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 splitted 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, (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.

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

- **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)
- **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.

```
# 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
 $ Performance.Rating      : Factor w/ 4 levels "Average","Below Average",..: 1 4 4 3 1 2
 $ 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
 $ 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
 $ 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…

```r
# EDA
```

Outliers

```r
# EDA
```

Visualization

```r
# 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 representations by expanding factors to a set of dummy variables. In order to avoid multicollinearity, one dummy variable will be dropped.

```r
# Create dummy variables for categorical data
data_dummy <- model.matrix(~., data = data)

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

# Check for column names
tail(colnames(data_dummy), 10)
```

```
 [1] "Remote.WorkYes"                 "Leadership.OpportunitiesYes"
 [3] "Innovation.OpportunitiesYes"    "Company.ReputationFair"
 [5] "Company.ReputationGood"         "Company.ReputationPoor"
```

```
 [7] "Employee.RecognitionLow"       "Employee.RecognitionMedium"
 [9] "Employee.RecognitionVery High" "AttritionStayed"
```

As expected for each categorical variable a new dummy column is created and one column for each category is dropped to avoid "dummy variable trap".

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_dummy[, !(colnames(data_dummy) %in% c("Employee.ID",
    "AttritionStayed"))]

y <- data_dummy$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    41
```

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

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

```
[1] 14900    41
```

```
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 
```

```
# 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 labeled 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

Basic Logistic Classifier

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