Data Science Capstone Project
HR Analytics Employee Attrition and Performance

Engin Turkmen

January 28, 2019

Contents

1. Introduction:

Attrition, in Human Resource terminology, refers to the phenomenon of the employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate, which simply measures the no of employees moving out of the company (voluntary resigning or laid off by the company).

In this project, I want to predict the attrition of the company's valuable employees, uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'.

My client is IBM human resources director. He is trying to figure out the roots of employee attrition and improve the performance of company. For that, he focuses on defining the parameters which cause the employee attrition via proactive approach and tries to overcome that/those with the project's outcome.

2. Data Acquisition and Cleaning:

2.1 Data Acquisition

I use IBM HR Analytics Employee Attrition & Performance data from Kaggle, which is created by IBM data scientists. Dataset is in the open source website and can be reached from this link. It has 1470 rows x 35 columns and contains numeric and categorical data types in columns. I loaded the dataset from this link in csv format and read it in the Jupyter notebook after importing necessary libraries.

2.2 Data Specification:

The dataset has 1470 rows and 35 columns. Rows are observations from each employee and columns are from different features which are obtained in order to explain the employee attrition. The features data types consist of 27 integers and 8 objects. List of attributes are presented below.

Response Variable : Attrition (int64)

Features:

Type(Int64)	Type(Object)
Age, DailyRate, DistanceFromHome, Education, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager	BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime,

For some features, It is important to figure out their identity.

Field	1	2	3	4
Education*	Below College	College	Bachelor	Master
Environment Satisfaction	Low	Medium	High	Very High
Job Involvement	Low	Medium	High	Very High
Job Satisfaction	Low	Medium	High	Very High
Performance Rating	Low	Good	Excellent	Outstanding
Relationship Satisfaction	Low	Medium	High	Very High
Work Life Balance	Bad	Good	Better	Best

* For ‘Education’ field, 5 stands for ‘Doctor’.

2.3 Data Cleaning:

I searched for missing values in every features of dataset, all features look like having 1470 non-null entries. However, missing values can be encoded in a number of different ways, such as by zeroes, or questions marks. For that reason, I checked both missing values and duplicate values in the dataset. Luckily, it was okay to continue to next step.

I observed 5 random sample records in the dataset to grasp the general intuition about whole picture. Besides that, I explored the statistical attributes of each features such as their mean, standard deviation, interquartile values in order to detect outliers. This research also gave me a general impression about unique and top values for each attribute in addition to their frequencies in the dataset. I made double checks on some of features in order to make sure that everything is good to go. Those results were also okay.

I inspected the useless features in order to drop in the dataset. “Over 18”, ”StandardHours”, and “EmployeeCount” had only one unique value for each observation and that did not impact or change anything in the data. For that reason, I dropped those three useless columns.

To be able to use effectively in the further steps, I reassigned the response variable (Attrition) which had “Yes” and “No” values previously. They were assigned to 1 and 0 respectively. After that, I moved the response variable to the last column place.

The dataset has 8 object types which are 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'. To be able have more memory usage and

become fast, I changed object type to category type in the dataset. At first memory usage was 402.0+ KB, and after changing the data types, it became 298.3 KB.

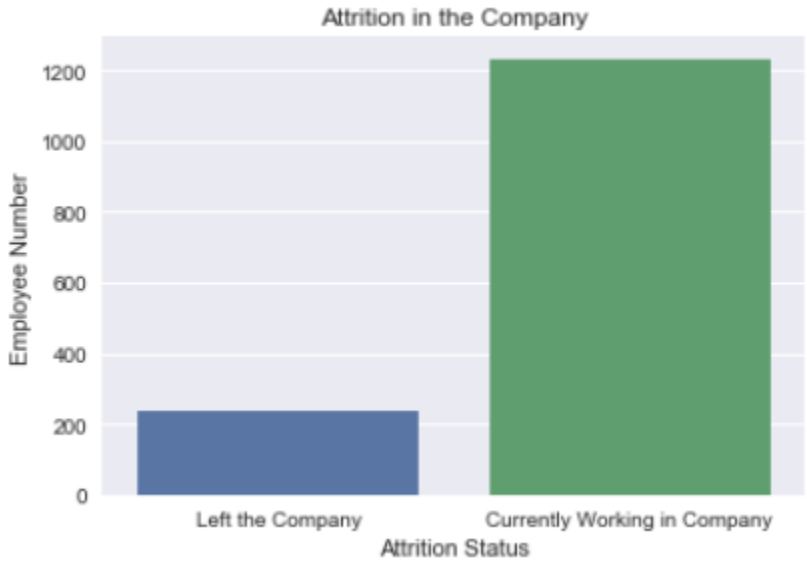
3. Exploratory Data Analysis:

3.1 Introduction:

We have 32 features consisting of both categorical as well as the numerical features. Response variable is 'Attrition' of the employees which can 1 and 0 (representing 'Yes' and 'No' respectively). This is what we will predict.

Now, I will try to analyze visually the trends in how and why employees are quitting their jobs. For that, I will deep dive into the details about features and their relationships between each other.

3.2 Target Variable (‘Attrition’):



In the company, there are 1470 employees.

237 employees who compose 16% of the total number of employees left the company for some reasons.

Besides that, 1233 employee is currently continuing to work in the same company.

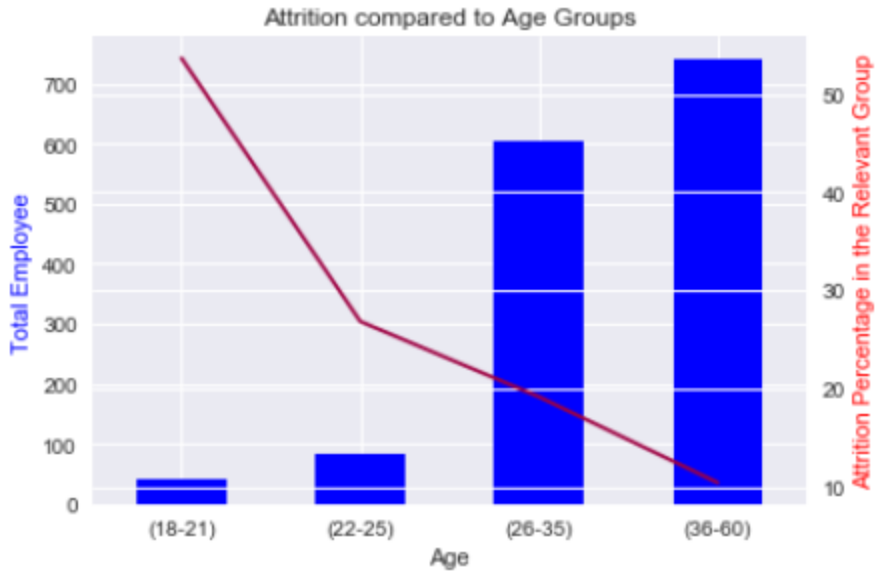
3.3 Features:

3.3.1 Age:

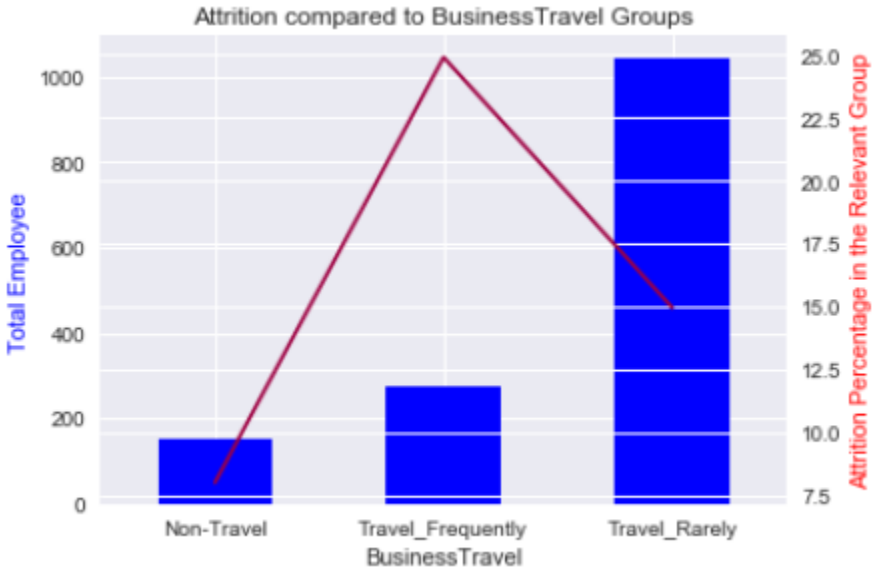
In **18-21 age group**, young employees are more likely to leave the company. Their attrition proportion to their age group is approximately 53.7% (22 out of 41) and that makes up 9% of all attrition (22 out of 237).

If we evaluate overall attrition number in the company, **26-35 age group's attrition number** is the highest comparing to other age groups. In this age group, we have 19.1 % of employee attrition(116 out 606). That makes up approximately 49% of all attrition in the company (116 out of 237).

35-60 age group generally prefers to secure their job in the same company.



3.3.2 Business Travel:



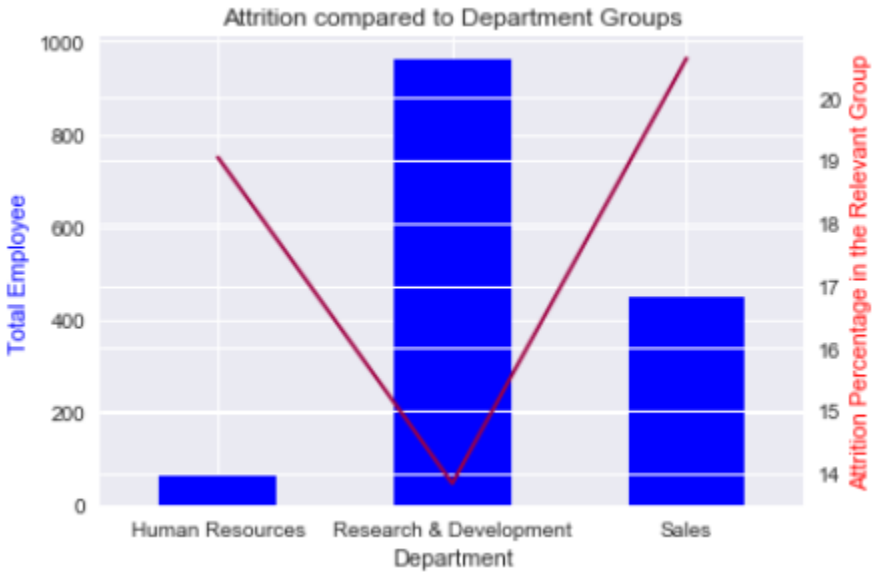
In the company, most of the employee travel rarely or don't travel according to their job description. That group compose the 81.1% of entire company(1193). The rest of the company employees which is 19.9% of them must travel frequently (277 out of 1470).

