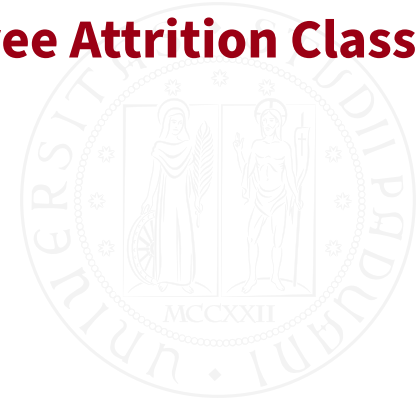
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

DIPARTIMENTO DI MATEMATICA "TULLIO LEVI-CIVITA"

STATISTICAL LEARNING FINAL PROJECT

Employee Attrition Classification



AUTHORS

Zeynep TUTAR - 2106038

Aysenur Oya ÖZEN - 0000000

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.

```
# 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, (Poor, Below Average, Good, Excellent)
- **Job Satisfaction:** The employee's satisfaction with their job: (Very Low, Low, Medium, High)
- **Performance Rating:** The employee's performance rating: (Low, Below Average, Average, High)
- **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: (High School, Associate Degree, Bachelor's Degree, Master's Degree, PhD)
- **Marital Status:** The marital status of the employee: (Divorced, Married, Single)

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

- **Job Level:** The job level of the employee: (Entry, Mid, Senior)
- **Company Size:** The size of the company the employee works for: (Small,Medium,Large)
- **Company Tenure:** The total number of years the employee has been working in the industry.
- **Remote Work:** Whether the employee works remotely: (Yes or No)
- **Leadership Opportunities:** Whether the employee has leadership opportunities: (Yes or No)
- **Innovation Opportunities:** Whether the employee has opportunities for innovation: (Yes or No)
- **Company Reputation:** The employee's perception of the company's reputation: (Very Poor, Poor,Good, Excellent)
- **Employee Recognition:** The level of recognition the employee receives:(Very Low, Low, Medium, High)
- **Attrition:** Whether the employee has left the company, encoded as 0 (stayed) and 1 (Left).

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.

```
# installing required libraries
library(car)
library(dplyr)
library(corrplot)
library(glmnet)
library(pROC)
```

```
# Descriptive statistics of DataFrame
summary(data)
```

```
  Employee.ID         Age           Gender      Years.at.Company
 Min.   :    1   Min.   :18.00   Female:33672   Min.   : 1.00
 1st Qu.:18625   1st Qu.:28.00   Male  :40826   1st Qu.: 7.00
 Median :37250   Median :39.00                  Median :13.00
 Mean   :37250   Mean   :38.53                  Mean   :15.72
 3rd Qu.:55874   3rd Qu.:49.00                  3rd Qu.:23.00
 Max.   :74498   Max.   :59.00                  Max.   :51.00
       Job.Role     Monthly.Income  Work.Life.Balance  Job.Satisfaction
 Education :15658   Min.   : 1226   Excellent:13432   High     :37245
 Finance   :10448   1st Qu.: 5652   Fair     :22529   Low      : 7457
 Healthcare:17074   Median : 7348   Good     :28158   Medium   :14717
 Media     :11996   Mean   : 7299   Poor     :10379   Very High:15079
 Technology:19322   3rd Qu.: 8876
                    Max.   :16149
    Performance.Rating Number.of.Promotions Overtime    Distance.from.Home
 Average      :44719    Min.   :0.0000       No :50157   Min.   : 1.00
```

