Untitled

Jade Lai

10/5/2022

I. Introduction

Data analytics sets itself apart from data analysis by providing guidance on actions and their expected outcomes. While data analysis offers insights into the current situation, data analytics predicts future results based on specific actions. Put simply, data analysis focuses on the past and present, while data analytics is oriented towards the future.

The following case serves as an example of predicting attrition for employees within a company.

Goal

How can we identify the relationship between the features? How can we categorize our employee and take action separately for each group? How can we improve the atrition rate? All these kind of questions can be answered by using data analyttics techniques

Result

By using statistics, programming and machine learning techniques, I can find the patterns hidden in the data. With this information I can built the model to predict the attrition among employee with provided accuracy

Project duration

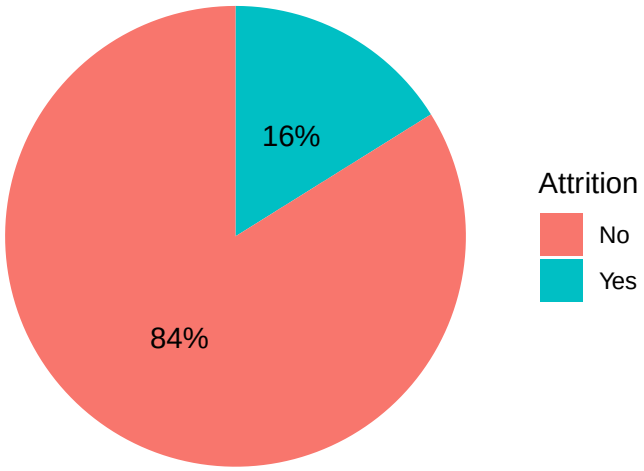
Project duration varies between 2 weeks - 2 months. The project starts by getting an understanding of the situation and gathering the right data. After collecting the right data, I start analyzing the data and share the results. Finally, I will make the algorithm or model with you can implement in your business

II. Exploration Data Analyst

The ability to predict when the employee leave the job is valuable for every business to develop the business management. Attrition rate is defined as the number of leaving employee divided by the number of current employee. In order to apply the modeling technique to predict attrition by Random Foreast athgorithm, we need to understand the employee behavior and characteristics which signal the risk of employee attrition.

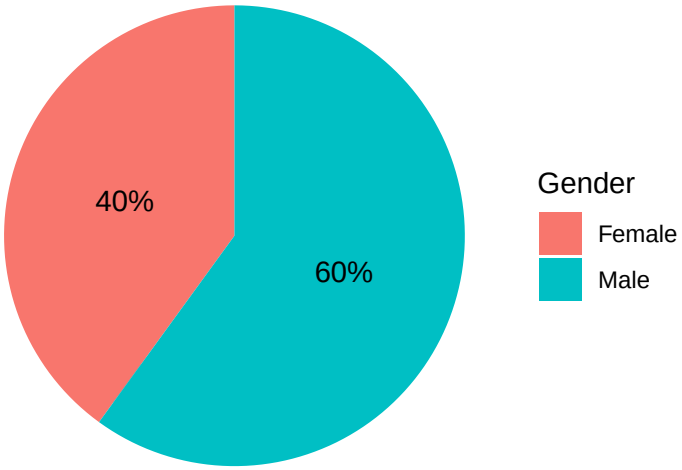
In this example, I will look into the HR employee attrition dataset.

