Employee Attrition - Factors and Prediction

IBM - HR Attrition Case Study

1. Introduction

Companies invest a lot of time and resources in employee recruiting and training, according to their strategic needs. Therefore, when an employee leaves the company, the organisation is not only losing a valuable employee, but also the resources, specifically money and HR staff effort, that were invested recruiting and selecting those employees and training them for their related tasks.

Employee attrition has an impact on productivity and profit, as the organisation must continuously invest in recruiting, training and developing new staff to fill vacant job positions. Training a new employee is a long and costly process and it is of full interest of the company to control and decrease the employee attrition rate.

2. Case Study

The goal of this project is to identify the main causes that contribute to an employee's decision to leave a company, but also to be able to predict whether a particular person will leave the company or not.

For this project I used the IBM HR Attrition Case Study. This is a fictional dataset that aims to identify important factors that might be influential in determining which employee might leave the company and who may not.

The dataset contains 1470 observations and 35 features. All features are related to the employees' working life and personal characteristics:

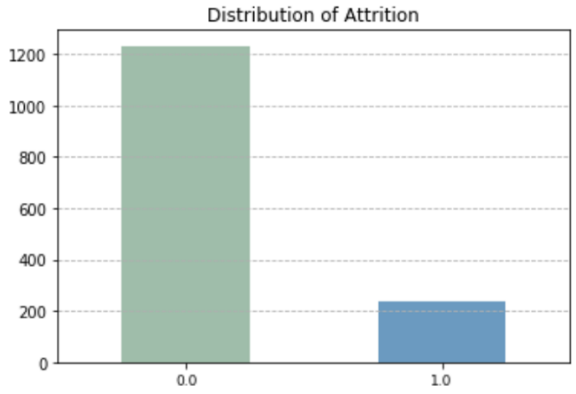
IBM - HR Attrition Dataset

Age	Employee Number	Monthly Income	Stock Option Level
Attrition	Environment Satisfaction	Monthly Rate	Total Working Years
BusinessTravel	Gender	Num Companies Worked	Training Times Last Year
Daily Rate	Hourly Rate	Over 18	Work Life Balance
Department	Job Involvement	Over Time	Years At Company
Distance From Home	Job Level	Percent Salary Hike	YearsIn Current Role
Education	Job Role	Performance Rating	Years Since Last Promotion
Education Field	Job Satisfaction	Relationship Satisfaction	Years With Curr Manager
Employee Count	Marital Status	Standard Hours	

Our target feature is the variable 'Attrition':

- **No -** represents an employee that did not leave the company
- **Yes -** represents an employee that left the company

The distribution of the target variable within the dataset, shows that out of 1470 employees, 16% (237 workers) left their jobs, while the remaining 84% (1233 workers) are still working with the company:



3. Feature Engineering

No null (*NaN*) or undefined values were found in the dataset.

Some of the features were redundant as they presented only one unique value ('Employee Count', 'Standard Hours', 'Employee Number', 'Over18'). All those columns have been removed as irrelevant for our analysis.