The highest attrition number with 156 belongs to the **employees who travels rarely**. That is approximately 15% of employees in that group (156 out of 1043). But when you put this number overall attrition, it makes up 65.8% of all attrition in the company(156 out of 237).

if we look at the attrition percentage of relevant travel group, the **employees who are traveling frequently** are in the danger zone. Because they have the highest attrition proportion, which is 24.9%, in their individual travel group(69 out of 277). That group's attrition rate composes of the 29.1% of overall attrition in the company (69 out of 237).

Employees who don't travel in their current role have the lowest attrition rate, which is 8%.

3.3.3 Department:



There are three departments in the company. **Research & Development Department** has the most attrition number in the company. 13.8% of **Research & Development Department** employee left the organization. In numbers, it is equal to 133, which makes us the 56.1% of all attrition in the company. Actually, this attrition is a big number for company, but compared with other departments, **Research & Development Department** has the lowest attrition rate in itself as an individual department.

Sales Department has mostly been affected by the attrition. Because 20.6% of its employees left the organization. This is the highest number compared to the other two departments. That attrition makes up 38.8% of the attrition in the company (92 out of 237).

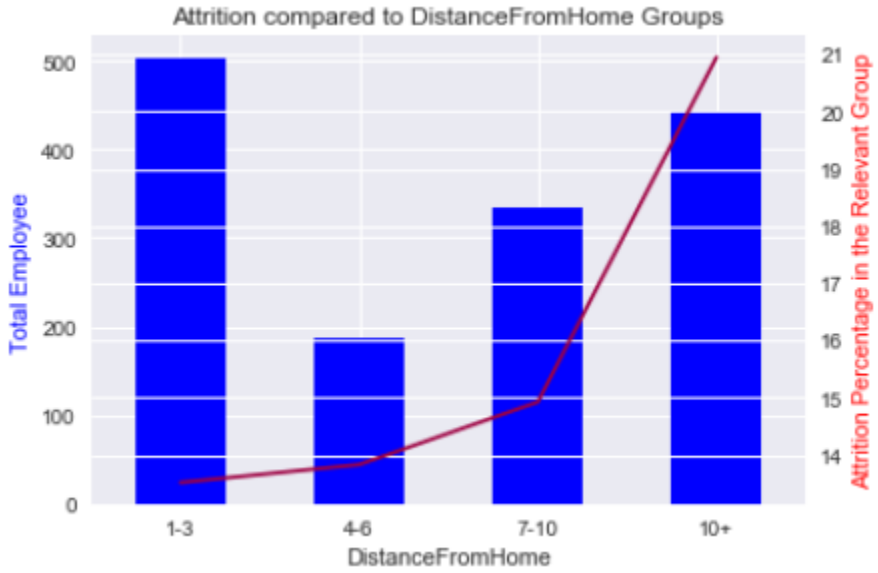
Human Resources Department follows the **Sales Department** in terms of being affected by attrition itself. 19% of that department employee left the company. But this is not that huge number in terms of whole attrition in company. **Human Resources Department** employee attrition makes up 5% of all attrition in the company (12 out of 237).

3.3.4 Distance From Home:

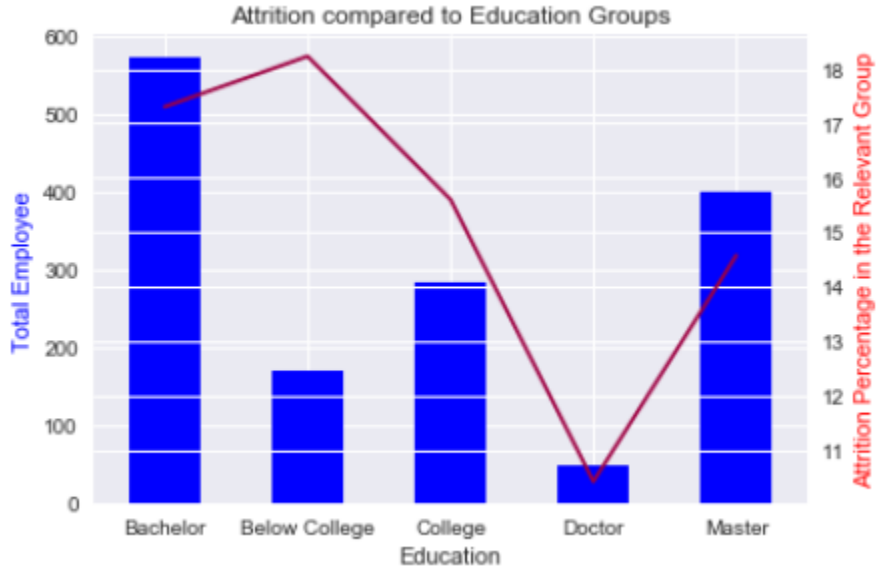
Employees whose homes are 1-3 miles far away from the company compose approximately 1/3 of the whole company employee and their attrition rate is 28.7% of all company (68 out of 237).

Also, **employees whose homes are 10+ miles far away from the company** compose approximately the other 1/3 of the whole company employee and their attrition rate is 39.2% of all company (93 out of 237).

Attrition rate within its own distance group seems to increase as the distance from home increases.



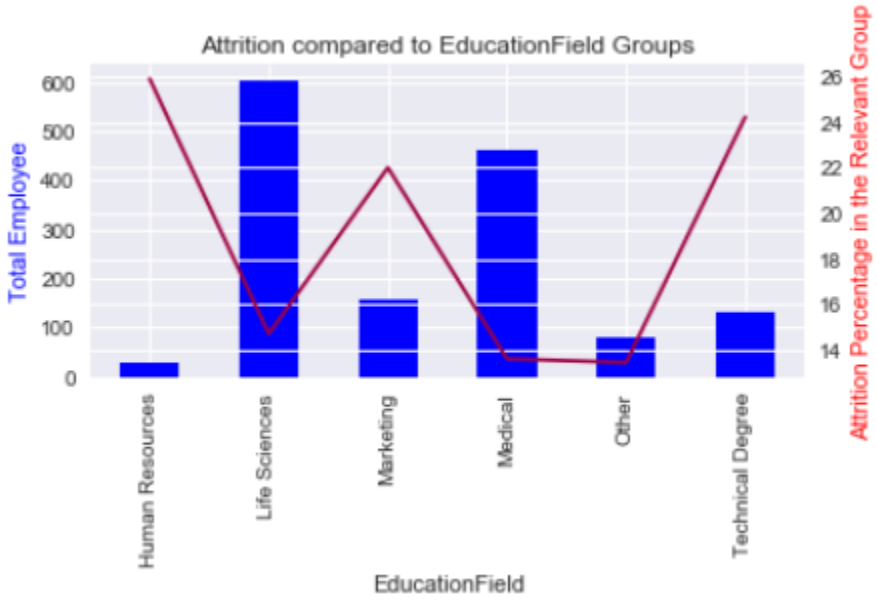
3.3.5 Education:



Employees who have bachelor's degree have the most attrition number (99 employees) in the company. That makes up 41.8% of all attrition in the company. **Employees who have Ph.D. degree** composes the least attrition number in the company.

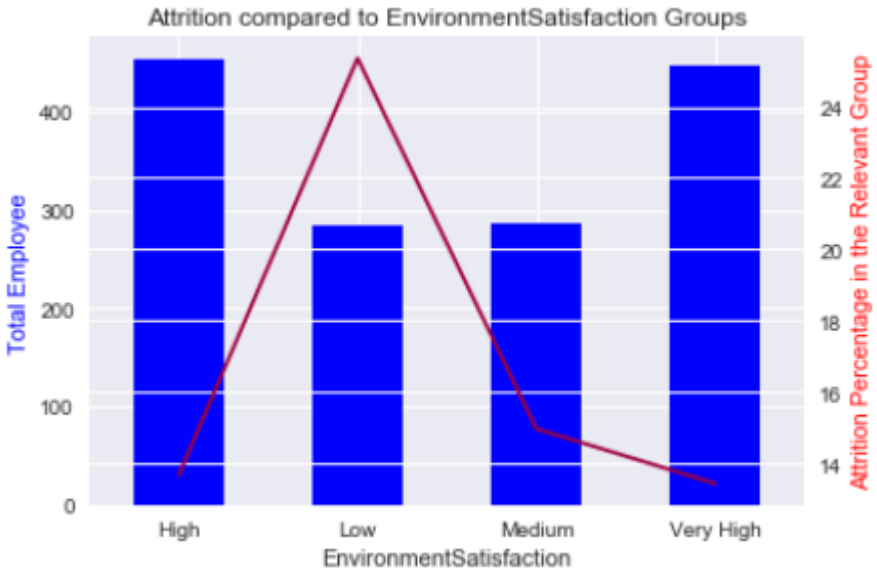
Employees who have the master, college, and below college degrees are follower of **employees who have bachelor's degrees** in terms of the attrition number in the company respectively.

3.3.6 Education Field:



Employees who have Life Science education level have the most attrition number which makes up the 37.5% of all attrition (89 out of 237). But that composes only 14.7% of attrition within Life Sciences field. **Medical** education level has the second highest attrition number which makes up the 13.57% of all attrition (63 out of 237). But that composes only 14.7% of attrition within Life Sciences field. Besides that, **Human Resources, Technical Degree, and Marketing** fields are mostly affected by the attrition respectively. Their approximately 22-26% employees left the company.

