

Individual Assignment

BY jurre foekens

Jurre Foekens | MDD Minor 3DMiB| 13-01-2023

Inhoud

[1 – Business Understanding 3](#_Toc124530394)

[2 – Data Understanding 4](#_Toc124530395)

[3 – Data Preparation 5](#_Toc124530396)

[4 – Modeling 6](#_Toc124530397)

[5 – Evaluation and Deployment 7](#_Toc124530398)

# 1 – Business Understanding

I found this dataset on: <https://www.kaggle.com/datasets/prachi13/employeeattritionrate>. It is about the employee dropout rate. The dataset contains information about employees at a company and whether they have left the company. This information can be used by a business to identify and address factors that affect employee retention and predict which employees are likely to leave. By understanding and addressing these factors, a business can keep its workers and make them more efficient, resulting in more money.

# 2 – Data Understanding

Below, I will show the data and then explain it:

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 1 - Dataset Attrition

The dataset has 13 variables/columns and 1470 oberservations/rows. Each row in dataset represents an employee.

* Below I will explain every column:
* Age: The age of the employee.
* Attrition: Whether the employee has left the company.
* Department: The department the employee works in.
* DistanceFromHome: The distance of the employee’s home from the office.
* Education: The level of education of the employee.
* EducationField: The field of education of the employee.
* EnvironmentSatisfaction: The employee’s satisfaction with the working environment.
* JobSatisfaction: The employee’s satisfaction with their job.
* MaritalStatus: The marital status of the employee.
* MonthlyIncome: The employee’s monthly income.
* NumCompaniesWorked: The number of companies the employee has worked for before.
* WorkLifeBalance: The employee’s work-life balance.
* YearsAtCompany: The number of years the employee has worked at the company.

Below, I will summarise the data and explain how the data fits together:

Afbeelding met tafel

Automatisch gegenereerde beschrijving

Figure 2 - Summary of the dataset

Looking at the table above, the following can be gleaned from it:

* Age: The age of the employees ranges from 18 to 60 with a mean of 36.92.
* Attrition: 237 employees have left the company and 1233 employees are still working there.
* Department: The majority of employees work in the Research & Development department.
* DistanceFromHome: The distance of employees’ homes from the office ranges from 1 to 29 miles with a mean of 9.19 miles.
* Education:
* EducationField: The majority of employees have a background in life sciences.
* EnvironmentSatisfaction: The employees are generally satisfied with their environment.
* JobSatisfaction: The employees are generally satisfied with their job.
* MaritalStatus: The most employees are married.
* MonthlyIncome:
* NumCompaniesWorked:
* WorkLifeBalance: The employees have different levels of work-life balance, with a range from 1 to 4 and a mean of 2.72. Which means that the WorkLifeBalance overall is good.
* YearsAtCompany: The number of years that the employees have worked at the company ranges from 0 to 40 and a mean of 7.01.

Since this is about attrition, I would like to see how often there is attrition and how often there isn’t.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 3 - Attrition frequency

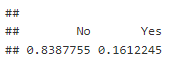


Figure 4 - Attrition frequency in %

The tables show that 1233 employees are staying and 237 leaving employees left the company. This means that about 84% of employees stayed and 16% left the company.

Now I will use visualisations to investigate what kind of people leave the company most often, and whether I can already see something of a correlation in this.

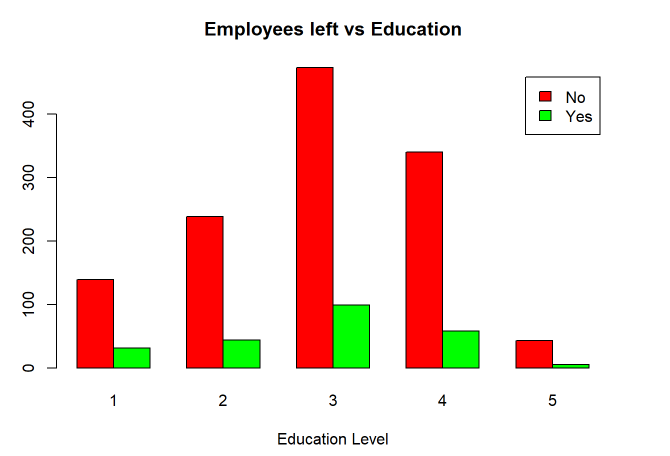


Figure 5 – Education

Those with a bachelor’s degree (level 3) have the highest attrition. Important to note with this is that there are few people who have a doctoral degree (level 5). This could have an impact on the amount that left in the Doctorate category. -Environment satisfaction: Here it can be seen that proportionally most of the people who leave the company are not or not really enjoying themselves (level 1 and 2).