```
 Below Average:11139    1st Qu.:0.0000       Yes:24341   1st Qu.:25.00
 High         :14910    Median :1.0000                   Median :50.00
 Low          : 3730    Mean   :0.8329                   Mean   :49.99
                        3rd Qu.:2.0000                   3rd Qu.:75.00
                        Max.   :4.0000                   Max.   :99.00
          Education.Level   Marital.Status  Number.of.Dependents  Job.Level
 Associate Degree :18649   Divorced:11078   Min.   :0.00         Entry :29780
 Bachelor's Degree:22331   Married :37419   1st Qu.:0.00         Mid   :29678
 High School      :14680   Single  :26001   Median :1.00         Senior:15040
 Master's Degree  :15021                    Mean   :1.65
 PhD              : 3817                    3rd Qu.:3.00
                                            Max.   :6.00
 Company.Size   Company.Tenure   Remote.Work Leadership.Opportunities
 Large :14912   Min.   :  2.00   No :60300   No :70845
 Medium:37231   1st Qu.: 36.00   Yes:14198   Yes: 3653
 Small :22355   Median : 56.00
                Mean   : 55.73
                3rd Qu.: 76.00
                Max.   :128.00
 Innovation.Opportunities Company.Reputation Employee.Recognition
 No :62394                  Excellent: 7414    High     :18550
 Yes:12104                  Fair     :14786    Low      :29620
                            Good     :37182    Medium   :22657
                            Poor     :15116    Very High: 3671


  Attrition
 Left  :35370
 Stayed:39128
```

```r
# Data types of columns
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 features

```r
# first column contains Employee IDs, so not necessary for analysis
data <- data[, !names(data) %in% "Employee.ID"]
```

2. Numeric and categorical value separation

```r
numeric_vars <- sapply(data, is.numeric)
categoric_vars <- sapply(data, function(x) is.factor(x) || is.character(x))

# Taking names of features
categoric_var_names <- names(data)[categoric_vars]
numeric_var_names <- names(data)[numeric_vars]

# Numeric val. 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 Company.Tenure
 Min.   : 1.00      Min.   :0.00         Min.   :  2.00
 1st Qu.:25.00      1st Qu.:0.00         1st Qu.: 36.00
 Median :50.00      Median :1.00         Median : 56.00
 Mean   :49.99      Mean   :1.65         Mean   : 55.73
 3rd Qu.:75.00      3rd Qu.:3.00         3rd Qu.: 76.00
 Max.   :99.00      Max.   :6.00         Max.   :128.00
```

3. Handling missing values

```
# Missing Values --- No null Values
na_summary <- sapply(data, function(x) sum(is.na(x)))
na_summary
```

```
                    Age                  Gender         Years.at.Company
                      0                       0                        0
              Job.Role          Monthly.Income       Work.Life.Balance
                      0                       0                        0
       Job.Satisfaction      Performance.Rating    Number.of.Promotions
                      0                       0                        0
              Overtime      Distance.from.Home         Education.Level
                      0                       0                        0
         Marital.Status    Number.of.Dependents               Job.Level
                      0                       0                        0
           Company.Size          Company.Tenure             Remote.Work
                      0                       0                        0
Leadership.Opportunities Innovation.Opportunities      Company.Reputation
                      0                       0                        0
    Employee.Recognition               Attrition
                      0                       0
```

Categorical Features

```
# Categorical val. dist.
categoric_var_names <- names(data)[categoric_vars]
for (var in categoric_var_names) {
    cat("\nDistribution of", var, ":\n")
    print(table(data[[var]]))
}
```

```
Distribution of Gender :

Female   Male
 33672  40826

Distribution of Job.Role :

 Education    Finance Healthcare      Media Technology
     15658      10448      17074      11996      19322

Distribution of Work.Life.Balance :

Excellent      Fair      Good      Poor
```

```
    13432     22529     28158     10379

Distribution of Job.Satisfaction :

     High       Low    Medium Very High
    37245      7457     14717     15079

Distribution of Performance.Rating :

      Average Below Average          High           Low
        44719         11139         14910          3730

Distribution of Overtime :

   No   Yes
50157 24341

Distribution of Education.Level :

 Associate Degree Bachelor’s Degree       High School   Master’s Degree
            18649             22331             14680             15021
              PhD
             3817

Distribution of Marital.Status :

Divorced  Married   Single
   11078    37419    26001

Distribution of Job.Level :

 Entry    Mid Senior
 29780  29678  15040

Distribution of Company.Size :

 Large Medium  Small
 14912  37231  22355

Distribution of Remote.Work :

   No   Yes
60300 14198

Distribution of Leadership.Opportunities :

   No   Yes
70845  3653
```