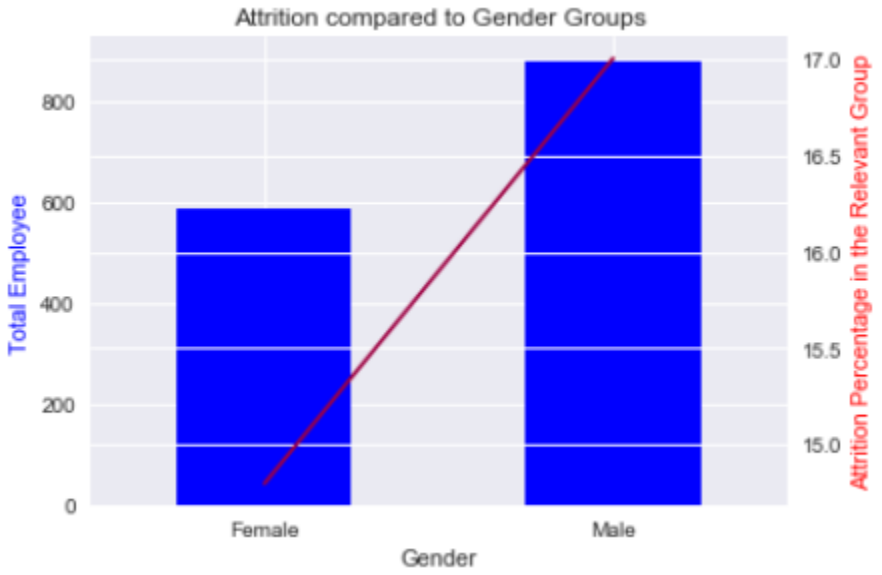
3.3.7 Environment Satisfaction:



As it may be expected, there is a high attrition rate in the **low satisfaction environment**. That composes the 30.4 % of the whole company's attrition.

Shockingly, in the **high and very high satisfaction environment**, there are still 13.7 % of the these each group's employees leave the company. That attrition composes of the 51.5 % of the whole company's attrition. This result might tell us that environment satisfaction is not the one of the main reasons for attrition in the company.

3.3.8 Gender:



Male employees are more likely to leave the company than **female employees**.

3.3.9 Job Involvement:

59% of all employee's job involvement in the company is in the **high** category(868 out of 1470). The highest attrition number is also observed in high job involvement category. 125 employees in this group, which composes the 52.7% of all attrition, left the company. But that is only 14.4% of high job involvement category.

Medium job involvement category is following the **high** category group in attrition number with 71 employees.

Low job involvement category has the highest employee leaving proportion within individual category when it is compared to the other categories. 33.7% of **Low** Job involvement group left the company.

3.3.10 Job Level:

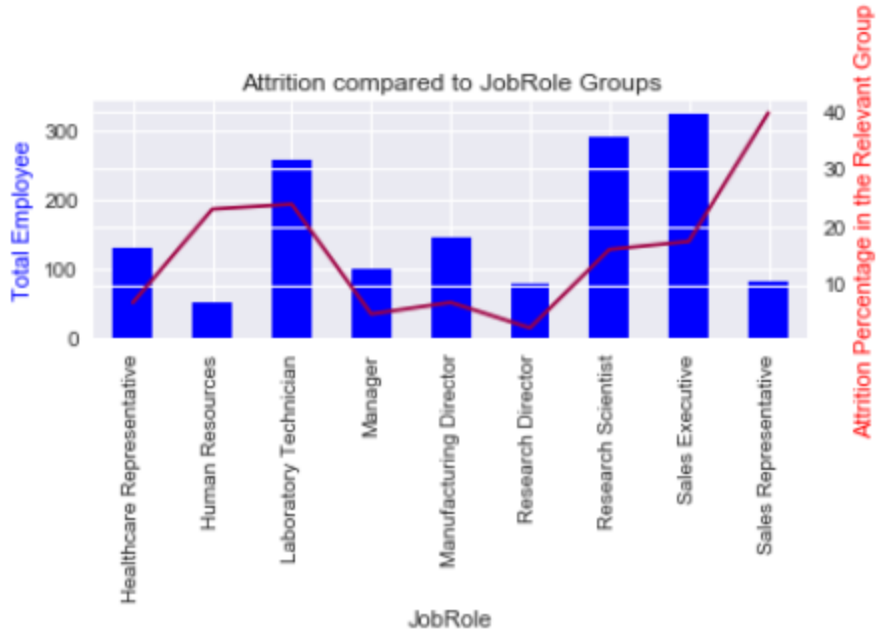
With an increase in job level, there is a decrease in attrition number throughout the company. The highest attrition is observed in the **job level-1**. 143 employees in the job level-1, who compose the 60.3% of all attrition, left the company.

3.3.11 Job Role:

Laboratory Technician has the most attrition number with the 26.2% of all attrition in the company (62 out of 237 employees). **Sales Executive** and **Research Scientist** are following the **Laboratory Technician** in attrition throughout the company with the 57 and 47 employees respectively. Those both job roles' attrition composes 44% of whole company's attrition.

Sales Representative role has been affected mostly by the attrition. Sales Representative has lost approximately 40% of its' employee. **Laboratory Technician** and **Human Resources** followed it in terms of losing employee as a job role.

On the other hand, **Research Director** job role has the lowest attrition number not only in the company (2.5%) but only within its own job role(0.8%).

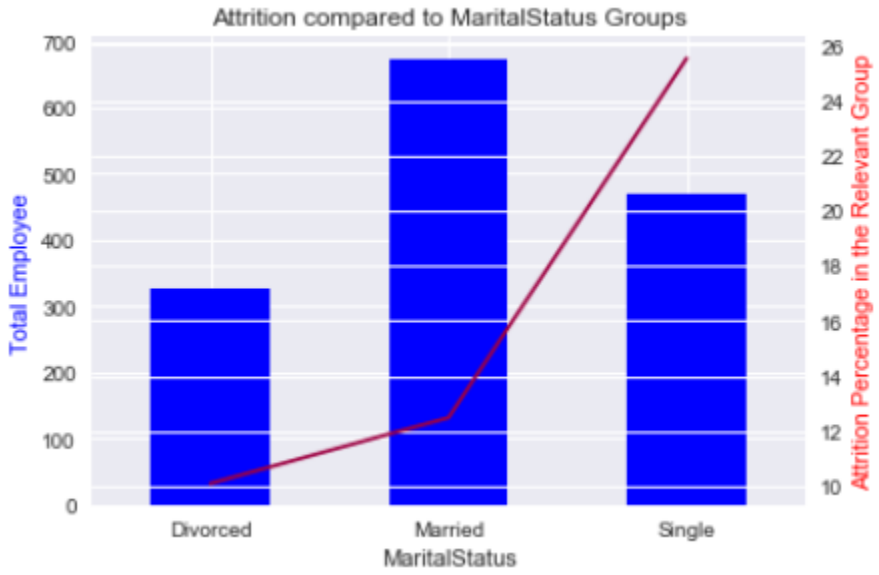


3.3.12 Job Satisfaction:

In high job satisfaction, surprisingly employees leave the company most and their attrition composes 30.8% of company's attrition. From this picture, I assume that job satisfaction should not be the main reason for employees to leave the company.

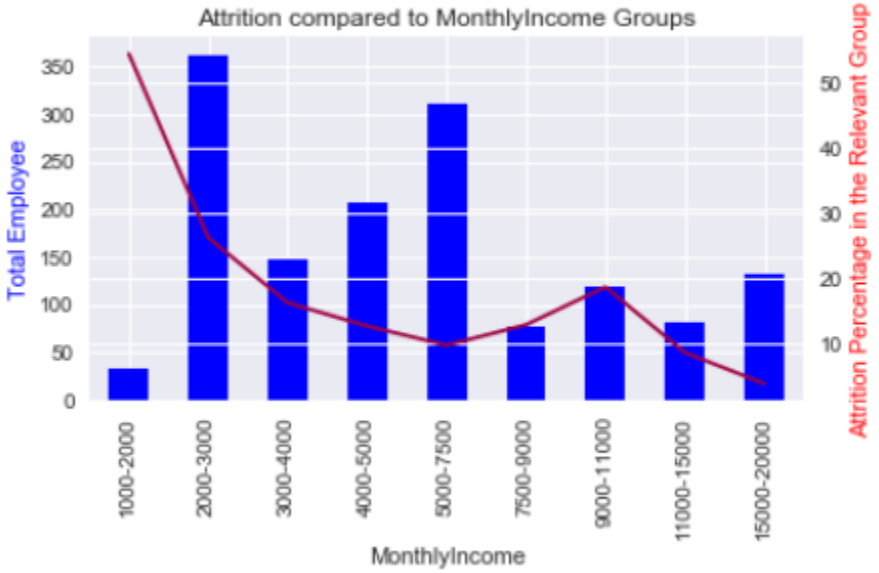
As it may be expected, in low job satisfaction, employees leave the company more than other groups except **high** satisfaction. They compose 27.8% of all attrition in the company.

3.3.13 Marital Status:



Single employees are more likely to leave the company. They have the highest attrition number and compose of the 50.6% employees who left the company. **Married** and **Divorced** employees are the followers of **Single** employees in the attrition number of the company respectively.