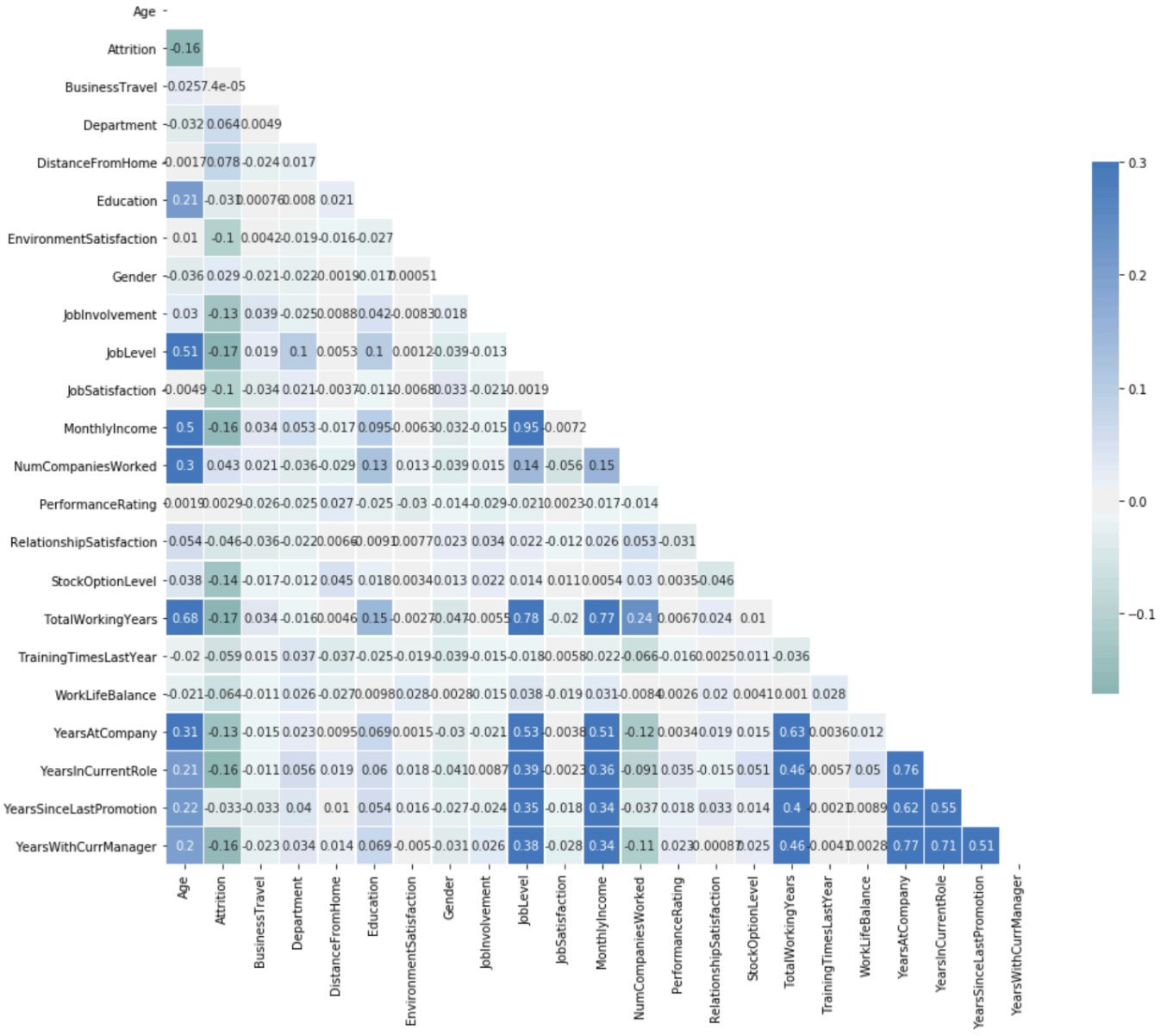
The dataset contained several variables with categorical values ('Business Travel', 'Department', 'Gender'). All these variables have been converted into numbers *(float)* so that the machine learning model could work.

4. Data Exploration - Key Findings and Insights

The below *Heatmap* graphically illustrates correlations among all variables.

Heatmaps are appealing to the eyes, and they tend to send clear messages about data almost immediately:

- **Grey** fields represent no correlation
- Green fields reveal a **negative** or indirect correlation -the variation of a characteristic inversely affects that of the other
- Blue reveals a **positive** or direct correlation - the variation of one characteristic directly affects the other



The correlations in the heatmap shows the following characteristics have high correlations:

- 'Monthly Income' and 'Job Level' have a high positive correlation, which means that employees with greater seniority generally tend to earn more
- The positive correlation among 'Years At Company', 'Years With Curr Manager', 'Years In Current Role' and 'Years since Last Promotion' highlights the absence of professional growth in the company.