1. What is the attrition rate?



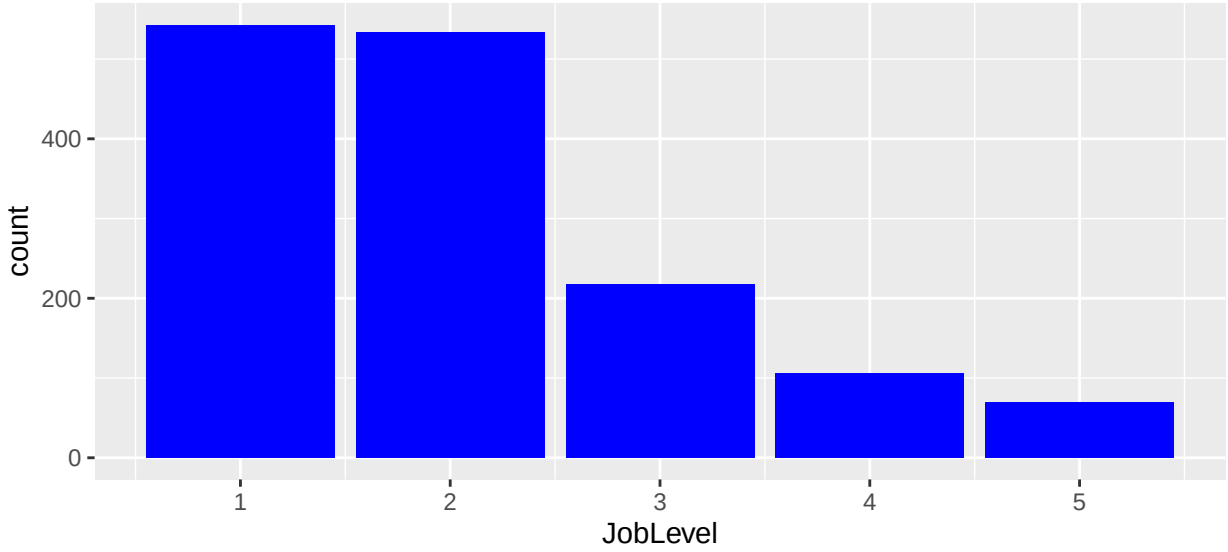
The employee attrition rate is around 16%. According to experts in the field of the human resources, the human resource consumption rate of each enterprise from 4% to 6% is a stable level. Therefore the rate of this company is at a dangerous level.The company should take action to decrease this rate.

2. What is the ratio of gender in company?



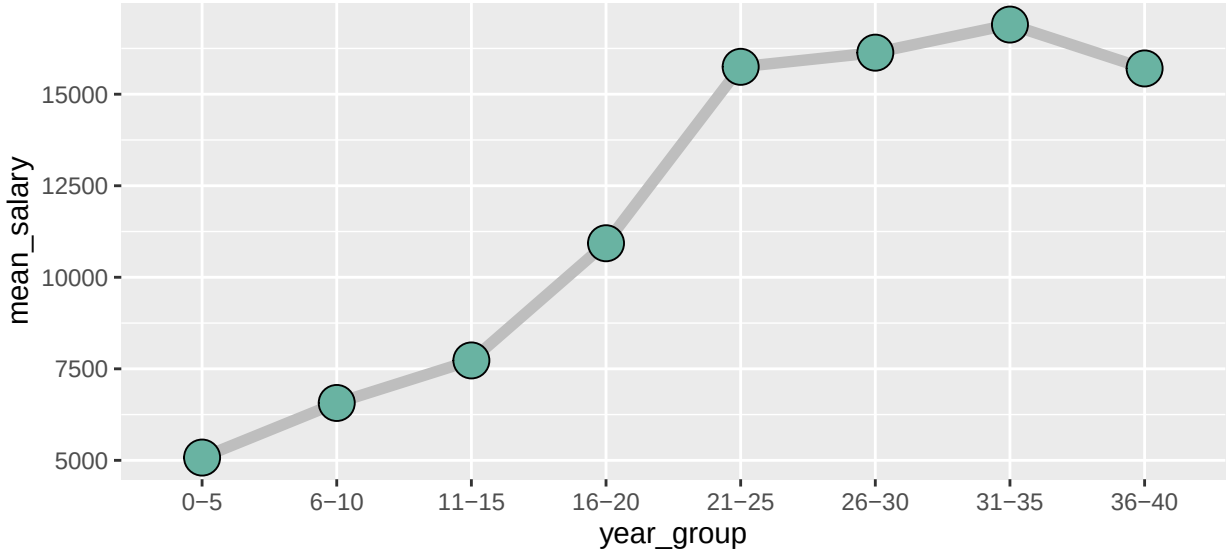
The percentage of male employees is higher than female employees (more than 20%)

3. Distribution of attrition by Job level



- Employees at Level 1 and 2 (Entry and Middle level) have a very high rate of leaving the company. They are usually very young people.
- Employees at Level 4 and 5 have a very low turnover rate.
- In conclusion, young employees who have just joined company have a very high rate to leave.