3.3.14 Monthly Income:



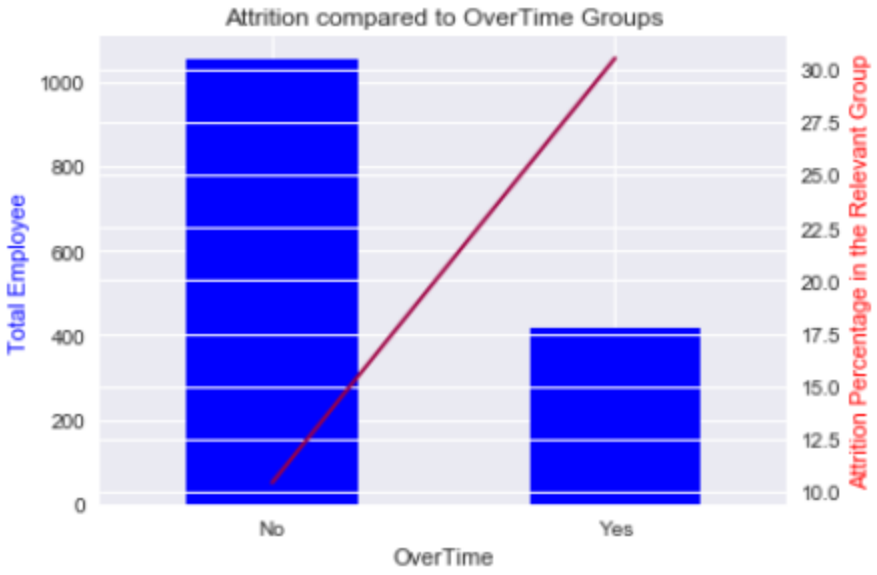
2000-**3000 dollars monthly income** level, there is a high attrition and it compose the 40% of attrition in the company. **1000-2000 dollars monthly income** level, there is a high attrition in its own income group level, which is 54.5%.

As the monthly income increase, it is observed that there is a decrease in attrition. But, in **9000-11000 dollars monthly income** level, there is a rise in attrition of its own monthly income group level.

3.3.15 Numbers Companies Worked:

35.4% of employees have **one company experience** before current company, and they are more likely to leave the company(18.8%). 21% of employees **worked in 5 and more companies** before this company and have 21.8% attrition.

3.3.16 Over Time:

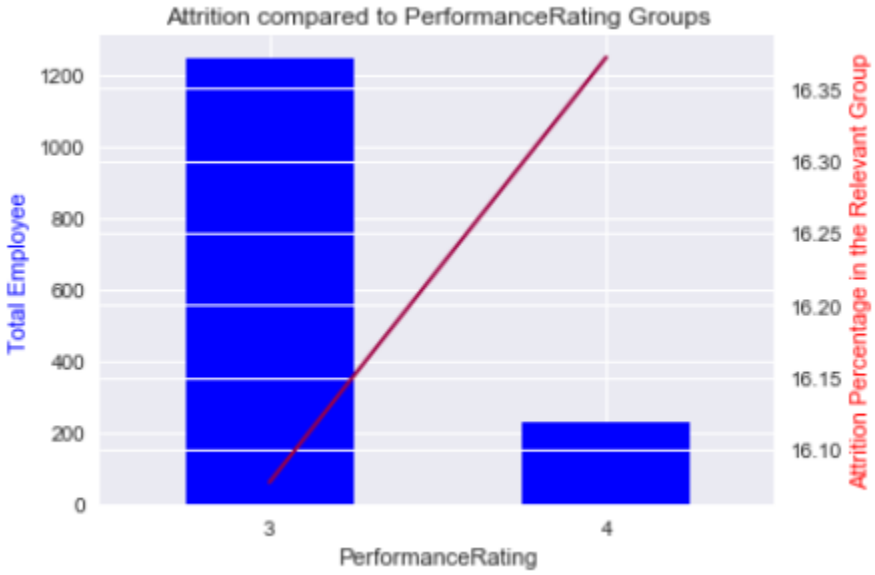


28.3% of employees have the overtime work in the company and they have higher attrition number than employees who don't have. There is not a significant difference between these two groups' attrition number. But if you compare individually both groups, over time employees are much more likely to leave the company.

3.3.17 Percent Salary Hike:

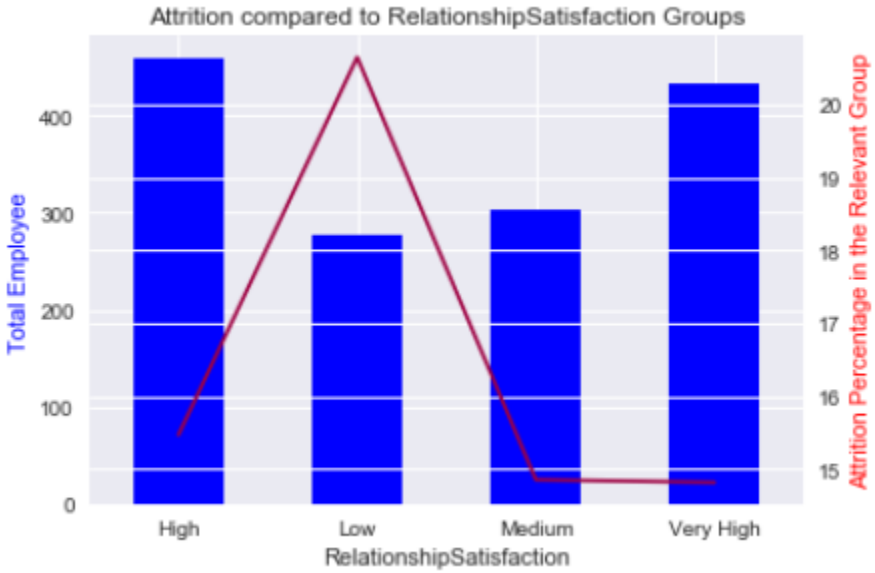
As it may be expected, the higher percent salary hike is, the more employees are likely and willingly to stay in the current company. The employees who have the highest percent salary hikes are more likely to leave the company. The reason for that might be due to the fact that they are more qualified and have the chance to find better position in other companies or due to the retirement.

3.3.18 Performance Rating:



Performance rating has two category such as 3 and 4. Not surprisingly, **performance rating 3 group** has the highest attrition number and compose 84.4% of all attrition in the company (200 out of 237 employees).

3.3.19 Relationship Satisfaction:



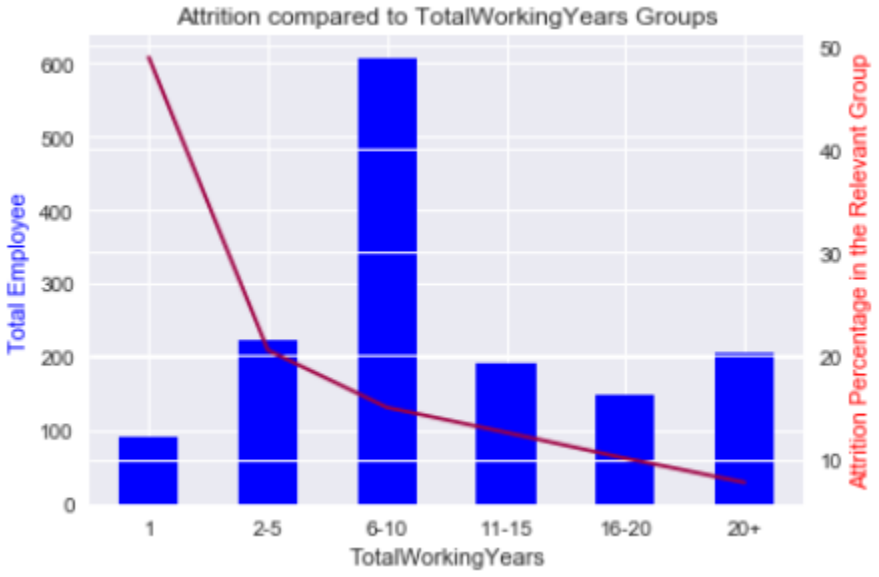
Relationship satisfaction is aligned with high and very high in the company. But, still **High** and **very high** relationship satisfaction level have the most attrition number respectively and compose of 52.7% all attrition in the company.

Besides that, the highest attrition percentage in the satisfaction group is **low** as it might be expected (20.7% of attrition in the low relationship satisfaction level.

3.3.20 Stock Option Level:

If **stock option level** is 0, there occurs a huge attrition in the company and it composes the 65% of the all attrition in the company. Besides, as the stock option level increase, there is a decrease in attrition number.

3.3.21 Total Working Years:



Employees who have one year or less working experience are more likely to leave the company and compose the 18.98% of all attrition throughout the company. In addition to that, **employees who have 6-10 years' experience** have also second highest attrition percentage throughout the company and it compose the 38.4% of all attrition.

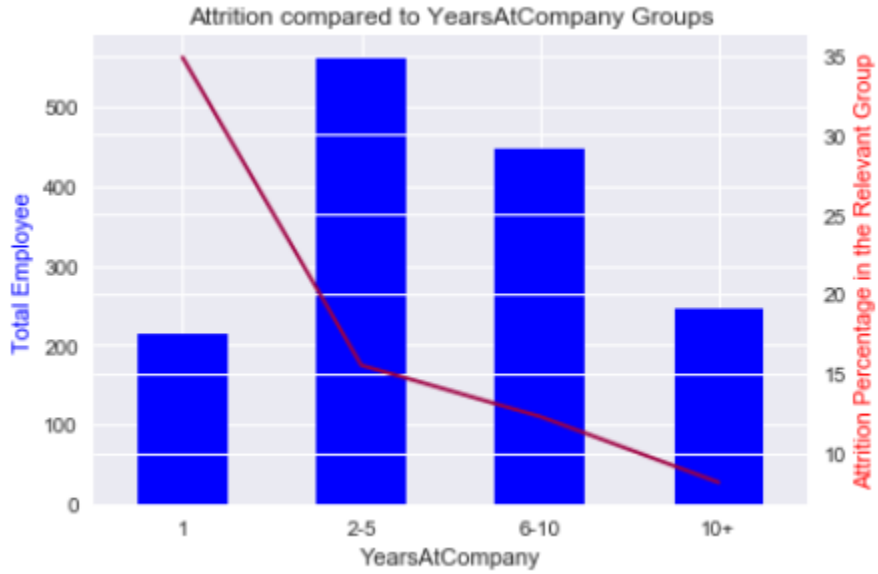
3.3.22 Training Times Last Year:

Employees who has **2 and 3 times training last year** has the most attrition number respectively and both of their attrition compose the 70.5% of all attrition in the company. Employees who don't have training time beforehand has the highest attrition number in its individual group.

