Università degli Studi di Padova

Dipartimento Matematica
Dipartimento di Matematica "Tullio Levi-Civita"

Statistical Learning Final Project

Employee Attrition Classification

Authors

Zeynep TUTAR - 2106038

Aysenur Oya ÖZEN - 2107501

Supervisor

Prof. Alberto ROVERATO

Academic Year:
2023/2024

Contents

Introduction to Dataset

The aim of this project is to develop two predictive models to determine employee attrition of a company. The dataset[1] used for this project is a simulated dataset designed for the analysis and prediction of employee attrition. It contains detailed information about various aspects of an employee's profile, including demographics, job-related features, and personal circumstances.The dataset contains 74,498 samples. Each record includes a unique Employee ID and features that influence employee attrition. The goal is to understand the factors contributing to attrition and develop predictive models to identify at-risk employees.

The dataset is already split into train and test but in order to better understand the data, it is crucial to analyse the dataset as a whole.

```r
# import the train and test datasets
data_train <- read.csv("data/train.csv", stringsAsFactors = TRUE)
data_test <- read.csv("data/test.csv", stringsAsFactors = TRUE)

# merge the datasets
data <- rbind(data_train, data_test)
attach(data)
```

Description of the Features

The features of the dataset are presented below:

- **Employee ID:** A unique identifier assigned to each employee.
- **Age:** The age of the employee, ranging from 18 to 60 years.
- **Gender:** The gender of the employee
- **Years at Company:** The number of years the employee has been working at the company.
- **Monthly Income:** The monthly salary of the employee, in dollars.
- **Job Role:** The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media.
- **Work-Life Balance:** The employee's perceived balance between work and personal life
- **Job Satisfaction:** The employee's satisfaction with their job
- **Performance Rating:** The employee's performance rating
- **Number of Promotions:** The total number of promotions the employee has received.
- **Distance from Home:** The distance between the employee's home and workplace, in miles.
- **Education Level:** The highest education level attained by the employee
- **Marital Status:** The marital status of the employee
- **Job Level:** The job level of the employee
- **Company Size:** The size of the company the employee works for

[1] https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data

- **Company Tenure:** The total number of years the employee has been working in the industry.
- **Remote Work:** Whether the employee works remotely
- **Leadership Opportunities:** Whether the employee has leadership opportunities
- **Innovation Opportunities:** Whether the employee has opportunities for innovation
- **Company Reputation:** The employee's perception of the company's reputation
- **Employee Recognition:** The level of recognition the employee receives
- **Attrition:** Whether the employee has left the company

Data Analysis

In order to develop predictive models, first it is necessary to perform exploratory data analysis (EDA) and modify the format of the data if necessary.

```r
# installing required libraries
library(class)
library(car)
library(corrplot)
library(glmnet)
library(pROC)
library(knitr)
library(leaps)
library(MASS)
```

```r
# Descriptive statistics
str(data)
```

```
'data.frame':   74498 obs. of  24 variables:
 $ Employee.ID             : int  8410 64756 30257 65791 65026 24368 64970 36999 32714 15944 ...
 $ Age                     : int  31 59 24 36 56 38 47 48 57 24 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 1 2 2 2 1 ...
 $ Years.at.Company        : int  19 4 10 7 41 3 23 16 44 1 ...
 $ Job.Role                : Factor w/ 5 levels "Education","Finance",..: 1 4 3 1 1 5 1 2 1 3 ...
 $ Monthly.Income          : int  5390 5534 8159 3989 4821 9977 3681 11223 3773 7319 ...
 $ Work.Life.Balance       : Factor w/ 4 levels "Excellent","Fair",..: 1 4 3 3 2 2 2 1 3 4 ...
 $ Job.Satisfaction        : Factor w/ 4 levels "High","Low","Medium",..: 3 1 1 1 4 1 1 4 3 1 ...
 $ Performance.Rating      : Factor w/ 4 levels "Average","Below Average",..: 1 4 4 3 1 2 3 3 3 1 ...
 $ Number.of.Promotions    : int  2 3 0 1 0 3 1 2 1 1 ...
 $ Overtime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 2 2 ...
 $ Distance.from.Home      : int  22 21 11 27 71 37 75 5 39 57 ...
 $ Education.Level         : Factor w/ 5 levels "Associate Degree",..: 1 4 2 3 3 2 3 4 3 5 ...
 $ Marital.Status          : Factor w/ 3 levels "Divorced","Married",..: 2 1 2 3 1 2 1 2 2 3 ...
 $ Number.of.Dependents    : int  0 3 3 2 0 0 3 4 4 4 ...
 $ Job.Level               : Factor w/ 3 levels "Entry","Mid",..: 2 2 2 2 3 2 1 1 1 1 ...
 $ Company.Size            : Factor w/ 3 levels "Large","Medium",..: 2 2 2 3 2 2 3 2 2 1 ...
 $ Company.Tenure          : int  89 21 74 50 68 47 93 88 75 45 ...
 $ Remote.Work             : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ Leadership.Opportunities: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Innovation.Opportunities: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 2 ...
 $ Company.Reputation     : Factor w/ 4 levels "Excellent","Fair",..: 1 2 4 3 2 2 3 1 2 3 ...
 $ Employee.Recognition   : Factor w/ 4 levels "High","Low","Medium",..: 3 2 2 3 3 1 3 2 3 2 ...
 $ Attrition               : Factor w/ 2 levels "Left","Stayed": 2 2 2 2 2 1 1 2 2 1 ...
```

