Reprot - IBM HR Analytics Employee Attrition Project

Introduction

Employees are the most valuable resources for any organization. The cost associated with professional training, the developed loyalty over the years and the sensitivity of some organizational positions, all make it very essential to identify who might leave the organization. Many reasons can lead to employee attrition.

Dataset

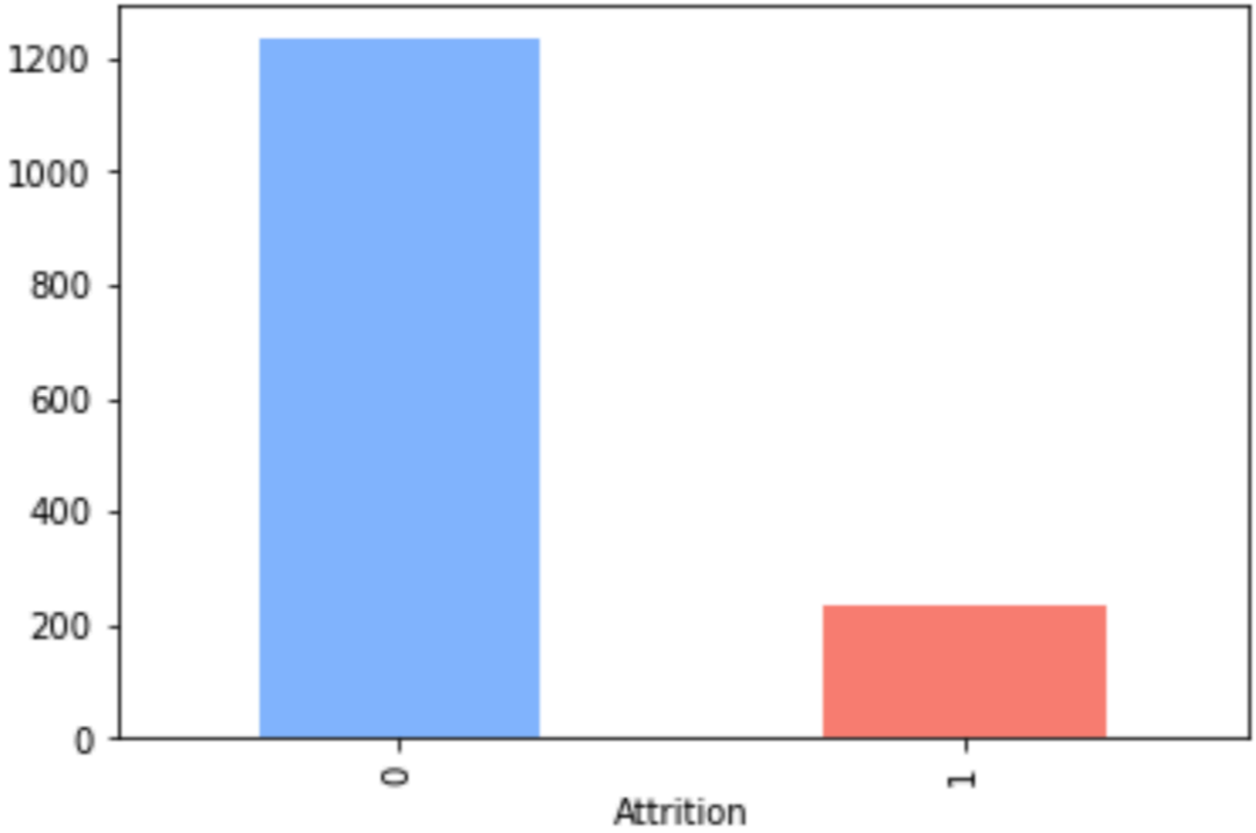
IBM HR Analytics Employee Attrition & Performance dataset is used. This dataset contains standard HR features such as age, education, gender and rate. It consists of a total of 1470 observations with 35 different attributes. IBM dataset, although it is fictional, has the characteristics which can represent real-world HR scenarios and its attributes can be readily available to the HR department in any organization. For example, the dataset includes attributes of the number of years since the last promotion, years spent in the company, number of companies the employee worked in and the training times in the last year. There was a total of 35 attributes, out of which two were the same for all data samples, i.e. standard hours and employee count.

Exploratory Data Analysis (EDA)