3.3.23 Work Life Balance:

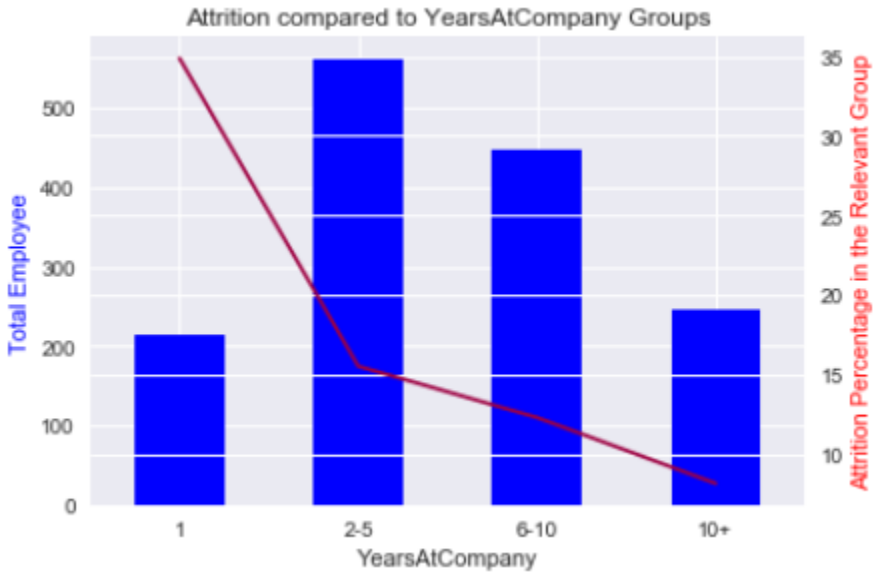
In general, work life balance is satisfactorily good throughout the company. But we have the highest attrition number and percentage throughout the company. Besides, bad work life balance group has highest attrition percentage in its individual group.

3.3.24 Years at Company:



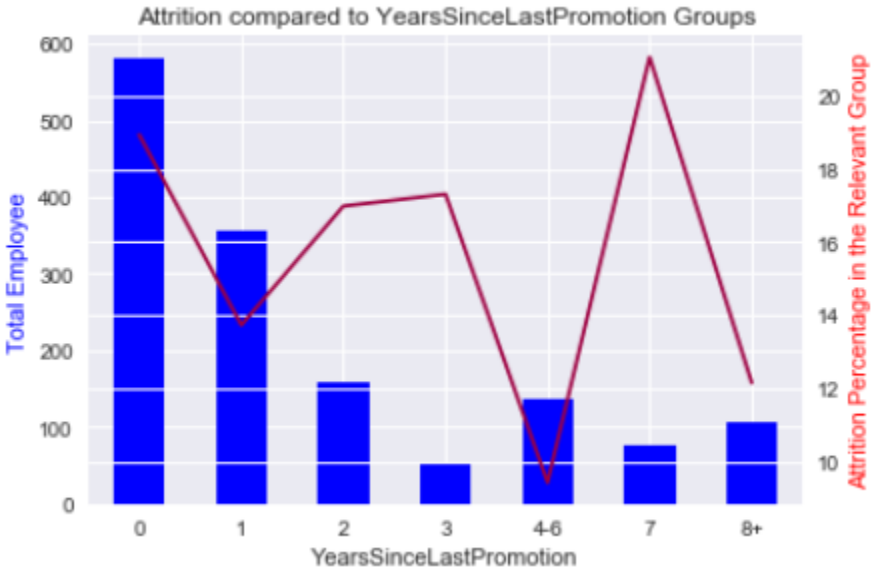
Employees who have one year or less working experience in the company has the highest attrition percentage in its individual experience group (34.9% of attrition in relevant group). Besides that, highest attrition number, which is 87 employees, is in the 2-5 years working experience at the company and that composes of the 36.7 % of all attrition in the company.

3.3.25 Years in Current Role:



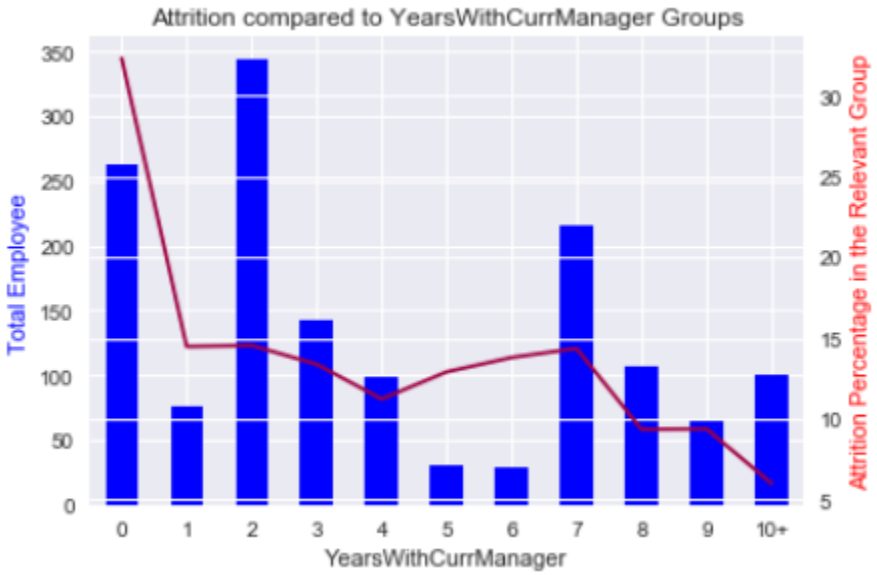
Employees who don't fulfill their first year and in their first year in their current role are more likely to leave the company. That might be result of challenge or not satisfied with the current role. Employees who have 2-5 years' experience in that company compose of the maximum attrition percentage and number in the company. Besides that, after years in current role, employees are willing to leave the company. That might be result of looking for better opportunities in other companies.

3.3.26 Years Since Last Promotion:



Employees who don't fulfill his one year since the last promotion in the company are more likely to leave the company(46.4% of all attrition). And employees who have one- and two-years' experience in the current company since the last promotion have the highest attrition number after the above group in the company respectively. 7 years since last promotion has the highest attrition in its individual group.

3.3.27 Years with Current Manager:



Most of the employee quit the company before completing their first year with their current manager. Other group who leaves the company most is the ones who work two years with current manager.

3.3.28 Other Features:

I also checked the 'Employee Number', 'Daily Rate', 'Hourly Rate' and 'Monthly Rate' features as I did the in previous features of dataset. But there is nothing significant to comment or visualize about these features. For that reason, I didn't include them in my notebook.

3.4. Feature/Variable Relationships:

In this section, I looked at how variables related to each other. There are various methods/visualizations for this. I used correlation matrix (heat map) for this purpose.

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation. A positive correlation is a relationship between two variables in which both variables either increase or decrease at the same time. An example would be height and weight. Taller people tend to be heavier. A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature). A zero correlation exists when there is no relationship between two variables. For example, there is no relationship between the amount of tea drunk and level of intelligence.

Strength of Correlation:

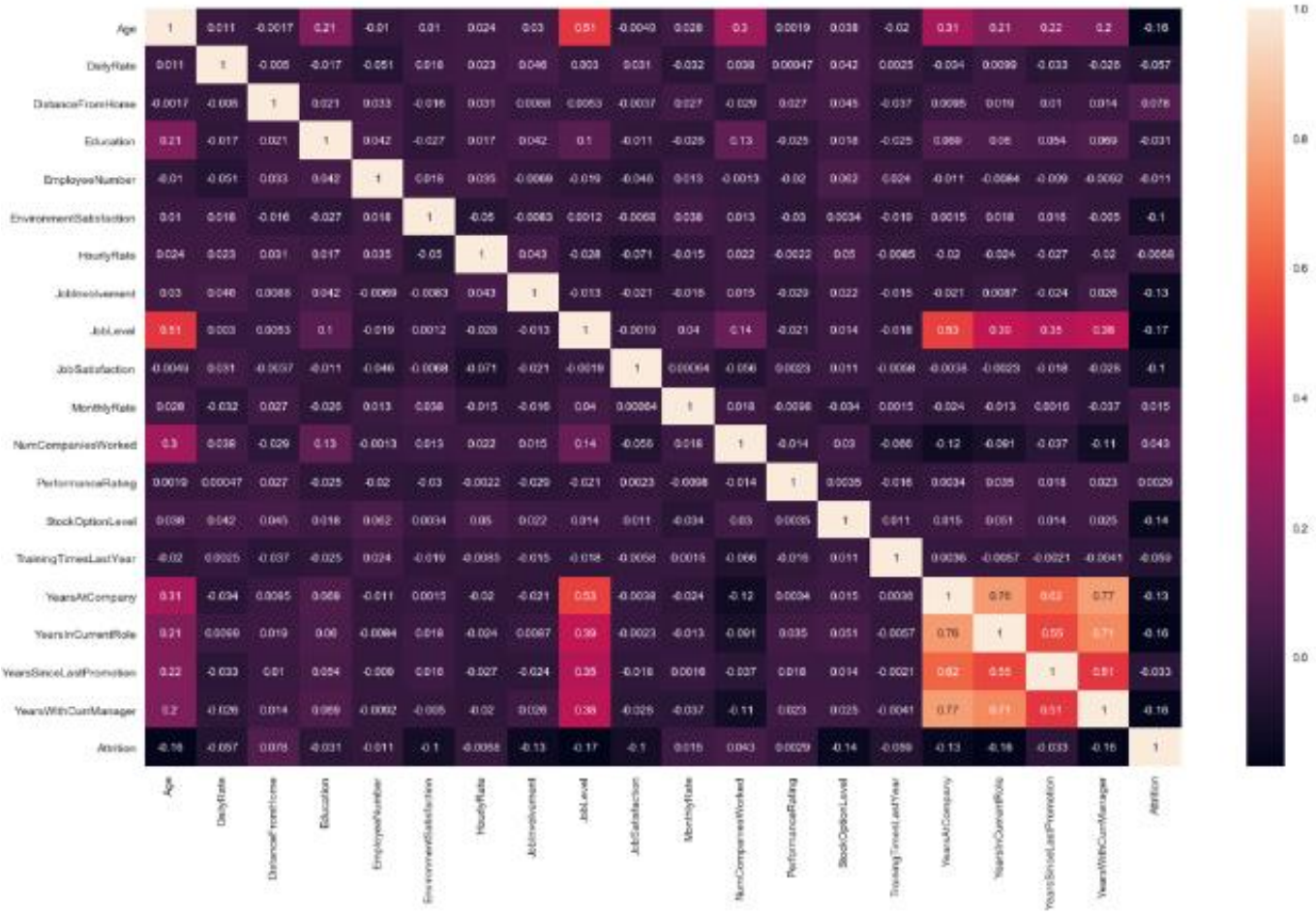
Perfect +1 -1

Strong +0.9, +0.8, +0.7 -0.9, -0.8, -0.7

Moderate +0.6, +0.5, +0.4 -0.6, -0.5, -0.4

Weak +0.3, +0.2, +0.1 -0.3, -0.2, -0.1

Zero 0



Based on the fact which is given strength of correlation chart, we can identify the features which have strong, moderate, weak and zero correlations between each other. I will just outline the strong and moderate correlations here.