Data Preprocessing

To prepare the dataset for further analysis, several data preprocessing steps are performed:

1. Removing Columns

Employee.ID and Company.Tenure dropped as they are not useful for predictive modeling. Company.Tenure column gives logically incorrect numerical values.

```r
# Drop Employee Id and Company.Tenure
data <- data[, !names(data) %in% "Employee.ID"]
data <- data[, !names(data) %in% "Company.Tenure"]

# Copy Data for further Analysis
data_detailed <- data
```

2. Numerical and Categorical Variables Separation

Numerical and categorical variables were separated for targeted analysis. Summary statistics were then used to provide a quick overview of the distribution of numerical features.

Summary statistics for numerical variables reveal that the ages of employees range from 18 to 59,indicating a workforce that spans multiple generations.Employees have been with the company for a wide range of 1 to 51 years suggesting a mix of new and long-term staff.Monthly incomes vary widely, from 1,226 to 16,149 units, indicating financial situation specific to the individual, industry or other variables. Lastly, the distance from home ranges from 1 to 99 units, which suggests that while some employees live close to their workplace, others may have longer commutes. This variability highlights the diverse nature of the workforce.

```r
# Numerical and categorical variables separation
numeric_vars <- sapply(data, is.numeric)
categoric_vars <- sapply(data, function(x) is.factor(x) || is.character(x))

# Taking names from numerical and categorical variables for
# distribution graph
categoric_var_names <- names(data)[categoric_vars]
numeric_var_names <- names(data)[numeric_vars]

# Numeric values summary
summary(data[, numeric_vars])
```

```
      Age        Years.at.Company Monthly.Income  Number.of.Promotions
 Min.   :18.00   Min.   : 1.00    Min.   : 1226   Min.   :0.0000
 1st Qu.:28.00   1st Qu.: 7.00    1st Qu.: 5652   1st Qu.:0.0000
 Median :39.00   Median :13.00    Median : 7348   Median :1.0000  
 Mean   :38.53   Mean   :15.72    Mean   : 7299   Mean   :0.8329  
 3rd Qu.:49.00   3rd Qu.:23.00    3rd Qu.: 8876   3rd Qu.:2.0000  
 Max.   :59.00   Max.   :51.00    Max.   :16149   Max.   :4.0000  
 Distance.from.Home Number.of.Dependents
 Min.   : 1.00      Min.   :0.00        
 1st Qu.:25.00      1st Qu.:0.00        
 Median :50.00      Median :1.00        
 Mean   :49.99      Mean   :1.65        
 3rd Qu.:75.00      3rd Qu.:3.00        
 Max.   :99.00      Max.   :6.00        
```

3. Missing Values Analysis

No missing values were found in the data set, so there was no need to apply removing of null values operation.

