IBM Employee Attrition Analysis

Abstract

This document reveals the analysis of IBM Employee Attrition dataset and find the reasons and factors which influence the attrition of employees.

Below are the analysis tasks to be performed:

1) Import attrition dataset and import libraries such as pandas, matplotlib.pyplot, numpy, and seaborn
2) Exploratory data analysis
    a. Find the age distribution of employees in IBM
    b. Explore attrition by age
    c. Explore data for left employees
    d. Find out the distribution of employees by the education field
    e. Give a bar chart for the number of married and unmarried employees
3) Build up a logistic regression model to predict which employees are like to attrite

Introduction

Attrition is a process by which employee leave the company through voluntary and involuntary means. Such as resignation or abandon.

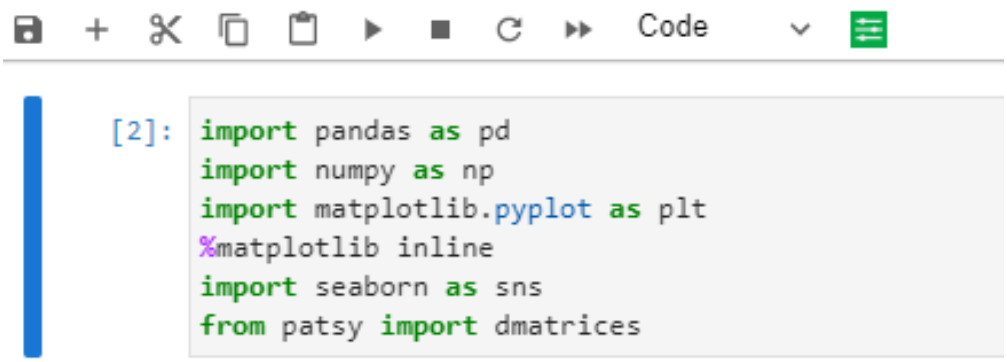
Objectives

The objective of this document is to identify the factors that contribute to the decision of employee leaving the organization and to be able to predict whether a certain employee will leave the organization by using machine learning model.

Importing libraries and dataset:

The libraries that was imported to work on the dataset are:

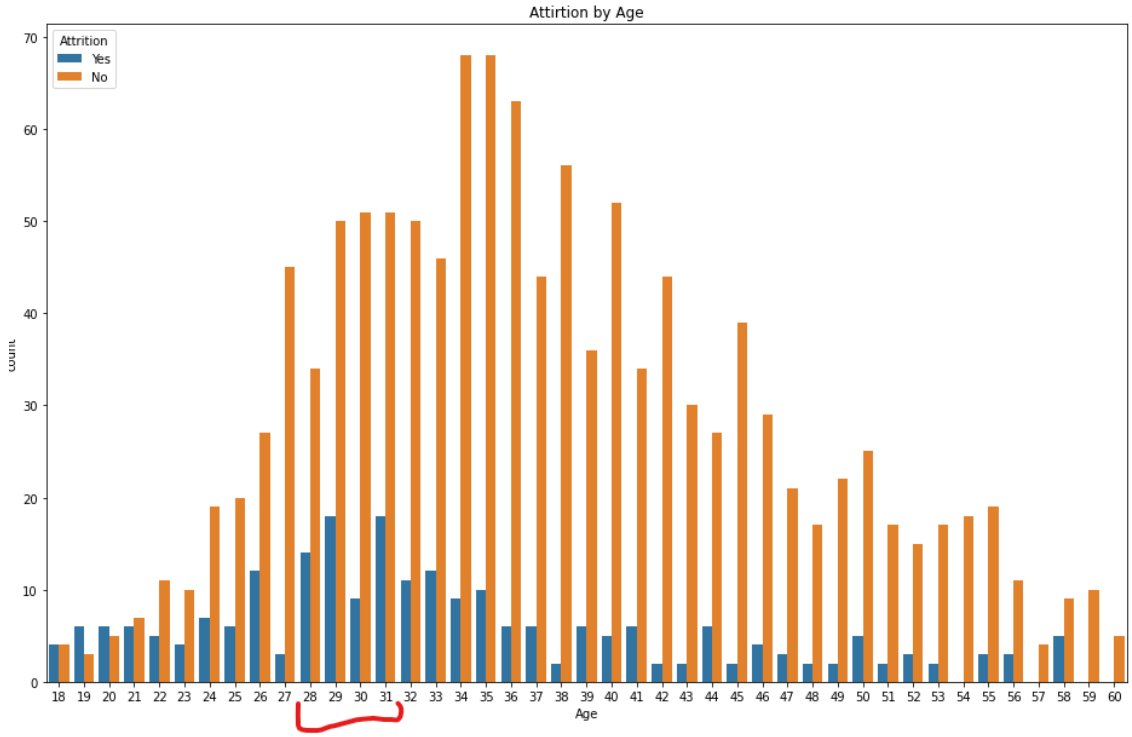
- Pandas – for analyzing, exploring and manipulating data
- Numpy – for supporting large, multi-dimensional arrays and matrices using functions
- Matplotlib.pyplot – for data visualization and graphical plotting
- Seaborn – for data visualization of statistical graphs
- Patsy – for describing statistical linear models and building design matrices



EDA (Exploratory data analysis):

The dataset was imported to the Jupyter application and imported using Pandas. The dataset was explored and following are the observations.