Features which have strong correlations:

Percent Salary Hike and Performance Rating,
Total Working Years, Monthly Income and Job Level,
Years at Company, Years with Current Manager, and Years in Current Role,

Features which have moderate correlations:

Age has moderate correlation with Total Working Years, Monthly Income, and Job Level,

Job Level has moderate correlation with Years at Company and Age,

Total Working Years has moderate correlation with Years with Current Manager, Years Since Last Promotion, Years in Current Role, Years at Company, and Age,

Years at Company has moderate correlation with Years Since Last Promotion, Total Working Years, Monthly Income, Job Level,

Years in Current Role has moderate correlation with Years Since Last Promotion, Total Working Years,

Years Since Last Promotion has moderate correlation with Years with Current Manager, Years in Current Role, Years at Company, Total Working Years,

Years with Current Manager has moderate correlation with Years Since Last Promotion, Total Working Years.

3.5 Summary:

There are 1470 employees in the company and 16% of them left the company. We have some data about employees to examine the attrition reasons. To sum our exploratory data analysis;

Young employees (18-25 years old) compose the 8.4% of the company and they are more likely to leave the company than other age groups. **As the employees get older, their attrition percentages drop.**

Employees who **travel rarely or don't travel** according to their job description compose the 81.1% of entire company. 18.8% of the employees **travels frequently** and they have the highest attrition percentage(25%).

30.3% of employees work in the **Sales department** and they have the highest attrition percentage (20.6%). **Human Resources** employees who compose the 4.3% of company, are the second highest attrition percentage(19%). **Research and Development Department** has 65.4% employee in the company and they have the lowest attrition percentage(13.8%).

As the **distance between company and employees' homes** increases, the attrition percentage increases.

11.6% of employees has **below college degree** and highest attrition percentage(18.2%). Other attrition percentages according to education levels: 39% of employees has the **bachelor's degree** and 17.3% attrition. 19.2% of employees has the **college degree** and 15.6% attrition. 27.1% of employees has the **master's degree** and 14.6% attrition. 3.3% of employees has the **doctorate degree** and 10.4% attrition.

According to the education field, 1.8% of employees has **Human Resources** education field and highest attrition(25.9%). 8.9% employees have **Technical degree** education field and 24.2% attrition. 10.8% of employees has **Marketing** education field and 22% attrition. Life Sciences, Medical and other education fields are affected respectively.

61.2% of employees has **environment satisfaction at very high/high level** in the company. 19.3% of employees has **low environment satisfaction** and 25.3% attrition.

Male employees(60% of company) are more likely to leave the company than **female employees**.

68.8% of employees has **high/very high job involvement** in the company. As the **job involvement** increases, the attrition percentage decreases respectively.

37% of employees has **job level-1** and 26.3% attrition. **Job level-3**(14.8% of the company) and **job level-2**(36.3% of the company) are affected mostly by %14.7 and 9.7% attrition after job level-1 respectively.

5.6% of employees works as **Sales Representative** and they have the highest attrition(39.8%). **Laboratory Technician** (17.6% of employees) and **Human Resources** (3.5% of employees) job role follows the Sales Representative attrition percentage by 23.9% and 23.07% respectively. Besides those job roles, **Sale Executive** (22.8% of the employees have 17.5% attrition) and **Research Scientist** (19.9% of employees have 16.1% attrition) have the higher attrition percentage than others.

61.3% of employees has **high/very high job satisfaction** in the company. 19.7% of employees have **low job satisfaction** and highest attrition percentage(22.8%). As the job satisfaction increase, attrition percentage decreases. **Medium**(19% of employees) and **high**(30% of employees) job satisfaction has approximately same attrition percentage(16.5%).

32% of employees are **single** and has the highest attrition percentage(25.5%).

26.9% of employees have **1000-3000 dollars monthly income** and highest attrition percentage(47.7%). Other monthly income employees have been affected by attrition in the same percentage level except **9000-11000 dollars monthly income** level. That group is 5.6% of employees and has 18.6% attrition.

35.4% of employees have **one company experience** before current company, and they are more likely to leave the company(18.8%). 21% of employees **worked in 5 and more companies** before this company and have 21.8% attrition.

28.3% of employees have **over time** and 30.5% attrition in the company while rest of the employees have only 10.4% attrition.

The higher percent salary hike is, the more employees are likely and willingly to stay in the current company.

Performance rating has two category such as 3 and 4. 84.6% of employees are in the **performance rating 3 group** and 16.08 attrition. The rest of the employees are in **performance rating 4 group** and their attrition percentage(%16.37) is a little bit higher than previous group.

18.8% of employees have **low relationship satisfaction** and highest attrition percentage(20.7%). The rest of the employees' attrition percentage is at 15% band level.

43% of employees has **zero stock option level** and highest attrition percentage(24.4%). There is a sharp decrease in attrition percentage until **stock option level-3**. %5.7 of employees has stock option level-3 and 17.4% attrition.

As the **total working years** increases, the attrition percentage decreases in the relevant experience groups.

70.6% of employees have **2-3 times training in the previous year** and have 17.9% and 14.1% attrition respectively. 3.7% of employees does **not have training times in the previous year** and has the highest attrition percentage(27.8%). 8.3% of employees have **4 times training last year** and have 21.1% attrition. The rest of the employees' attrition percentage is around between 9% and 12%.

84.2% of employees have **good or better work life balance**, and 16.7% and 14.2% attrition respectively. 5.4% of employees have **bad work life balance** and the highest attrition percentage(31.3%).

As the **years at company and current role** increase, attrition percentage decreases in the company.

39.5% of employees leaves the company **before fulfilling one year since their last promotion**. 24.3% of employees **fulfilling one year since their last promotion** has the 13.7 attrition percentage. **7 years since last promotion employee group** (5.2%) has the highest attrition(21.1%) in its individual group.

Most of the employee(17.9%) quit the company **before completing their first year with their current manager**(32.3% attrition). Other group who leaves the company most is the ones who work **two years with current manager**(23.4% of employees with 14.5% attrition).

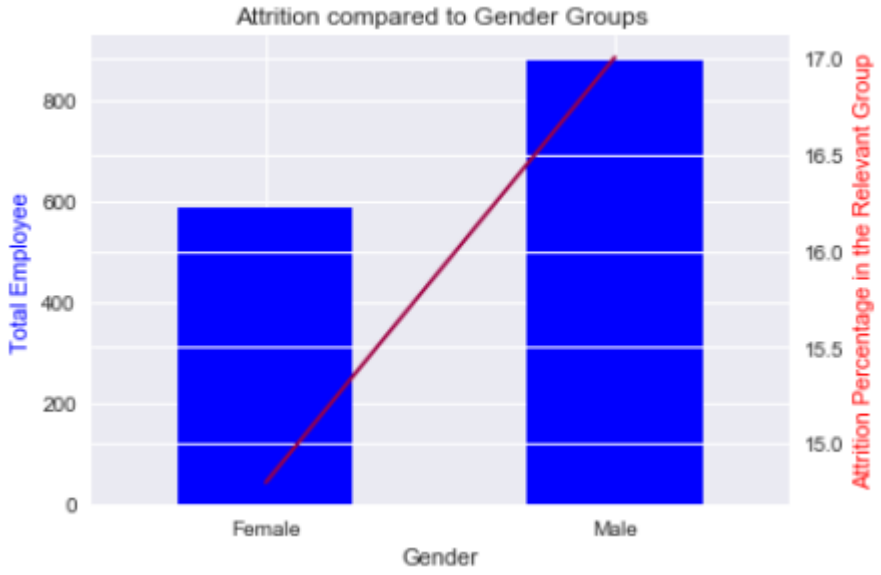
I also checked the **'Employee Number', 'Daily Rate', 'Hourly Rate' and 'Monthly Rate'** features as I did the in previous features of dataset. But there is nothing significant to comment or visualize about these features.

Feature/Variable relationships are explicitly expressed above in section 3.4.

4. Hypothesis Testing (Examining Attrition in Gender throughout the Company):

4.1 Selecting Appropriate Test:

Gender	Total Employee	Attrition Number	% of Attrition in the Relevant Group	% of Attrition in the Company
Female	588	87	14.795918	36.708861
Male	882	150	17.006803	63.291139



Since the rate of attrition for two groups is compared, a two-proportion z-test is appropriate. Central Limit Theorem (CLT) states that regardless of the population, the distribution of sample averages tends to be normal. This holds for sample sizes greater than or equal to 30. Since the size of the data is big enough and 'female' and 'male' attrition randomly to the resumes when presented to the employer, CLT could be applied here assuming that samples are representative of the population.

4.2 The Null and Alternate Hypotheses:

As observed above, it is obvious that male attrition is more than female's attrition but is the difference significant? Here is the important point to define the null and alternative hypothesis:

In the context of provided information, the alternative hypothesis assumes that male attrition is pervasive enough to be a factor in attrition throughout the company.

Null Hypothesis: There is no difference in the proportion of attrition for male and female employees in the company.