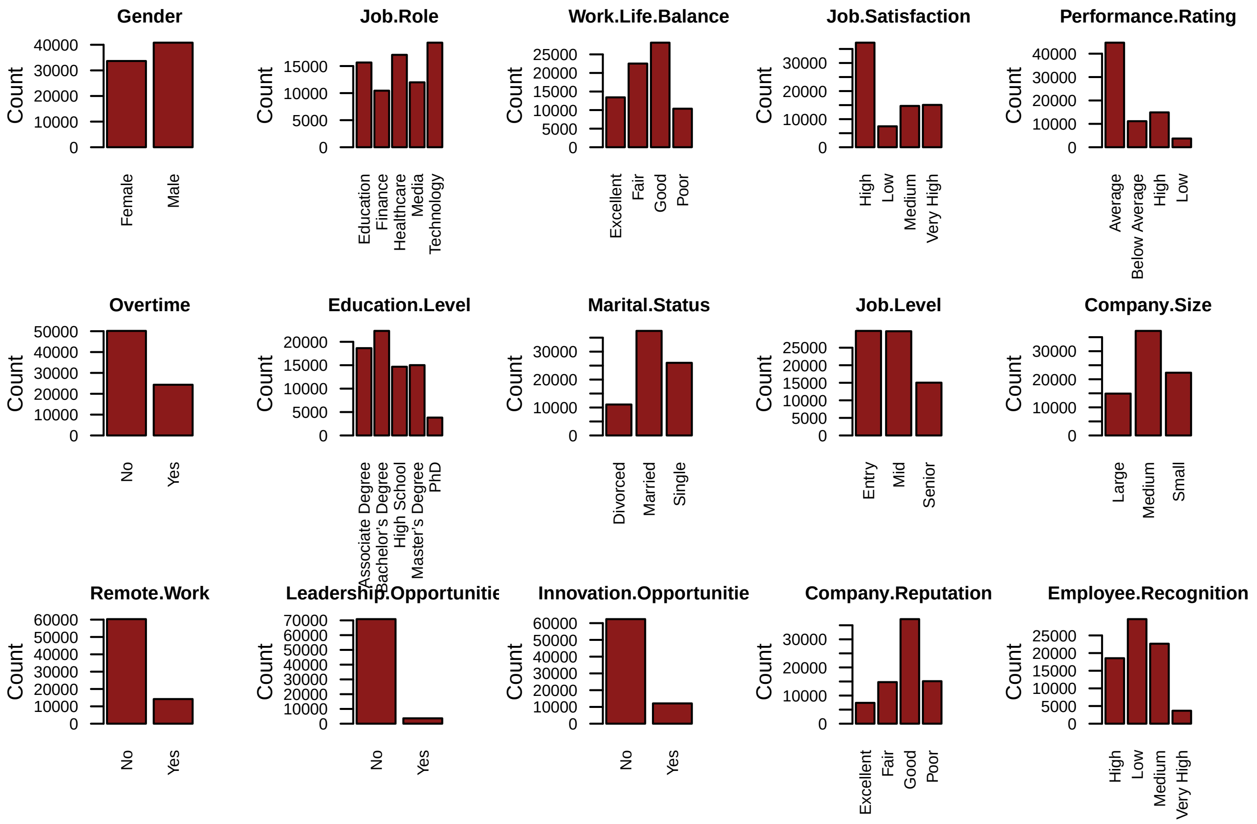
```r
# Checking if there are missing values in the entire dataset
total_na <- sum(is.na(data))
cat("Total number of missing values:", total_na, "\n")
```

```
Total number of missing values: 0
```

4. Categorical Variables Distribution

Inferences from the bar charts of categorical variables: The fact that the gender distribution is almost equal shows that there is no gender bias in the data.Job roles skew towards technology and media, indicating a focus on these sectors.Most employees rate their work-life balance as good or excellent; This indicates high satisfaction; however, job satisfaction varies, with some rating it as low or medium.Performance ratings are mostly average, indicating a need for better performance management.The fact that most employees have not received any promotions highlights potential career development issues.Overtime is rare, indicating manageable workloads.Education levels vary, and many have bachelor's degrees.Marital status indicates more married workers.Job levels are mostly entry and intermediate, indicating a younger workforce.Company size is evenly distributed, indicating different operational scales.The rarity of remote work points to traditional office culture.Limited leadership and innovation opportunities suggest areas for development.The company's reputation is mostly good, but employee recognition is low to moderate, indicating room for improvement.