Figure 6 - Satisfaction level

Here it can be seen that proportionally most of the people who leave the company are not or not really enjoying themselves (level 1 and 2). The people who are enjoying themselves more may also leave the company occasionally. However, proportionally, this happens much less.

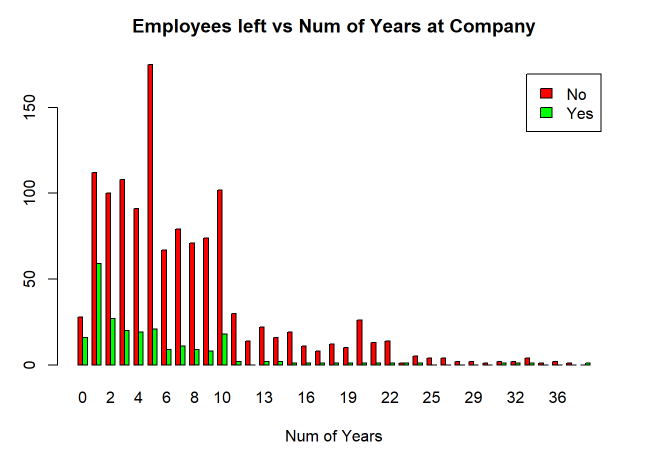


Figure 7 – Years worked at the company

Here it can be seen that most people do not work at the same company for more than 10 years. It can also be seen here that the people who do work at the same company for more than 10 years also do not leave the company.

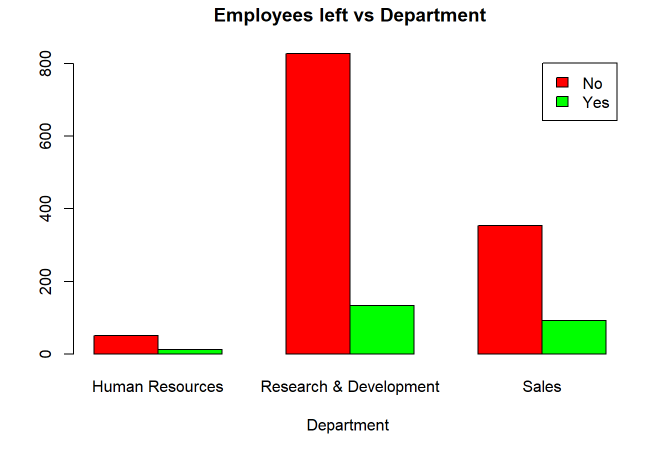


Figure 8 – Department

Most people who leave the company work in the Sales and Research & Development departments. However, it is important to note here that the HR department is comparatively smaller than the other departments.

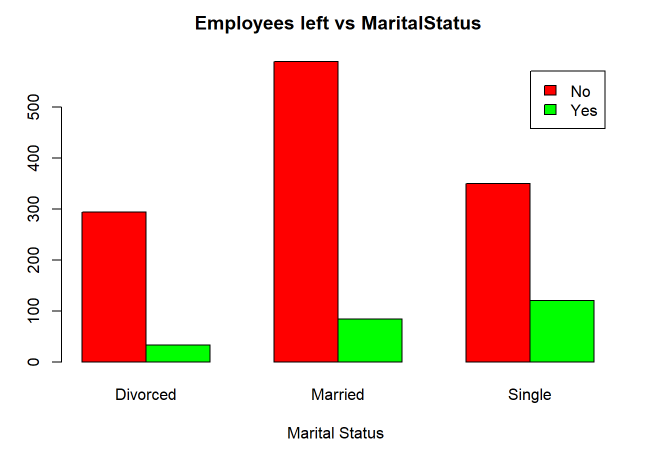


Figure 9 - Marital Status

Of those who are single, they are most likely to leave the company. Those who are divorced are less likely to leave a company.