$H_o : \hat{p}_{male\_attrition} - \hat{p}_{female\_attrition} = 0$

Alternative Hypothesis: There is a significant difference in the proportion of attrition for male and female employees in the company.

$H_a : \hat{p}_{male\_attrition} - \hat{p}_{female\_attrition} \neq 0$

Significance Level: 95% Confidence:

4.3 The Frequentist Statistical Approach:

Female attrition mean:	0.14795918367346939
Male attrition mean:	0.17006802721088435
Difference of mean between male and female employees' attrition:	0.022108843537414963
z-score:	1.1292547809155016
p-value:	0.2587903704911598
Margin of Error:	0.03837338930564671
Confidence Interval:	[-0.01626455 0.06048223]

The p-value is above the significance level (0.05). So, I fail to reject the null hypothesis. We can conclude that there is no enough evidence to reject the assumption of no difference in the proportion of attrition for male and female employees in the company.

5. Machine Learning Modeling

5.1 Feature Engineering and Selection

We have explored the trends and relationships within the data, now we can work on engineering a set of features for our models. We can also use the results of the EDA to inform this feature engineering. For this dataset, we have two main things to do before running a model:

- Decide what features we should keep:

We learned the following from EDA which can help us in engineering/selecting features:

'EmployeeNumber' is evidently irrelevant features, so we can remove them. Also, the 'PerformanceRating' had no effect on attrition. As we noticed before, all employees were graded as either 3 or 4. In this analysis we noticed that turnover was practically the same in both cases.

We can drop some highly corelated features as they add redundancy to the model but since the correlation is very less in general let us keep all the features for now. In case of highly

corelated features we can use something like Principal Component Analysis(PCA) to reduce our feature space. We could probably find many other columns to drop, but we can do that later by analyzing the results from our first ML model.

- Transform categorical data into numerical:

The categorical data must be converted into numbers for the Machine Learning model to work. This can be done through sklearn (label encoding and one hot encoding) or pandas. We used pandas approach in this particular case. In our dataset, we transformed textual columns which are 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', and 'Overtime' into a numeric continuous one and encoded the categorical data into a binary representation.

5.2 Data Preprocessing

Separate response variable and features:

Firsts, we separated response variable and features as X and y.

Splitting the dataset into the Training set and Test set:

Then, we divided the dataset into the training and test sets. We have used 25% of dataset as testing set and 75% of the dataset as the training set. We also set the random state of the split to ensure consistent results.

Feature Scaling:

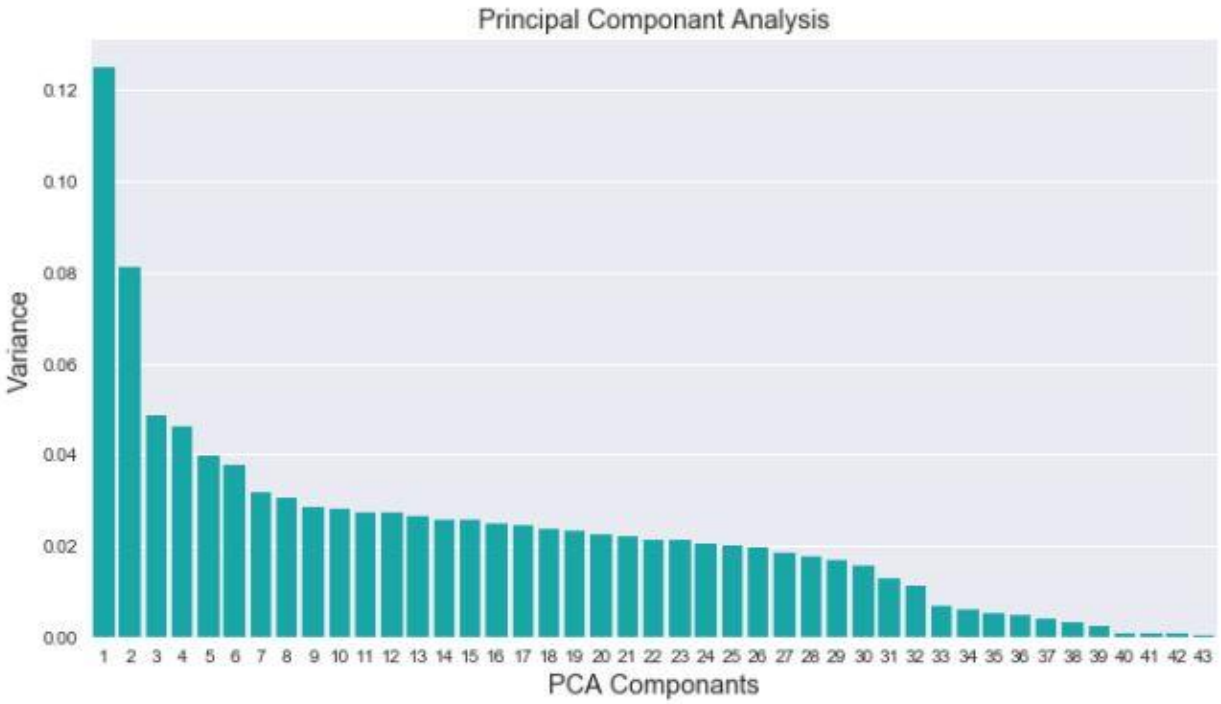
The scikit library provides various types of scalers including MinMaxScaler and the StandardScaler. In our case, we used the StandardScaler to scale the data.

Handling imbalanced dataset:

We have a imbalanced dataset with majority of observations being of one type ('No' which is '0' in the dataset) in our case. We have about 84 % of observations having 'No' and only 16 % of 'Yes' and hence this is an imbalanced dataset. To deal with such a imbalanced dataset we have to take certain measures, otherwise the performance of our model can be significantly affected. We have two approaches to curb such datasets: oversampling, which increase the number of observations corresponding to the minority class, or under-sampling which decrease the number of observations for the majority class. We have used an oversampling technique known as the SMOTE(Synthetic Minority Oversampling Technique) which randomly creates some 'synthetic' instances of the minority class so that the net observations of both the class get balanced out.

Principal Component Analysis

We have used Principal Component Analysis (PCA) which is a dimension-reduction tool and reduces a large set of variables to a small set that still contains most of the information in the dataset.



We don't see any sharp drop off in the percentage of variance explained from this screen plot, suggesting no natural cut off point in keeping certain dimensions and discarding others. Since the number of the components is not small and the linear relationships among them are not strong, it is hard to interpret the principal components.

If we look at the cumulative components scores, it shows that we may explain more than 90% of the variation with at least 27 features in our model.

5.3 Selecting the Right Evaluation Metric

As the data imbalance is emphasized above, the evaluation of the classifier performance must be carried out using adequate metrics in order to take into account the class distribution and to pay more attention to the minority class. When the positive class is smaller and the ability to detect correctly positive samples is our main focus (correct detection of negatives examples is less important to the problem) we should use precision and recall. For our particular case, based on this thought I will use f1 score which is harmonic average of precision and recall as my evaluation metric.

5.4 Applying Machine Learning Algorithms

Since we must predict a binary class, we will be using classification models for training & predicting Employee Attrition. We need to keep in mind that our focus should be to have a better accuracy of predicting attrition i.e. Attrition = 1 which in confusion matrix will be "True Positive".

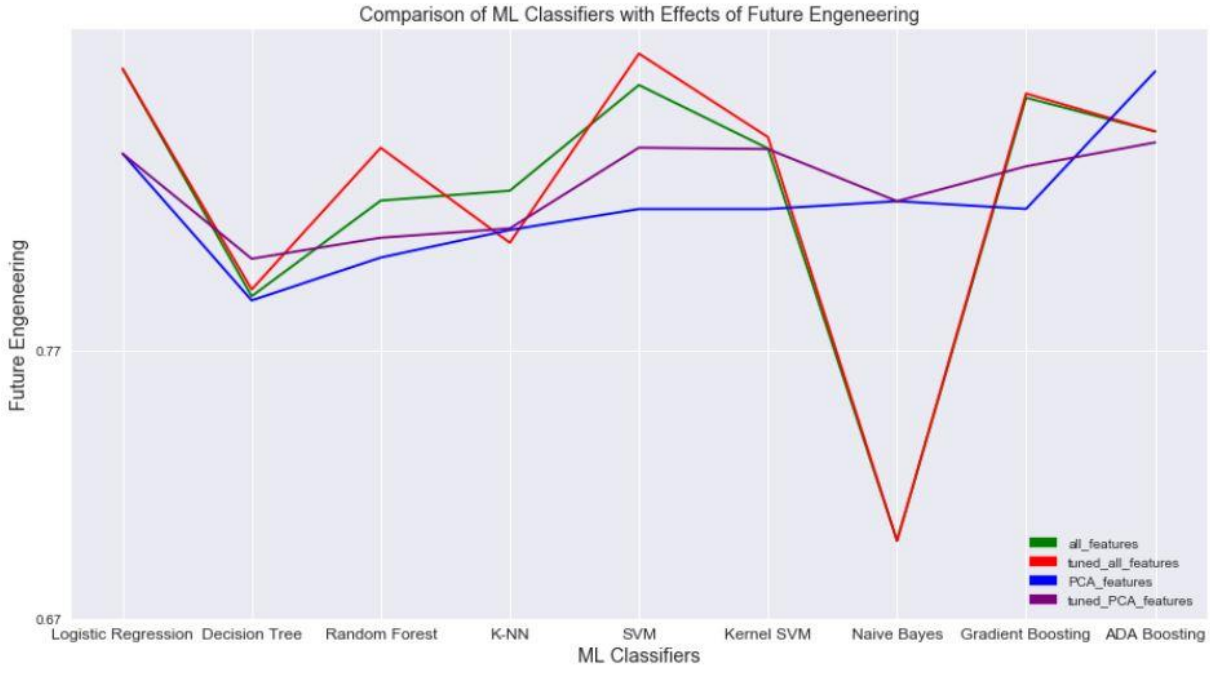
However, we should not forget the prediction accuracy of not qualifying for attrition i.e. Attrition = 0 which will be "True Negative" in confusion matrix.