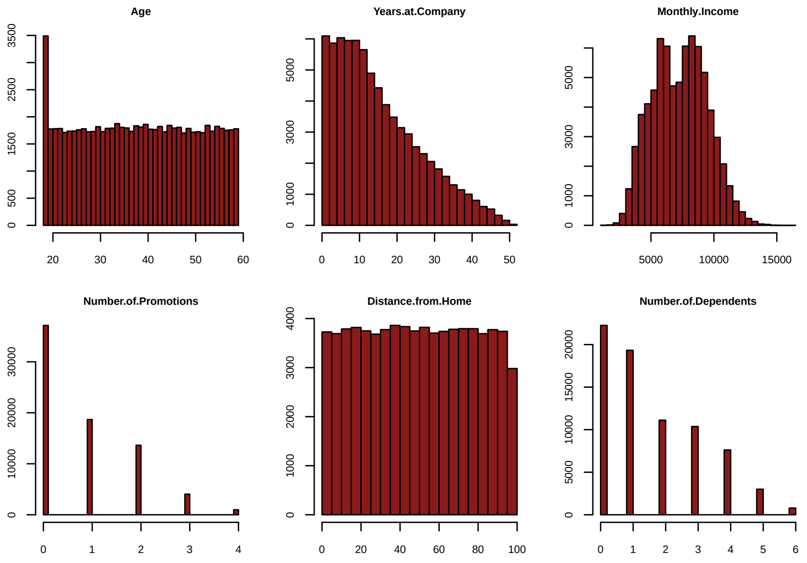
```r
# Categorical variables distribution
par(mfrow = c(3, 5), mar = c(5, 4, 2, 2) + 0.1)
for (cat_var in categoric_var_names[-16]) {
    barplot(table(data[[cat_var]]), main = paste(cat_var), ylab = "Count",
        col = "firebrick4", cex.names = 0.7, las = 2, cex.main = 0.9, cex.axis = 0.8)
}
```



5. Numerical Values Distribution

Inferences from the histograms for the numerical variables: The age distribution is fairly uniform, indicating a wide age range among employees.The years at the company show a right-skewed distribution, with most employees having shorter tenures and fewer employees staying beyond 30 years.Monthly income has a roughly normal distribution, peaking around 5,000 to 10,000 units, suggesting that most employees earn within this range.The distance from home distribution is relatively uniform, indicating that employees live at various distances from their workplace.

```r
# Numerical variables distribution
par(mfrow = c(2, 3), mar = c(3, 3, 2, 1))
for (num_var in numeric_var_names) {
    hist(data[[num_var]], main = paste(num_var), xlab = "", ylab = "", col = "firebrick4",
        breaks = 30, cex.main = 0.8, cex.axis = 0.8, cex.lab = 0.8)
}
```



6. Target Value Distribution

The pie chart depicting attrition distribution shows that 47.5% of employees left the company, while 52.5% stayed. This indicates a relatively balanced split between those who left and those who remained. Such a near-equal distribution suggests that significant turnover, highlighting the need for strategies to improve retention. Understanding the factors contributing to attrition could help in developing targeted initiatives to retain employees and enhance overall workforce stability.

```r
# Target value distribution
par(mfrow = c(1, 2))
barplot(table(data$Attrition), main = "Attrition Count", xlab = "Attrition",
    ylab = "Count", col = c("firebrick4", "rosybrown2"))

# Target values distribution with pie chart
attrition_table <- table(data$Attrition)
attrition_df <- as.data.frame(attrition_table)
colnames(attrition_df) <- c("Attrition", "Count")
attrition_df$Percentage <- round(100 * attrition_df$Count/sum(attrition_df$Count),
    1)
pie(attrition_df$Count, labels = paste(attrition_df$Attrition, " - ",
↪   attrition_df$Percentage,
    "%"), col = c("firebrick4", "rosybrown2"), main = "Attrition Distribution",
    cex = 1, radius = 1)
```