4.How do monthly salary change according the year working at company?



- Following the line graph, during the period of working with the company, the average of monthly salary increase gradually overtime. This proves that when working for a long time, the employee will have a decent salary.

- However, there is a decrease of salary for working from 35-40 but not much. The reason may be that after a long period working in this company, these employees are nearing retirement age, so their productivity falls or they have ceded senior positions for young leaders, ...

III. Chi Square test

Definition

Chi-Square test in R is a statistical method which used to determine if two categorical variables have a significant correlation between them. The two variables are selected from the same population

We use Chi-Square test to test the correlation between working over time and attrition rate. If a correlation is being found, we can plan for improving the attrition rate.

Particularly in this test, we have to check the p-values. Moreover, like all statistical tests, we assume this test as a null hypothesis and an alternate hypothesis.

The main thing is, we reject the null hypothesis if the p-value that comes out in the result is less than a predetermined significance level, which is 0.05 usually, then we reject the null hypothesis.

H0: There is no relationship between the attrition and overtime

HA: There is a significant relationship between attrition and overtime

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  attrition$Attrition and attrition$OverTime
## X-squared = 87.564, df = 1, p-value < 2.2e-16
```

We have a high chi-squared value and a p-value of less than 0.05 significance level. So we reject the null hypothesis and conclude that there is a significant relationship between attrition and overtime.

IV. Predictive model for attrition

The purpose of this analysis was to identify the effective model to predict the employee attrition. In this report, we compare between Random forest model and CART (A classification and Regression Tree) model.

The measures use in comparing two models will be: Accuracy, Kappa value.

1. Random Forest

Random Forest is a highly effective algorithm for attrition classification, which involves predicting and understanding employee turnover within an organization. By leveraging the power of ensemble learning, Random Forest combines multiple decision trees to provide accurate and reliable predictions regarding attrition. It considers various features such as employee demographics, job-related factors, performance metrics, and satisfaction surveys to identify patterns and factors contributing to attrition. The algorithm's ability to handle a large number of input variables and handle complex interactions makes it ideal for analyzing and predicting attrition.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  367  63
##        Yes   3   9
##
##                Accuracy : 0.8507
##                  95% CI : (0.814, 0.8826)
##     No Information Rate : 0.8371
##     P-Value [Acc > NIR] : 0.2416
##
##                   Kappa : 0.1759
##
##  Mcnemar's Test P-Value : 3.803e-13
##
##             Sensitivity : 0.12500
##             Specificity : 0.99189
##          Pos Pred Value : 0.75000
##          Neg Pred Value : 0.85349
##              Prevalence : 0.16290
##          Detection Rate : 0.02036
##    Detection Prevalence : 0.02715
##       Balanced Accuracy : 0.55845
##
##        'Positive' Class : Yes
##
```

2. CART Model (Classification and Regression Trees)

CART Model (Classification and Regression Trees) is a versatile algorithm used for attrition classification, aiming to predict and understand employee turnover within an organization. CART models utilize decision trees, which recursively split the data based on the most informative features, to create a hierarchical structure of rules for classification. This algorithm is particularly useful for attrition classification as it can handle both categorical and continuous variables, making it suitable for analyzing various factors influencing attrition, such as employee demographics, performance metrics, and job-related factors. By constructing a binary tree structure, CART models offer interpretable rules that help identify the key drivers of attrition. With its flexibility, simplicity, and capability to handle complex interactions, CART is an effective tool for organizations seeking to gain insights into attrition patterns and make informed decisions to mitigate employee turnover.

