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.

```r
# first column contains Employee IDs, so not necessary for summary
# Descriptive statistics of DataFrame
summary(data[, -1])
```

```
      Age           Gender       Years.at.Company        Job.Role    
 Min.   :18.00   Female:33672   Min.   : 1.00    Education :15658  
 1st Qu.:28.00   Male  :40826   1st Qu.: 7.00    Finance   :10448  
 Median :39.00                  Median :13.00    Healthcare:17074  
 Mean   :38.53                  Mean   :15.72    Media     :11996  
 3rd Qu.:49.00                  3rd Qu.:23.00    Technology:19322  
 Max.   :59.00                  Max.   :51.00                      
 Monthly.Income  Work.Life.Balance  Job.Satisfaction     Performance.Rating
 Min.   : 1226   Excellent:13432   High     :37245   Average      :44719   
 1st Qu.: 5652   Fair     :22529   Low      : 7457   Below Average:11139   
 Median : 7348   Good     :28158   Medium   :14717   High         :14910   
 Mean   : 7299   Poor     :10379   Very High:15079   Low          : 3730   
 3rd Qu.: 8876                                                             
 Max.   :16149                                                             
 Number.of.Promotions Overtime    Distance.from.Home          Education.Level
 Min.   :0.0000       No :50157   Min.   : 1.00      Associate Degree :18649  
 1st Qu.:0.0000       Yes:24341   1st Qu.:25.00      Bachelor's Degree:22331  
 Median :1.0000                   Median :50.00      High School      :14680  
 Mean   :0.8329                   Mean   :49.99      Master's Degree  :15021  
 3rd Qu.:2.0000                   3rd Qu.:75.00      PhD              : 3817  
 Max.   :4.0000                   Max.   :99.00                               
  Marital.Status  Number.of.Dependents  Job.Level      Company.Size  
 Divorced:11078   Min.   :0.00         Entry :29780   Large  :14912  
```

```
 Married :37419   1st Qu.:0.00         Mid    :29678   Medium:37231
 Single  :26001   Median :1.00         Senior:15040   Small :22355
                  Mean   :1.65
                  3rd Qu.:3.00
                  Max.   :6.00
 Company.Tenure   Remote.Work Leadership.Opportunities Innovation.Opportunities
 Min.   :  2.00   No :60300   No :70845                No :62394
 1st Qu.: 36.00   Yes:14198   Yes: 3653                Yes:12104
 Median : 56.00
 Mean   : 55.73
 3rd Qu.: 76.00
 Max.   :128.00
 Company.Reputation Employee.Recognition  Attrition
 Excellent: 7414    High     :18550      Left  :35370
 Fair     :14786    Low      :29620      Stayed:39128
 Good     :37182    Medium   :22657
 Poor     :15116    Very High: 3671
```

```r
# Columns in DataFrame
colnames(data[, -1])
```

```
 [1] "Age"                      "Gender"
 [3] "Years.at.Company"         "Job.Role"
 [5] "Monthly.Income"           "Work.Life.Balance"
 [7] "Job.Satisfaction"         "Performance.Rating"
 [9] "Number.of.Promotions"     "Overtime"
[11] "Distance.from.Home"       "Education.Level"
[13] "Marital.Status"           "Number.of.Dependents"
[15] "Job.Level"                "Company.Size"
[17] "Company.Tenure"           "Remote.Work"
[19] "Leadership.Opportunities" "Innovation.Opportunities"
[21] "Company.Reputation"       "Employee.Recognition"
[23] "Attrition"
```

```r
# Data types of columns
str(data[, -1])
```

```
'data.frame':   74498 obs. of  23 variables:
 $ 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 ...
```

```r
# Number of unique values in each column
unique_values <- apply(data, 2, function(x) length(unique(x)))
print(unique_values)
```

```
             Employee.ID                      Age                   Gender 
                   74498                       42                        2 
        Years.at.Company                 Job.Role           Monthly.Income 
                      51                        5                     9842 
       Work.Life.Balance         Job.Satisfaction       Performance.Rating 
                       4                        4                        4 
    Number.of.Promotions                 Overtime       Distance.from.Home 
                       5                        2                       99 
         Education.Level           Marital.Status     Number.of.Dependents 
                       5                        3                        7 
               Job.Level             Company.Size           Company.Tenure 
                       3                        3                      127 
             Remote.Work Leadership.Opportunities Innovation.Opportunities 
                       2                        2                        2 
      Company.Reputation     Employee.Recognition                Attrition 
                       4                        4                        2 
```