1) The employees between age 28 to 31 years are the most attrite

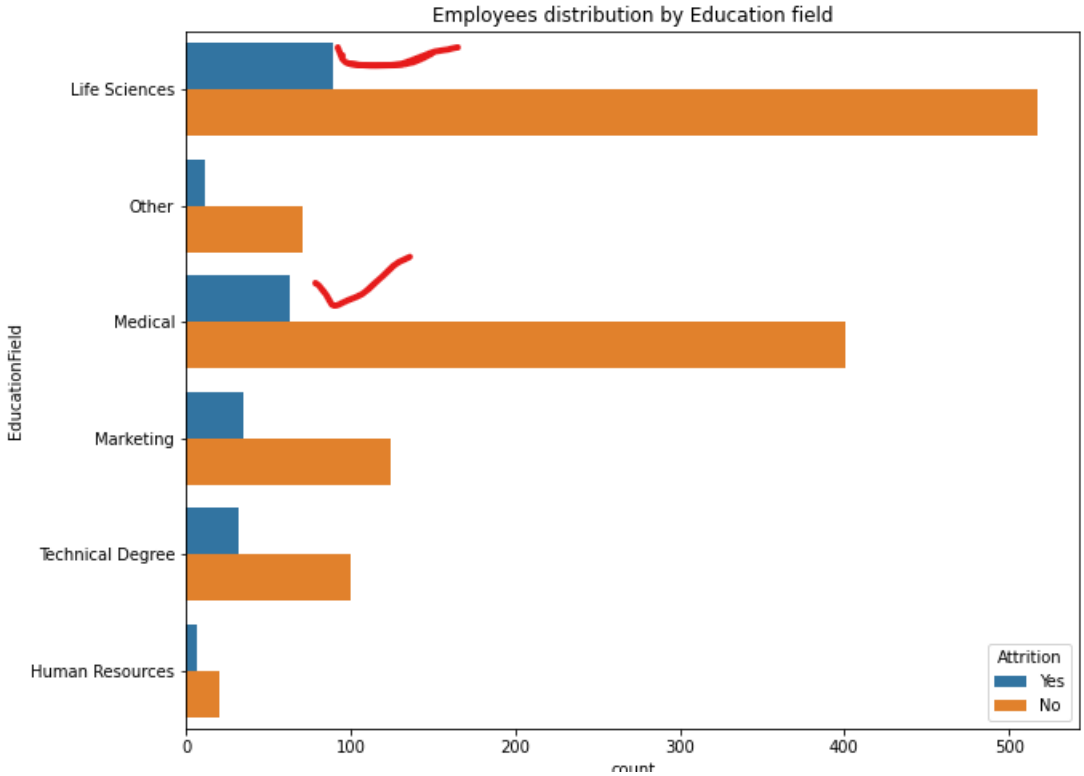


2) The number of attrite employees are 237 than 1233 who did not attrite. That means the Attrition rate is > 16% which is higher than industry standard of 10%.

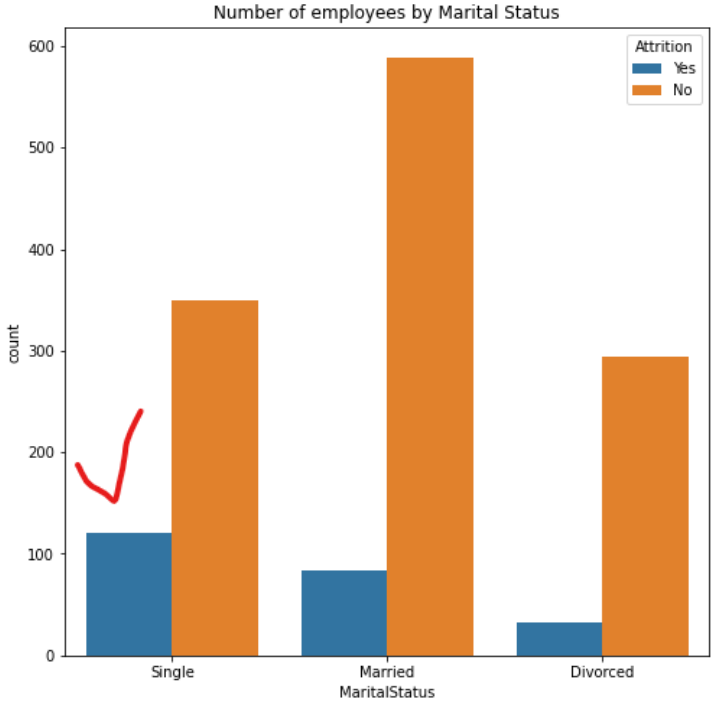
```
dataset['Attrition'].value_counts()
```

```
No     1233
Yes     237
Name: Attrition, dtype: int64
```

3) Also, the employees from education field of Life sciences and Medical are the top 2 contributors



4) Whereas Singles are the most attrite employees i.e. greater than 100 almost 50% of the attrition.