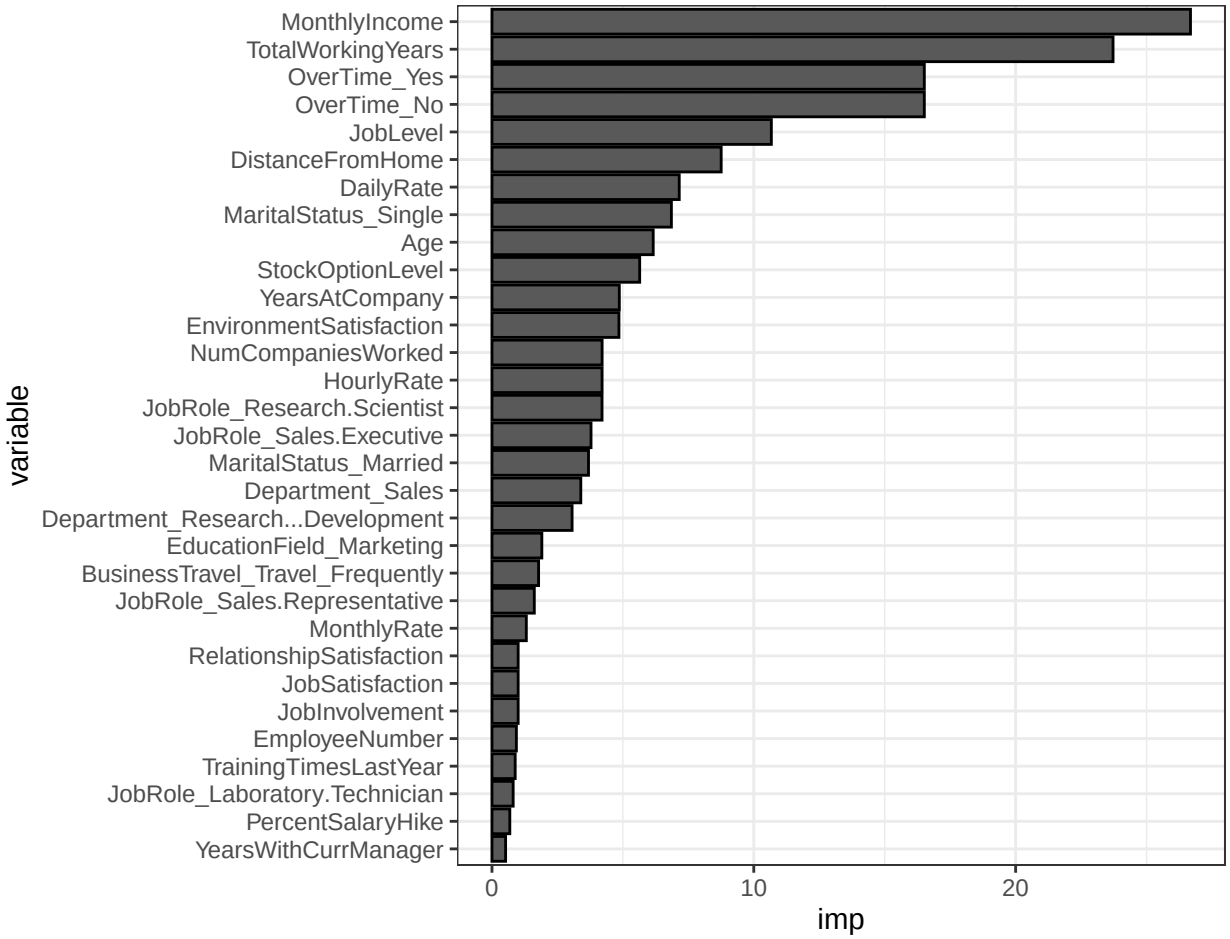
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  356  55
##        Yes  14  17
##
##                Accuracy : 0.8439
##                  95% CI : (0.8066, 0.8765)
##     No Information Rate : 0.8371
##     P-Value [Acc > NIR] : 0.3787
##
```