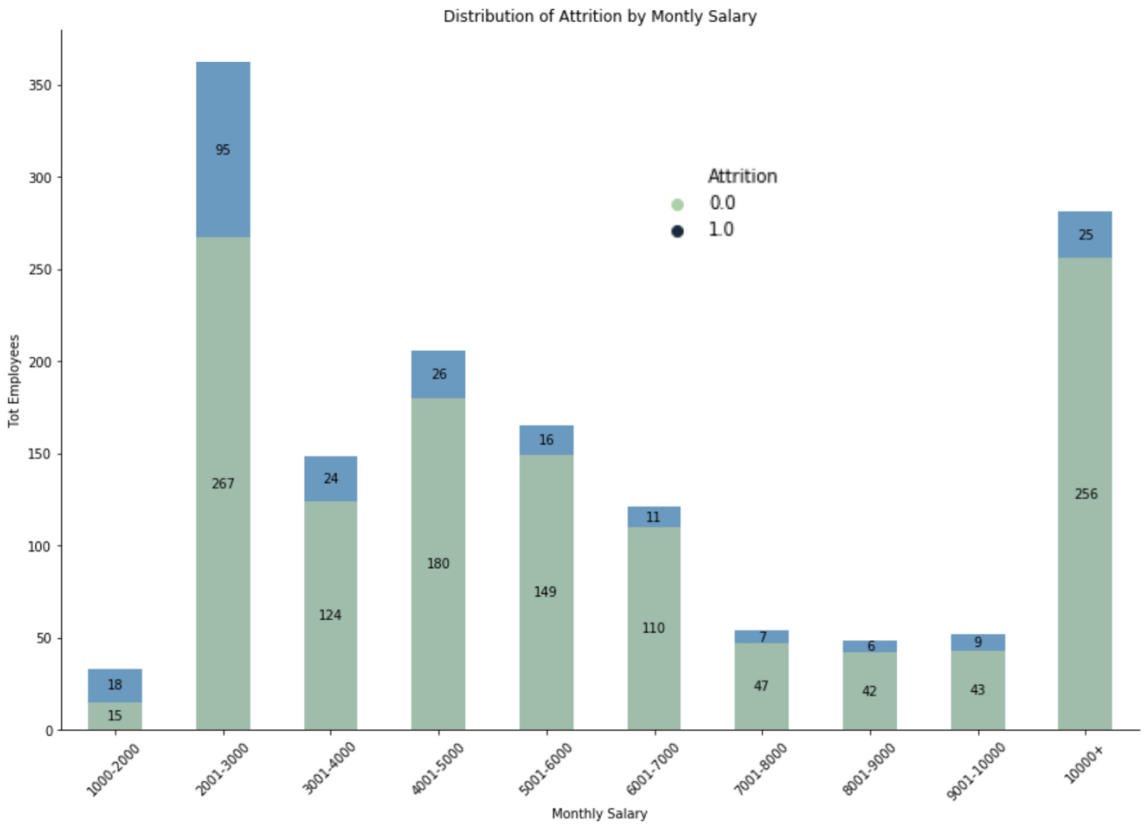
As 'Attrition' is negative correlated with all these features, it can be assumed that the lack of promotions and consequently spending years at the same Job Level without a salary increase is the main reason for employees to leave the company.

In order to have a better understanding of the correlation, the first step is to proceed with a descriptive analysis that will help to observe the distribution of the target variable within the dataset.