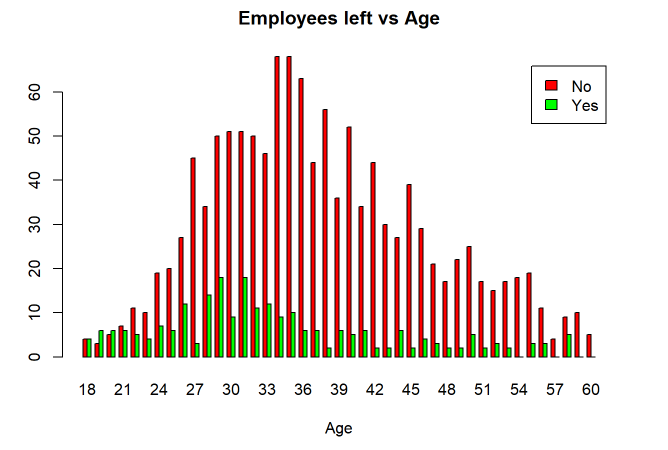


Figure 10 – Age

On average, most people leaving the company are around 30 years old and the most people that work for the company are within 30 to 45 years old.

# 3 – Data Preparation

To better define the data, I have assigned numbers to the categorial (in)dependent variables in the code below. I am doing this because R can use letters as input, but this can make interpreting the results more difficult.

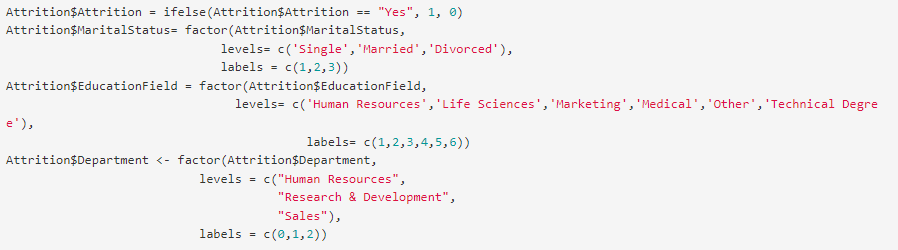


Figure 11 – Assigned numbers to categorial (in)dependent variables

As seen in the code above, the variables are still factors. Because of that, I am going to convert all columns to numeric in the code below.

Afbeelding met tekst, document, schermafbeelding

Automatisch gegenereerde beschrijving

Figure 12 - Classifying variables as numeric

Now I am going to create a model to use cook’s distance. I am using the GLM function because Attrition is a binary variable (yes or no). That’s also why the ‘binomial’ family statement is used, this is because the outcome variable is binary.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 13 - Model to calculate outliers

In the code below, I am creating a visualization to evaluate the performance of a logistic regression model that has been built using the Attrition dataset. First, I calculate the standardized residuals by taking the residuals of the model using the “deviance” method and dividing them by the square root of the ratio of deviance to the residual degrees of freedom. Then, I create a new data frame by combining the original Attrition dataset with the standardized residuals. Finally, I use the ggplot2 package to create a scatter plot of the fitted values of the model against the standardized residuals. I also add a horizontal line at y = 0 to represent a perfect fit and label the x and y axis accordingly.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 14 - Code for plotting the residuals

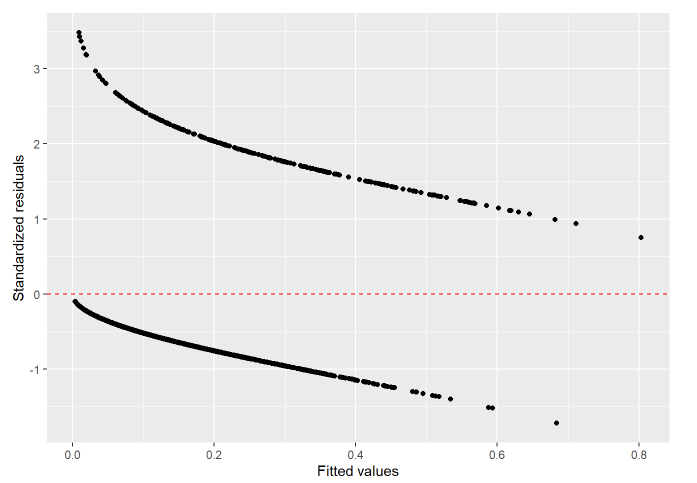


Figure 15 - Residuals plotted

Now I identified the observations standardized residuals greater than 3 and extracting them from the data frame, creating a new data frame with only those observations.