```
##                   Kappa : 0.2573
##
##  Mcnemar's Test P-Value : 1.469e-06
##
##             Sensitivity : 0.23611
##             Specificity : 0.96216
##          Pos Pred Value : 0.54839
##          Neg Pred Value : 0.86618
##              Prevalence : 0.16290
##          Detection Rate : 0.03846
##    Detection Prevalence : 0.07014
##       Balanced Accuracy : 0.59914
##
##        'Positive' Class : Yes
##
```

3. Comparation

After careful consideration, we would opt for the CART model over the Random Forest algorithm. Although Random Forest exhibits a high accuracy of approximately 0.85, its Kappa value is significantly low at only 0.17. In contrast, the CART model demonstrates an accuracy of about 0.84 with a comparatively higher Kappa value of 0.2573. Hence, based on both accuracy and Kappa value, the CART model emerges as the preferred choice.

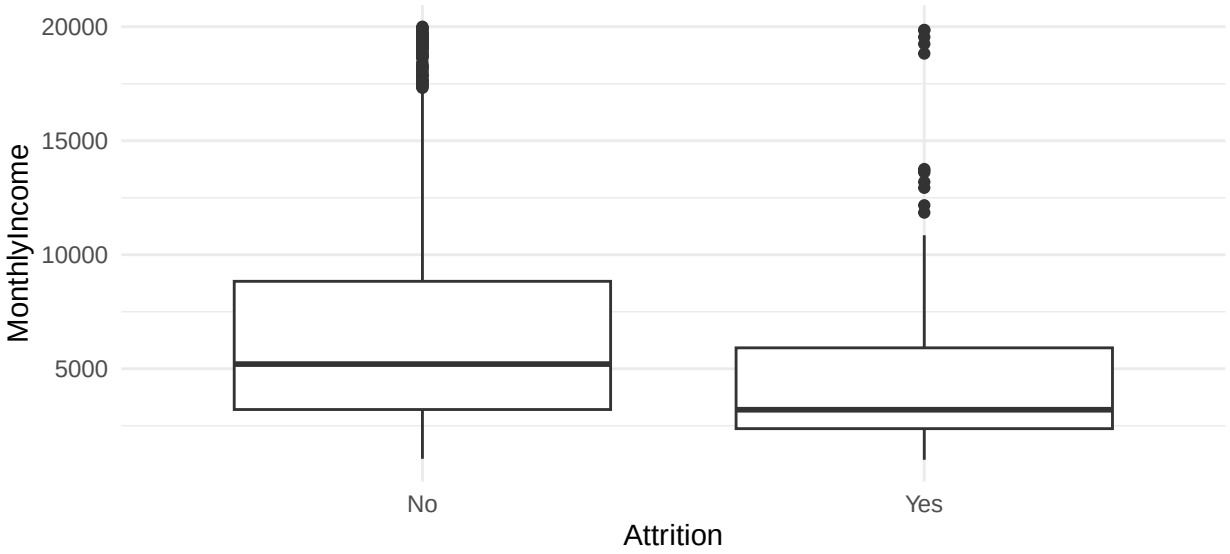
4. Feature Importance



The bar chart above shows the importance of some features such as MonthlyIncome, TotalWorking Years and Overtime variables. Combining with the statistic test and descriptive analysis above, we deeply dive in the variable and suggest recommendation for business

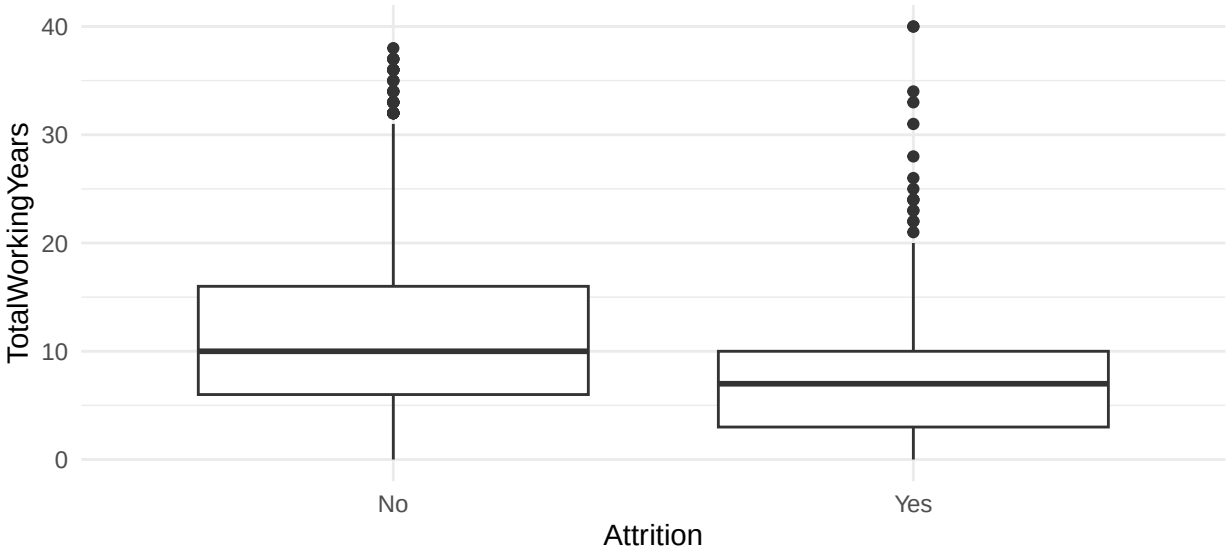