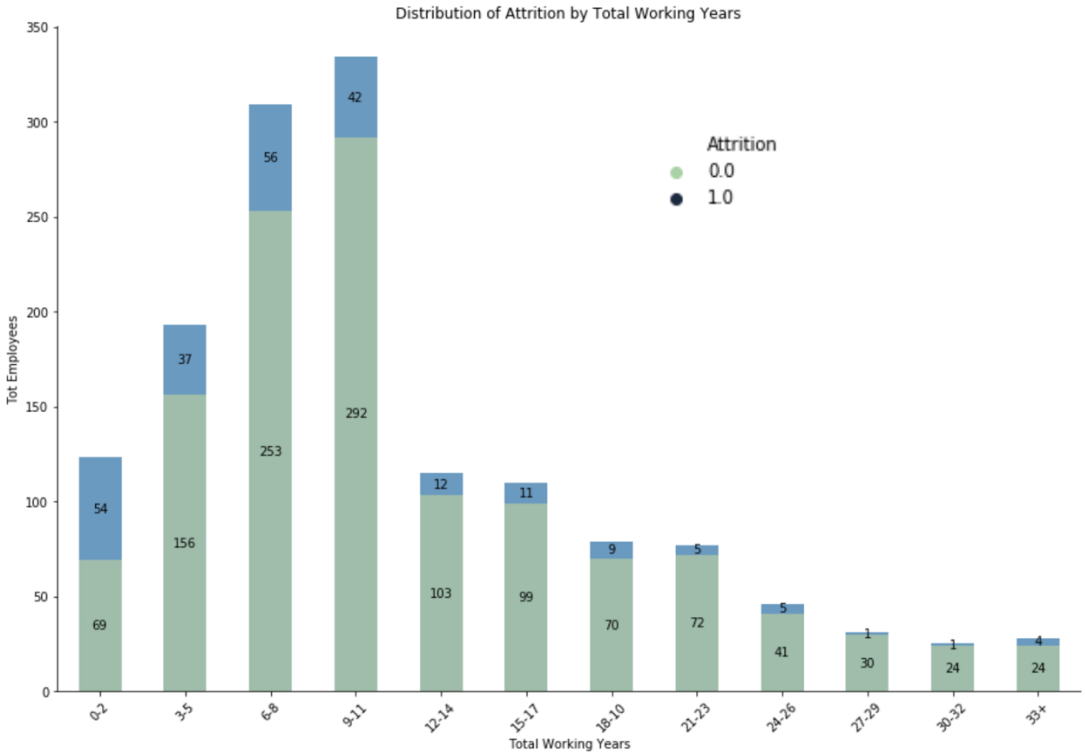
5. Descriptive Analysis

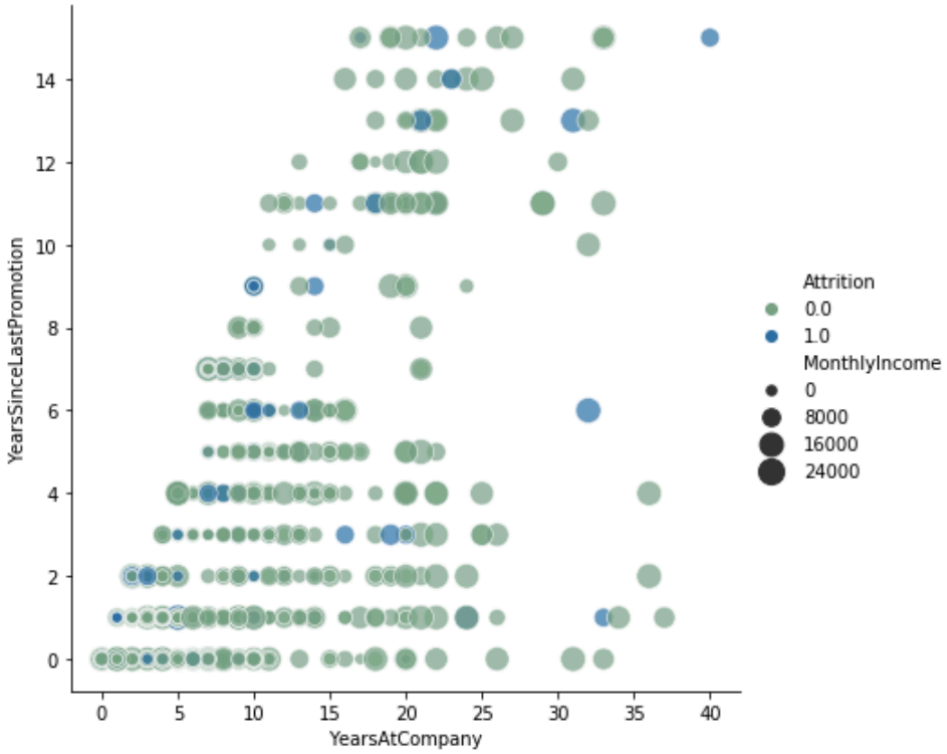
Attrition progressively decreases for higher salaries. In fact, the highest attrition rate overall is found in the '$2001–$3000' range with a percentage of 40%.

It is worth noting that the highest attrition is in the cluster '$1000–$2000' with a percentage of 54.5%.



It is possible to observe that attrition progressively decreases as the overall time working with the company increases and that the '0–2' cluster presents 44% of resigned employees within the category.