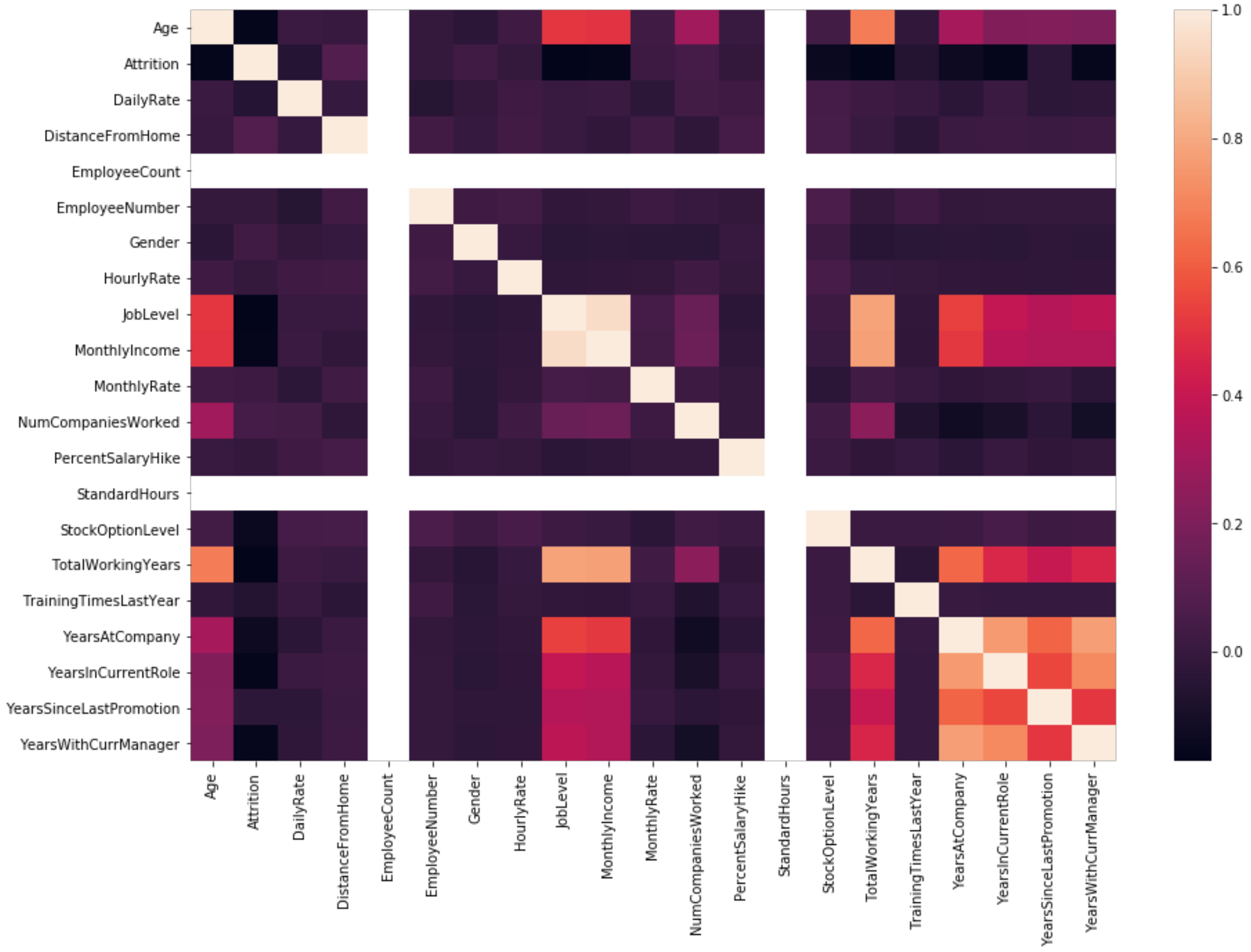
Univariate Exploration

I'll start by looking at the distribution of the main variable of interest: attrition. I found most of dataset are not attrition employees

I'll start by looking at the distribution of the main variable of interest: attrition. I found most of dataset are not attrition employees



Bivariate Exploration



Main findings from correlation

1. Monthly income is highly correlated with Job level.
2. Job level is highly correlated with total working hours.
3. Monthly income is highly correlated with total working hours.
4. Age is also positively correlated with the Total working hours.

Data Processing

1. Missing value imputation:

we found that the IBM HR dataset used in this paper contains no missing values. Thus, this step was skipped.

2. Data type conversion:

some machine learning such as logistic regression, are not able to deal with categorical variables. Thus, it is essential to convert these variables into numerical format.

3. Feature selection:

Drop 4 irrelevant columns, i.e:EmployeeCount, EmployeeNumber, Over18 and StandardHour. So, we have to remove these for more accuracy.

Modeling

The models used in this project are:

- Random Forest Model
- Logistic Regression Model
- Gradient Boosting Model

Evaluation Metrics

In IBM Attrition dataset analytics, the distribution of employees who left and those who stayed is imbalanced. For the "attrition" attribute, only 237 out of 1470 were positive, i.e. the employee who left. This imbalance should be taken into account when evaluating the proposed models. Therefore, in this work, five evaluation metrics were employed to provide a complete coverage and unbiased analysis of the results.

- Accuracy: calculated as the percentage of the correctly classified data samples by the model
- Precision: calculated as the number of TP divided by the sum of TP and false positives TP
- Recall: calculated as the number of TN divided by the sum of TP and false positives FP
- F1 Score: defined as the harmonic mean of precision and recall

Results

Based on the below findings, The Gradient Boosting has the best Accuracy and best performance.

Logistic Model

Logistic AUC = 0.62 precision recall f1-score support

```
           0       0.89      0.62      0.73       247
           1       0.23      0.62      0.34        47

    accuracy                           0.62       294
```

Accuracy score: 0.6156462585034014

Random Forest Model

Random forest AUC = 0.64 precision recall f1-score support

```
           0       0.89      0.85      0.87       247
           1       0.36      0.43      0.39        47

    accuracy                           0.79       294
```

Accuracy score: 0.7857142857142857

Gradient Boosting Classifier

Gradient Boosting AUC = 0.53 precision recall f1-score support

```
           0       0.85      0.93      0.89       247
           1       0.26      0.13      0.17        47

    accuracy                           0.80       294
```

Accuracy score: 0.8027210884353742

Future Work

I believe that a higher rate can be reached at the level of the train set and the test set if we tried to work on the different model's combinations that have similarities in the universe or have some significant hyperparameters.