V. Recommendation

1. Monthly income



The data suggests that leaving people have lower average monthly income. This make sense because money is the big motivation, and employees may receive offers with better pay from other organization. Therefore, if the company want to try to keep employees for a longer period, it might be a good idea to offer them a raise.

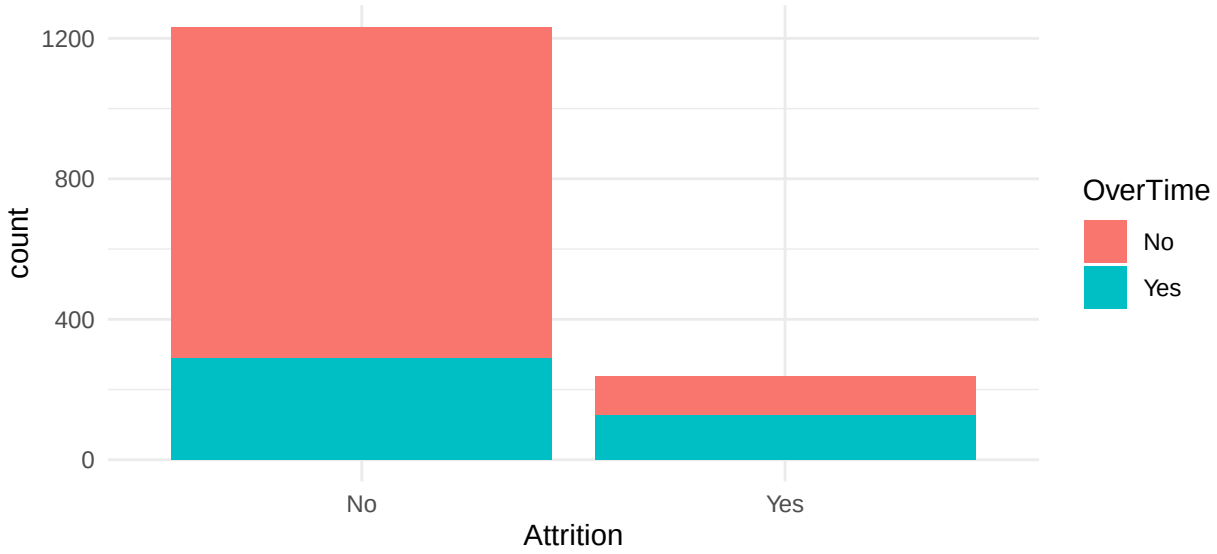
2. Total Years of Working



The data displays employees with more extensive work history correlates with lower attrition. It could be due to various reasons. Maybe because experienced employees are already established their career, and want to settle down or it could be some reasons like: having higher income or becoming a senior role as a result

of their working history. Therefore, organization should consider working experience in deciding who to hire over other factors such as education, skills, and core competencies.

Overtime



It is obvious that the Overtime is the second greatest predictor of employee attrition. This might be because of a lack of work and life balance, or various factor. Therefore, the company could try to reduce the amount of overtime their employee worked.