Build up Logistic regression model to predict employee attrition:

I found that the standard deviation of the “Age, Monthly Income, Distance from Home and the Yearsatcompany” are higher and were the top 4 factors which is impacting Attrition of employees.

```
[10]: ## Build Logistic Regression Model to predict which employees are likely to attrite
      dataset.describe()
```

[10]:

	Age	DistanceFromHome	Education	EnvironmentSatisfaction	JobSatisfaction	MonthlyIncome	NumCompaniesWorked	WorkLifeBalance	YearsAtCompany
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	2.912925	2.721769	2.728571	6502.931293	2.693197	2.761224	7.008163
std	9.135373	8.106864	1.024165	1.093082	1.102846	4707.956783	2.498009	0.706476	6.126525
min	18.000000	1.000000	1.000000	1.000000	1.000000	1009.000000	0.000000	1.000000	0.000000
25%	30.000000	2.000000	2.000000	2.000000	2.000000	2911.000000	1.000000	2.000000	3.000000
50%	36.000000	7.000000	3.000000	3.000000	3.000000	4919.000000	2.000000	3.000000	5.000000
75%	43.000000	14.000000	4.000000	4.000000	4.000000	8379.000000	4.000000	3.000000	9.000000
max	60.000000	29.000000	5.000000	4.000000	4.000000	19999.000000	9.000000	4.000000	40.000000

The data type of Attrition, Department, Education field and Marital status columns are object type i.e strings.

```
dataset.head()
```

	Age	Attrition	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	JobSatisfaction	MaritalStatus	MonthlyIncome	NumCompaniesWorked	WorkLifeBalance	YearsAtC
0	41	Yes	Sales	1	2	Life Sciences	2	4	Single	5993	8	1	
1	49	No	Research & Development	8	1	Life Sciences	3	2	Married	5130	1	3	
2	37	Yes	Research & Development	2	2	Other	4	3	Single	2090	6	3	
3	33	No	Research & Development	3	4	Life Sciences	4	3	Married	2909	1	3	
4	27	No	Research & Development	2	1	Medical	1	2	Married	3468	9	3	

Since, I had to use regression model to predict employee attrition, I replaced the object data type to “int64”.

```
[19]: dataset['EducationField'].replace('Life Sciences',1, inplace=True)
      dataset['EducationField'].replace('Medical',2, inplace=True)
      dataset['EducationField'].replace('Marketing',3, inplace=True)
      dataset['EducationField'].replace('Other',4, inplace=True)
      dataset['EducationField'].replace('Technical Degree',5, inplace=True)
      dataset['EducationField'].replace('Human Resources',6, inplace=True)
      dataset['Department'].replace('Research & Development',1, inplace=True)
      dataset['Department'].replace('Sales',2, inplace=True)
      dataset['Department'].replace('Human Resources',3, inplace=True)
      dataset['MaritalStatus'].replace('Married',1, inplace=True)
      dataset['MaritalStatus'].replace('Single',2, inplace=True)
      dataset['MaritalStatus'].replace('Divorced',3, inplace=True)
```

I exploited the Logistic Regression to predict the relationship between predictors “employee characteristics” and predicted variable “employee attrition” where the dependent variable is binary i.e.(1:Yes, 0:No). I compare the relationships between each category to understand which type of person is more likely to quit.

Below is the linear regression model summary for the reference.

```
model3=sm.OLS(y,x)
result3=model3.fit()
print(result3.summary())
from sklearn import metrics

print(metrics.accuracy_score(y_test, predicted))
print(metrics.roc_auc_score(y_test, probability[:, 1]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.049
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     12.47
Date:                Tue, 17 Jan 2023   Prob (F-statistic):           9.62e-14
Time:                        12:25:24   Log-Likelihood:                -578.62
No. Observations:                1470   AIC:                             1171.
Df Residuals:                    1463   BIC:                             1208.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept          0.2682      0.051      5.232      0.000       0.168       0.369
Age               -0.0051      0.001     -4.618      0.000      -0.007      -0.003
Department         0.0424      0.017      2.492      0.013       0.009       0.076
DistanceFromHome   0.0036      0.001      3.068      0.002       0.001       0.006
Education         -0.0008      0.009     -0.090      0.928      -0.019       0.018
EducationField     0.0151      0.007      2.113      0.035       0.001       0.029
YearsAtCompany    -0.0058      0.002     -3.596      0.000      -0.009      -0.003
==============================================================================
Omnibus:                      415.266   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              842.071
Skew:                           1.730   Prob(JB):                    1.40e-183
Kurtosis:                       4.332   Cond. No.                          221.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
0.8435374149659864
0.6502502887947632
```

The summary result of the linear regression model confirms that the Age, Distance from home and years at company are the factors with which company may know the reasons of employee attrition.