Data Preprocessing

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

1. Converting categorical features to factors
2. Removing features
3. Handling na values
4. etc…

```
# EDA
```

Outliers

```
# EDA
```

Visualization

```
# EDA
```

As a result of the analysis, the following observations were made regarding the characteristics of the data:

Features vs. Target

Categorical Features vs. Target

Numerical Features vs. Target

Correlation Matrix

Partial Correlation Matrices

Data Preparation

After completing the data analysis steps, it is necessary to prepare the data for model development.

Handling Categorical Features

In order to use the categorical features in the model, we need to convert categorical features to numeric (ordinal or nominal) representations.

```
# Ordinal mappings:
balance.map <- c(Poor = 1, Fair = 2, Good = 3, Excellent = 4)
data$Work.Life.Balance <- balance.map[as.numeric(data$Work.Life.Balance)]

satisfaction.map <- c(Low = 1, Medium = 2, High = 3, `Very High` = 4)
data$Job.Satisfaction <- satisfaction.map[as.numeric(data$Job.Satisfaction)]

performance.map <- c(Low = 1, `Below Average` = 2, Average = 3, High = 4)
data$Performance.Rating <- performance.map[as.numeric(data$Performance.Rating)]
```

```r
education.map <- c(`High School` = 1, `Associate Degree` = 2, `Bachelor's Degree` = 3,
    `Master's Degree` = 4, PhD = 5)
data$Education.Level <- education.map[as.numeric(data$Education.Level)]

level.map <- c(Entry = 1, Mid = 2, Senior = 3)
data$Job.Level <- level.map[as.numeric(data$Job.Level)]

reputation.map <- c(Poor = 1, Fair = 2, Good = 3, Excellent = 4)
data$Company.Reputation <- reputation.map[as.numeric(data$Company.Reputation)]

recognition.map <- c(Low = 1, Medium = 2, High = 3, `Very High` = 4)
data$Employee.Recognition <- recognition.map[as.numeric(data$Employee.Recognition)]

size.map <- c(Small = 1, Medium = 2, Large = 3)
data$Company.Size <- size.map[as.numeric(data$Company.Size)]

# Nominal mappings: Create dummy variables for nominal data
data_numeric <- model.matrix(~., data = data)

# Convert the resulting matrix back to a data frame
data_numeric <- as.data.frame(data_numeric)[, -1]  # -1 to remove the intercept column
```

Train-Test-Split

Before splitting the data into training and test, first features and target should be defined.

```r
# Splitting data into features and target:
X <- data_numeric[, !(colnames(data_numeric) %in% c("Employee.ID", "AttritionStayed"))]

y <- data_numeric$AttritionStayed
```

Now, we can split the dataset for modelling.

```r
set.seed(42)

trainIndex <- sample(1:nrow(X), 0.8 * nrow(X))

# 80% of data is used for training
X.train <- X[trainIndex, ]
y.train <- y[trainIndex]

# 20% of data is used for testing
X.test <- X[-trainIndex, ]
y.test <- y[-trainIndex]
```

Before moving to modelling step, it is beneficial to check the dimensions and balance of the datasets.

```r
# Number of samples in train data
dim(X.train)
```

```
[1] 59598    26
```

```r
train.size <- dim(X.train)[1]

# Number of samples in test data
dim(X.test)
```

```
[1] 14900    26
```

```r
test.size <- dim(X.test)[1]

# Proportion of stayed employees for train data
prop.table(table(y.train))
```

```
y.train
        0         1 
0.4752341 0.5247659 
```

```r
# Proportion of stayed employees for test data
prop.table(table(y.test))
```

```
y.test
       0        1 
0.472953 0.527047 
```

We can observe that the train and test datasets are balanced within themselves. Also the train data is representative of test data.

