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")
data_test <- read.csv("data/test.csv")

# 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)

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

- **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)
- **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
summary(data[, -1], )
```

```
      Age           Gender          Years.at.Company   Job.Role
 Min.   :18.00   Length:74498       Min.   : 1.00    Length:74498
 1st Qu.:28.00   Class :character   1st Qu.: 7.00    Class :character
 Median :39.00   Mode  :character   Median :13.00    Mode  :character
 Mean   :38.53                      Mean   :15.72
 3rd Qu.:49.00                      3rd Qu.:23.00
 Max.   :59.00                      Max.   :51.00
```

```
Monthly.Income  Work.Life.Balance  Job.Satisfaction   Performance.Rating
Min.   : 1226   Length:74498       Length:74498       Length:74498      
1st Qu.: 5652   Class :character   Class :character   Class :character  
Median : 7348   Mode  :character   Mode  :character   Mode  :character  
Mean   : 7299                                                           
3rd Qu.: 8876                                                           
Max.   :16149                                                           
Number.of.Promotions   Overtime           Distance.from.Home Education.Level
Min.   :0.0000       Length:74498       Min.   : 1.00      Length:74498   
1st Qu.:0.0000       Class :character   1st Qu.:25.00      Class :character
Median :1.0000       Mode  :character   Median :50.00      Mode  :character
Mean   :0.8329                          Mean   :49.99                     
3rd Qu.:2.0000                          3rd Qu.:75.00                     
Max.   :4.0000                          Max.   :99.00                     
Marital.Status     Number.of.Dependents  Job.Level          Company.Size      
Length:74498       Min.   :0.00         Length:74498       Length:74498      
Class :character   1st Qu.:0.00         Class :character   Class :character  
Mode  :character   Median :1.00         Mode  :character   Mode  :character  
                   Mean   :1.65                                               
                   3rd Qu.:3.00                                               
                   Max.   :6.00                                               
Company.Tenure   Remote.Work        Leadership.Opportunities
Min.   :  2.00   Length:74498       Length:74498            
1st Qu.: 36.00   Class :character   Class :character        
Median : 56.00   Mode  :character   Mode  :character        
Mean   : 55.73                                              
3rd Qu.: 76.00                                              
Max.   :128.00                                              
Innovation.Opportunities Company.Reputation Employee.Recognition
Length:74498             Length:74498       Length:74498        
Class :character         Class :character   Class :character    
Mode  :character         Mode  :character   Mode  :character    



 Attrition        
Length:74498      
Class :character  
Mode  :character  
```

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

Categorical to Numerical Feature Conversion

Train-Test-Split

Feature and Output Samples

Predictive Classification Models

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