In this section; Logistic Regression, Decision Tree Classification, Random Forest Classification, K-NN Classification, Support Vector Machine (SVM) Classification, Kernel Support Vector Machine (SVM) Classification, Naïve Bayes Classification, Gradient Boosting Classification, and ADA Boost Classification algorithms are applied to the dataset.

I will apply these algorithms into all features available and also features which explain the 90% of total importance via PCA. I will also try to improve model accuracy via hyperparameter tuning. I have coded four functions to apply to the models throughout this section.

5.5 Model Comparison

	all_features	tuned_all_features	PCA_features	tuned_PCA_features	Total Improvement(%)
Logistic Regression	0.875442	0.875442	0.843742	0.843742	0.0000
Decision Tree	0.790452	0.792969	0.788889	0.804466	0.0177
Random Forest	0.826190	0.845914	0.804907	0.812311	0.0239
K-NN	0.829883	0.810444	0.815210	0.815783	0.0000
SVM	0.869366	0.881111	0.823034	0.845970	0.0135
Kernel SVM	0.845668	0.849900	0.823034	0.845450	0.0050
Naive Bayes	0.699065	0.699065	0.825964	0.825964	0.1815
Gradient Boosting	0.864627	0.866205	0.823078	0.838986	0.0018
ADA Boosting	0.852034	0.852034	0.874469	0.847921	0.0263



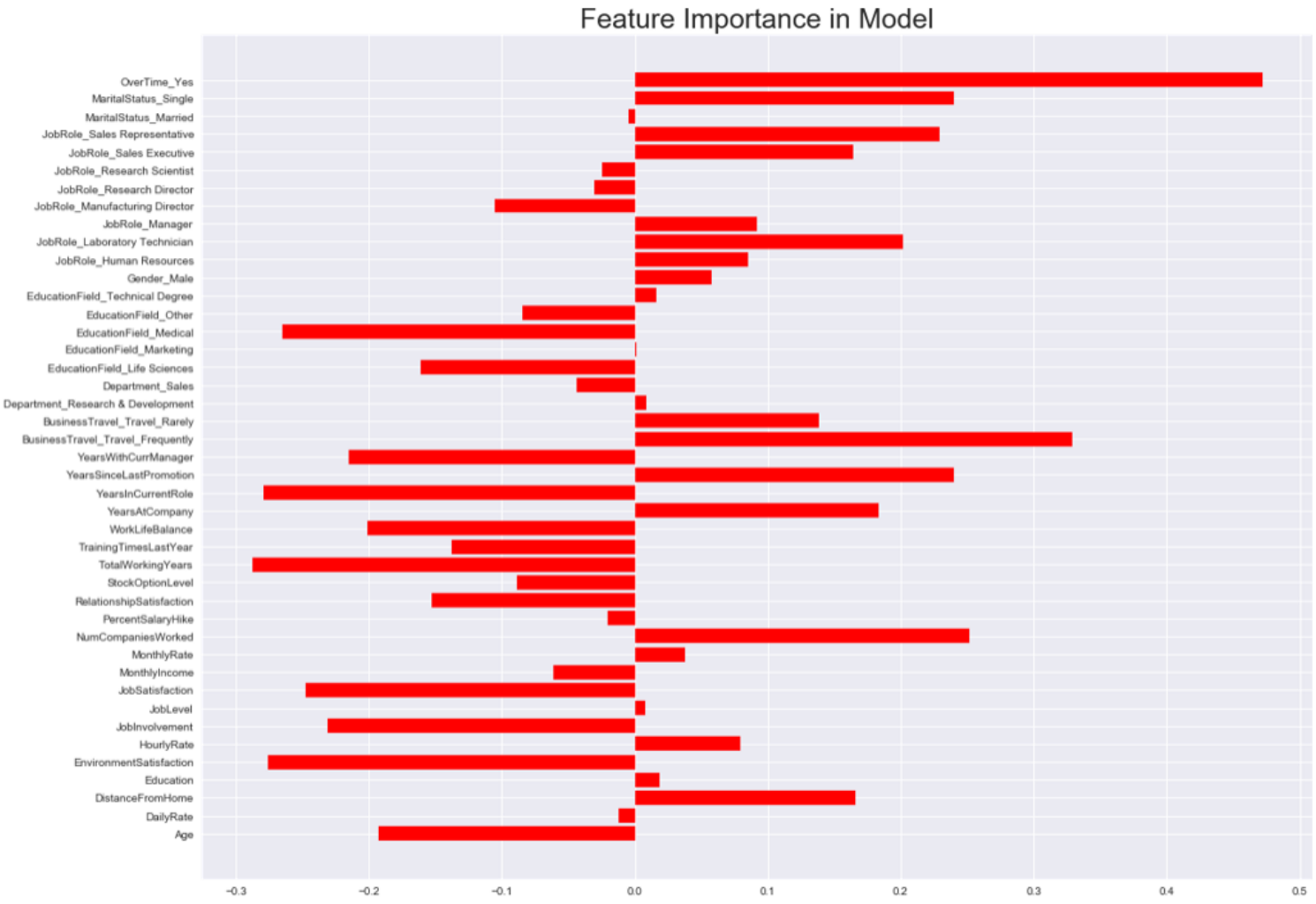
Handling Target Class Imbalance

Since we have already noted the imbalance in the values within the target variable, let us implement the SMOTE method in the dealing with this skewed value in order to see whether we may improve our accuracy score.

After SMOTE mechanism to improve target class imbalance and identifying best hyper-parameters, f1 score of our model did not show an improvement and decreased to 0.8276. Besides that, we should keep in mind that oversampling will generate artificial observations which may be tricky to evaluate the accuracy of the model.

6. Conclusion

6.1 Most important Features



In the modeling and comparison section, we have seen the hyperparameter tuned support vector machine model has the highest f1 score (0.88) with tuned model parameters C:0.1 and gamma: 0.001. Even we try principal component analysis to see the effects of less features in the modelling, it didn't improve model performance. Besides that, we use SMOTHE technique to improve accuracy metric, but it didn't contribute any increase in model performance either.

As observed in the plot of feature importance above, it seems that our SVM Classifier has decided to rank the features of OverTime highest, which is followed by BusinessTravel status. Both has positive coefficients which mean than attrition will show increase as overtime and business travel increase.

We may evaluate the most important factors which may yield to attrition via this chart above. On the right section of the chart, we see the positive features' coefficients which increase the attrition in the company. Among those; overtime, business travel, numbers of companies worked status, marital status, years since last promotion, and job role status are the ones which come forward.

To the left side of the chart, we see the negative features' coefficients which decrease attrition in the company. If the employer keeps those features stronger, it is sure to see decrease in the company. These features may be assessed as the preventive action items to keep the attrition rate low in the company. Among those; total working years, years in current role, environment satisfaction, medical education field, job satisfaction, job involvement, years with current manager, work life balance, and age are the ones which come forward.

Both those indicated positive and negative coefficients are the most important factors which the company should focus and take necessary measures to reduce the attrition.

6.2 Recommendations

1. 28.3% of employees have over time work in the company and 30.5% of those employees leave the company. As it is also reflected in the model, employees working overtime are significantly more likely to resign. Therefore, the company should understand the reason why they are working overtime. Is it for too high workload or are employees' qualifications not enough to complete the scheduled tasks on time? Maybe there might be some other reasons behind that. Our recommendation will be to understand the reason(s) for overtime with detail research and take appropriate measures to reduce the factors behind this attrition factor.

2. 18.8% of the employees travels frequently, and they have the highest attrition percentage(25%). The company should question what makes traveling a burden on their employees. The company should balance the travel status and if necessary, there might be some adjustments on the job description in terms of traveling. The company may use some extra incentives to motivate their employees who are supposed to travel.

3. 21% of employees worked in 5 and more companies before this company and have 21.8% attrition. This is an area where HR should be aware of. HR should question the employee candidate why the employee quits the previous job and get in touch with previous company to have information about the applying employees. Besides that, the company should take precautionary measurements to keep their employees in their current role after they hire new employees.

4. 32% of employees are single and has the highest attrition percentage(25.5%). The company should be aware of this important factor and have strategy to deal with this groups' performance.

5. If the year increases since the employees' last promotion, the attrition percentage also displays increase. Especially, 7 years since last promotion employee group (5.2%) has the highest attrition(21.1%) in its individual group. For that reason, the company review their promotion policy, and maybe define the company's expectations from their employees and make clear to all employees how and when they may be promoted.

6. 5.6% of employees works as Sales Representative and 17.6% of employees works as Laboratory Technician. They have 39.8% and 17.6% attrition percentage respectively. These two-job roles should be questioned, and the company should find the reason(s) why these job roles face more attrition rate than all others and take necessary actions.

7. Beside those factors above, there are some other indicators which keep employees in the company. These factors are stated below.

- Total working years in the company,

- Years in current role,

- Environment satisfaction,

- Medical education field,

- job satisfaction,

- Job involvement,

- Years with current manager,

- Work life balance,

- Age.

The company should primarily try to increase the effectiveness of those factors. As a result, it will yield to the decrease in the attrition rate.

6.3 Next Steps to Improve Model

1. In our model, we have only 1470 observations from the company, which is poor to create a strong machine learning algorithm in order to predict the attrition beforehand and act depending on this prediction. For that reason, we need to have more data from the company to improve our model's prediction accuracy. There are couple of ways to do that.

- The company should prepare a well-designed survey for the employees who leave the company. According to those survey's result, they may create other feature(s) for their data, which will enable to improve model accuracy.

- Besides that, the company should also prepare surveys for the employees who stay longer in the company since they should be aware of what makes the them to work longer in the company.

- The company should take random/immediate surveys whenever they come across a problem area in terms of human resources.

- The company apply regular surveys to see the trend in the company. For example, once in a month, twice in a year, etc.

2. We can do more permutations & combinations, feature engineering, feature selection, hyper-parameters tuning, class imbalance, etc. to improve the accuracy score. But this won't make so much difference in the accuracy score at hand now. Modelling improvement depends more on the increase in the observation quantity. Namely, the company should focus on more to collect the reasonable data from their employees.