levels 2 -none- character  
## isnumeric 31 -none- logical   
## call 4 -none- call

# confusion matrix to find correct and incorrect predictions  
table(Attrition=nb\_Pred, true=test$Attrition)

## true  
## Attrition 0 1  
## 0 407 43  
## 1 77 61

Using the formula stated earlier, Accuracy = TP /(TP+FP), [TP is the number of true positives, and FP is the number of false positives] the accuracy rate turns out to be 79%. (468/588 = ~.7959)

## Association Rule Mining

### 3-1. Data preparation for the model

Because of the uneven distribution as indicated below (Figure 11), it was hard to measure how much of impact that Performance Rating and Over Time have on employee’s attrition rate. The data points were largely skewed to the left; to gain fair findings of the result, the two variables were normalized by random sampling.

Graphical user interface

Description automatically generated

Figure 11. Histogram of PerformanceRating and OverTime variables

### 3-2. Generate the Association rules with rhs = “Attrition=No”.

The first Association Rule Mining was run on parameters support level = 0.07, and have consequent, rhs = “Attrition=No”. Association Rules are created by searching data for frequently produced items, patterns, using criteria support, confidence, and lift. The support measure was first chosen because it is an indication of how frequently the items appear in the data set. Confidence indicates the likeness of occurrence of the relationships. For example, the rule found for those who stayed, who had a salary increase. The confidence level for this was 1, which indicates there is a high likelihood this relationship will hold. Technically, confidence is the conditional probability of occurrence of consequent item (rhs) given the antecedent (lhs).

Confidence ({X} ® {Y}) =

Lift is the ratio of confidence to support. A positive lift value indicates there is a positive correlation. Two different algorithms will be run to see if they bring out different frequent item sets. Also, having “Attrition=No” as our threshold delivered many more results to examine than having “Attrition=Yes”. This is because there are more data of employees who stayed at the company (i.e., “Attrition = No”).

The factors that cause employees to stay at the company based on the support measure were (Figure 12. Table of 10 Association rules sorted by support decreasing): Mode of Work = Office, Performance Rating = 3, OverTime = No, Work Accident = Yes, Job Involvement = 3, and Stock Option Level = 1. These are counter-intuitive to what companies would expect from employees. Findings show that employees leave if they had to work in office and had a work injury. To prove this finding, different parameters should be implemented.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | LHS | RHS | support | Lift | count |
| 1 | OverTime=No, Mode\_of\_work=OFFICE | Attrition=No | 0.25 | 1.26 | 25 |
| 2 | PerformanceRating=3, Mode\_of\_work=OFFICE | Attrition=No | 0.25 | 1.26 | 25 |
| 3 | OverTime=No, PerformanceRating=3, Mode\_of\_work=OFFICE | Attrition=No | 0.25 | 1.26 | 25 |
| 4 | JobInvolvement=3, OverTime=No, Work\_accident=Yes | Attrition=No | 0.23 | 1.26 | 23 |
| 5 | JobInvolvement=3, PerformanceRating=3, Work\_accident=Yes | Attrition=No | 0.23 | 1.26 | 23 |
| 6 | JobInvolvement=3, OverTime=No, PerformanceRating=3, Work\_accident=Yes | Attrition=No | 0.23 | 1.26 | 23 |
| 7 | Gender=Female, StockOptionLevel=1 | Attrition=No | 0.22 | 1.26 | 22 |
| 8 | JobInvolvement=3, OverTime=No, Mode\_of\_work=OFFICE | Attrition=No | 0.22 | 1.26 | 22 |
| 9 | JobInvolvement=3, PerformanceRating=3, Mode\_of\_work=OFFICE | Attrition=No | 0.22 | 1.26 | 22 |
| 10 | JobInvolvement=3, OverTime=No, PerformanceRating=3, Mode\_of\_work=OFFICE | Attrition=No | 0.22 | 1.26 | 22 |