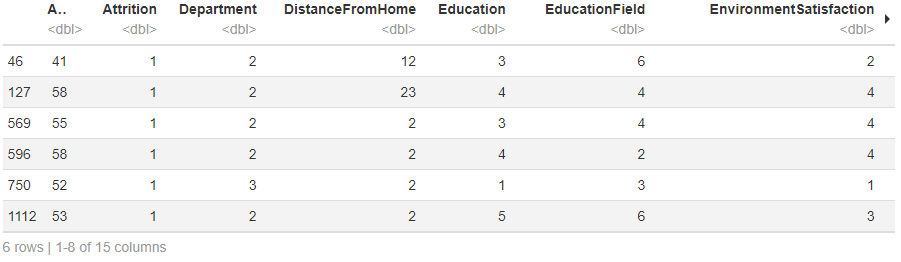


Figure 16 - Standardized residuals > 3

The below code filters the data frame by removing observations with standardized residuals greater than 3, which indicates poor model fit for those observations. This cleaning step aims to improve the model’s performance by removing outliers.



Figure 17 - Removing the outliers

I also checked the dataset for missing- and duplicate values. Those are both zero.



Figure 18 - Missing values



Figure 19 - Duplicate values

Now I am going to investigate whether there is mulitcollinearity between the independent variables. There is talk of multicollinearity between two independent variables if the correlation is between them is higher than 0.7. If there is, this could make the results less reliable. I used the abs() method to get the absolute value that is positive value doesn’t change and negative value converted into positive value. I do this to ensure that negative multicollinearity is not overlooked.

As shown in the table below, there is no multicollinearity between the independent variables.

Afbeelding met tafel

Automatisch gegenereerde beschrijving

Figure 20 - Check for multicollinearity

# 4 – Modeling

In the code below I am creating a random split the dataset into two new datasets. The “training” dataset will contain 80% of the rows from the original “Attrition” dataset, while the “testing” dataset will contain the remaining 20% of the rows. This split is done using the createDataPartition() function, which randomly selects rows to be included in the “training” dataset. The split is done only once (times = 1) and the results are not put in a list (list = FALSE).

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 21 - Splitting the data

First, I created a model with all variables and see how that does.

Afbeelding met tafel

Automatisch gegenereerde beschrijving

Figure 22 - Summary of model with all variables

In the summary above I can see which variables are insignificant. This are the variables with a p-value greater than 0.05. Based on the p-values from the summary of the linear regression model, the following variables are insignificant:

* Education
* EducationField
* Department

Then I created a model without the insignificant variables in it.

Afbeelding met tekst, tafel

Automatisch gegenereerde beschrijving

Figure 23 - Summary of model with significant variables only

The independent variables are all significant, as indicated by the small p-values (all less than 0.05).

# 5 – Evaluation and Deployment

In the first row of code below I am creating a new column in the data frame called predicted and assigning the values of the predictions made by the model. If the predicted value is greater than 0.5, it is assigned the value of 1, otherwise it is assigned the value of 0.

Below the code I will explain what the table means and how accurate this model is.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 24 - Code for creating the confusion matrix

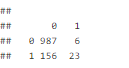


Figure 25 - Result of conclusion matrix

* The number 987 in the top left corner represents the number of observations that were actually negative (0) and were correctly predicted as negative by the model.
* The number 6 in the top right corner represents the number of observations that were actually negative (0) but were incorrectly predicted as positive (1) by the model.
* The number 156 in the bottom left corner represents the number of observations that were actually positive (1) but were incorrectly predicted as negative (0) by the model.
* The number 23 in the bottom right corner represents the number of observations that were actually positive (1) and were correctly predicted as positive by the model.

The accuracy model is used following the following formula: (number of true positives + number of true negatives) / (number of true positives + number of false positives + number of true negatives + number of false negatives).

The accuracy of this model is = (987 + 23) / (987 + 6 + 156 + 23) = (1010) / (1172) = 0.86 or 86%.

The accuracy of this model is quite good, as it means that the model is correctly predicting the outcome for 86 out of 100 observations.