6. Building Machine Learning Models

After having analysed and prepare the data, we proceeded with the design of the prediction model to identify employees that would potentially leave the company/

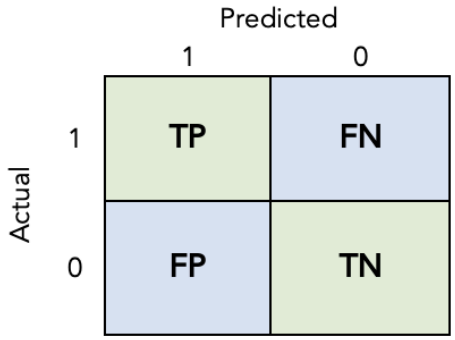
First step will be to divide the original dataset was divided into two parts:

- 70% of the data (1029) where used as **Train-set** in order to allow the model to learn the relationships hidden in the data
- The remaining 30% of the data (441) will be used as **Test-set** in order to the test and evaluate the general performance of the model and to calculate errors between predicted and actual results.

The classification algorithms considered in this section are: **Logistic Regression**, **Random Forest** and **SVM**.

In this case study, we are interested in predicting the greatest number of people who could leave the company by minimising the number of false negatives. In order to do so, a confusion matrix has been produced for each of our classification algorithms.

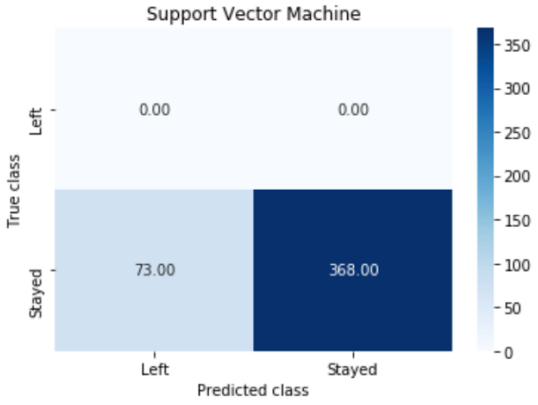
Confusion matrix is a table that is used to describe the performance of the classification model:



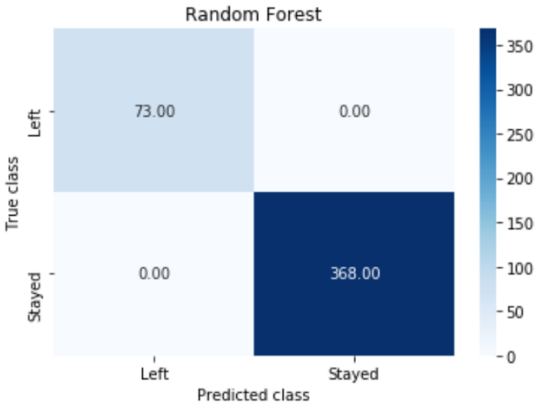
- **True Positive (TP):** We predicted an employee left and he actually did
- **True Negative (TN):** We predicted an employee stayed and he actually did
- **False Positive (FP):** We predicted an employee left but he actually stayed
- **False Negative (FN):** We predicted an employee stayed but he actually left

The following are the results of the confusion matrix applied to the selected classification models:

	precision	recall	f1-score	support
0.0	0.83	0.98	0.90	368
1.0	0.14	0.01	0.03	73
accuracy			0.82	441
macro avg	0.49	0.50	0.46	441
weighted avg	0.72	0.82	0.76	441



	precision	recall	f1-score	support
0.0	0.83	1.00	0.91	368
1.0	0.00	0.00	0.00	73
accuracy			0.83	441
macro avg	0.42	0.50	0.45	441
weighted avg	0.70	0.83	0.76	441



	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	368
1.0	1.00	1.00	1.00	73
accuracy			1.00	441
macro avg	1.00	1.00	1.00	441
weighted avg	1.00	1.00	1.00	441

7. Conclusion

We can safely say that Random Forest is the model that performed the best with a 100% accuracy as it predicted correctly all the employees that left the company.

8. Suggestions for next steps

The aim of this project was to apply machine learning techniques in order to identify the factors that may contribute to an employee leaving the company and, above all, to predict the likelihood of individual employees leaving the company.

Results obtained demonstrate that the main attrition variables are monthly income and lack of a promotion system. This is only a starting point in the development of increasingly efficient employee attrition classifiers as the periodically update of the data, the identification of new significant features and the availability of additional information on employees would improve the overall knowledge of the reasons why employees leave their companies.