Predictive Classification Models

Predictive classification models are a type of machine learning algorithm used to predict the category or class label of new, unseen instances based on historical data. These models are trained using a labelled dataset where the input features (independent variables) are associated with known class labels (dependent variable). The goal of the model is to learn the relationship between the features and the class labels so that it can accurately classify new data points into one of the predefined categories.

In this project we aim to find the risk of an employee leaving the company (class 0) and the factors affecting employee retention. So we will develop several classification models and examine their performances.

Logistic Regression

The logistic regression model estimates the odds of the dependent variable occurring and applies the logit (log-odds) transformation to express this relationship.

$$g(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \in (-\infty, +\infty)$$

Basic Logistic Classifier

```r
logit.out <- glm(y.train ~ ., data = X.train, family = binomial)

summary(logit.out)
```

```
Call:
glm(formula = y.train ~ ., family = binomial, data = X.train)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.505e+00  9.372e-02 -16.057  < 2e-16 ***
Age                          5.657e-03  9.538e-04   5.931 3.00e-09 ***
GenderMale                   5.510e-01  1.967e-02  28.008  < 2e-16 ***
Years.at.Company             1.263e-02  1.117e-03  11.305  < 2e-16 ***
Job.RoleFinance              5.782e-02  4.596e-02   1.258  0.20839
Job.RoleHealthcare           3.478e-02  4.017e-02   0.866  0.38663
Job.RoleMedia                1.053e-01  3.433e-02   3.068  0.00216 **
Job.RoleTechnology           5.270e-02  4.601e-02   1.145  0.25202
Monthly.Income               1.164e-05  7.834e-06   1.486  0.13740
Work.Life.Balance           -1.814e-01  1.038e-02 -17.478  < 2e-16 ***
Job.Satisfaction            -1.235e-01  7.955e-03 -15.522  < 2e-16 ***
Performance.Rating          -9.274e-02  1.020e-02  -9.091  < 2e-16 ***
Number.of.Promotions         2.248e-01  9.933e-03  22.631  < 2e-16 ***
OvertimeYes                 -3.351e-01  2.079e-02 -16.123  < 2e-16 ***
Distance.from.Home          -8.487e-03  3.442e-04 -24.660  < 2e-16 ***
Education.Level              1.280e-01  8.077e-03  15.844  < 2e-16 ***
Marital.StatusMarried        2.625e-01  2.808e-02   9.348  < 2e-16 ***
Marital.StatusSingle        -1.400e+00  3.032e-02 -46.162  < 2e-16 ***
Number.of.Dependents         1.320e-01  6.315e-03  20.907  < 2e-16 ***
Job.Level                    1.143e+00  1.422e-02  80.401  < 2e-16 ***
Company.Size                -1.026e-01  1.392e-02  -7.369 1.72e-13 ***
Company.Tenure               8.568e-05  4.270e-04   0.201  0.84095
Remote.WorkYes               1.612e+00  2.754e-02  58.548  < 2e-16 ***
Leadership.OpportunitiesYes  1.071e-01  4.508e-02   2.376  0.01751 *
Innovation.OpportunitiesYes  1.204e-01  2.650e-02   4.545 5.49e-06 ***
Company.Reputation          -1.200e-01  1.118e-02 -10.731  < 2e-16 ***
Employee.Recognition        -3.518e-03  1.141e-02  -0.308  0.75783
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82474  on 59597  degrees of freedom
Residual deviance: 63074  on 59571  degrees of freedom
AIC: 63128

Number of Fisher Scoring iterations: 4
```

The above results indicate that some features are insignificant to Attrition, such as; Education.LevelBachelor's Degree, Company.SizeMedium,

Logistic Regression with Backward Variable Selection

Logistic Regression with Shrinkage Method

ROC Curve & Comparison of Logistic Classifiers

Another Classification Model

Model Results

Performance Metrics and Confusion Matrix