Figure 12. Table of 10 Association rules sorted by support decreasing

Chart

Description automatically generated with low confidence

Figure 13. Plot of the best 20 Association Rules sort by support

Secondly, to further validate the findings of employee’s main reason to stay at the company, another Association Rules will be run targeting those who stay. But this time, the algorithms will be run based on lift - the ratio of confidence to support. (Figure 14. Table of 10 Association rules sorted by lift decreasing)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | LHS | RHS | support | Lift | count |
| 1 | Year\_of\_Hire\_G=[2015,2017) | Attrition=No | 0.08 | 1.26 | 8 |
| 2 | ageG=[47,60] | Attrition=No | 0.10 | 1.26 | 10 |
| 3 | YearsWithCurrManager=[3, 5) | Attrition=No | 0.10 | 1.26 | 10 |
| 4 | JobRole=Manufacturing Director | Attrition=No | 0.11 | 1.26 | 11 |
| 5 | MonthlyIncomeG=[10552,19999] | Attrition=No | 0.13 | 1.26 | 13 |
| 6 | PercentSalaryHike=14 | Attrition=No | 0.14 | 1.26 | 14 |
| 7 | JobInvolvement=3,  Year\_of\_Hire\_G=[2015,2017) | Attrition=No | 0.08 | 1.26 | 8 |
| 8 | Department=Research & Development, ageG=[47,60] | Attrition=No | 0.08 | 1.26 | 8 |
| 9 | BusinessTravel=Travel\_Rarely, ageG=[47,60] | Attrition=No | 0.08 | 1.26 | 8 |
| 10 | YearsAtCompany=[ 4, 6),  YearsWithCurrManager=[3, 5) | Attrition=No | 0.09 | 1.26 | 9 |

Figure 14. Table of 10 Association rules sorted by lift decreasing

When the rules were sorted by lift, different factors turned out to motivate employees to stay at the company:

* Years at company is between 4 and 6 years
* Years with current manager is between 3 to 5
* Monthly Income between $10,552 and $19,999
* Percent Salary Hike is 14%
* Hired between 2015 and 2017
* Age between 47 and 60
* Job Role = Manufacturing Director
* Department = Research & Development

The lift measures show how frequently consequent (in this case, “Attrition=No”) shows up accounting both antecedent and consequent. Support is a measure to tell how much an item appears [i.e., frequency (A, B)/Total]. These are sensible results; for instance, those who are in a higher income group, tend to be happier at the company. From [Figure *2*. Histogram of Monthly Income](#Figure2), those who have monthly income between $10,552 and $19,999 are in the highest income group.

Chart

Description automatically generated with low confidence

Figure 15. Plot of the best 20 Association Rules sort by lift

# Results

Upon running K-means Cluster Analysis, Decision Tree Model, and Association Rule Mining, there were a few factors that all these models shared in common: OverTime, Stock Option Level, Monthly Income, and Year of Hire. K-means clustering was conducted to find out there were 3 clusters that share similarities. This analysis was illustrated by the Elbow Plot and Silhouette method. To predict attributes that motivates employees to stay at current company, Decision Tree Algorithms was implemented. From Decision Tree Model, Over Time, Monthly Income, Stock Option Level, and Year of Hire were attributes that have largest impact on whether employees stays or not. This is significant because Over Time, Monthly Income, and Year of Hire were found to be highly correlated with employee’s decision to stay at current company from the correlation analysis. The accuracy of Decision Tree Model was 81%. As another means to validate classification models of supervised learning, Naïve Bayes Model was conducted to validate the assumption of independence of the variables. The accuracy of the prediction model turned out to be 84%. Using Association Rule Mining, correlation among different variables was assessed. Different attributes appeared to be meaningful when the rules were analyzed by different parameters. However, the findings that Monthly Income, Year of Hire, Over Time, and Stock Option Level have high correlation with employee’s decision to stay at current company still held valid as discovered in other models.

# Conclusions

A few variables in employee’s attrition were found to be significant after conducting Correlation Analysis, Cluster Analysis, Decision Tree Model, and Association Rule Mining. The common factors that all these models found were:

* Whether an employee works overtime
* Employee’s Monthly Income
* Employee’s Stock Option Level
* Employee’s Year of Hire

These attributes were statistically significant as different machine learning models produced in common. Besides the above-listed attributes, Years of Hire is between 2015 and 2017, Years at Company is 4 to 6, Marital Status, and Total Working Years were shown to be highly correlated to employees’ attrition rate. Even though the company cannot influence variables such as Year of Hire, Marital Status, the company can attract and motivate employees by enhancing the benefit packages such as giving the employees wedding anniversary gifts or sending them on a vacation on their 2-year anniversary of working at the company. Other ways of improving working conditions are enhancing Stock Option Level, Monthly Income, and preventing over time work. These factors could be dependent on the profits of the company, but only when the company maintains a lower attrition rate, in other words, if the company finds a way to encourage good and hard-working employees to stay longer at the company, they can potentially increase their future profits. Also, this model should be assessed periodically because there are many outward factors (i.e. macro-economic condition, pandemic, etc.) that can affect the company’s financial health and its employee attrition rate.