Wharton Analytics Fellow Data Challenge

By Thanyaphorn Thangthanakul

IBM HR Analytics Employee Attrition & Performance

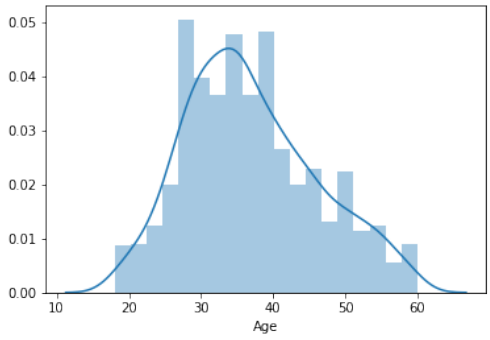
EDA: Who are the employees?

The given data has 1470 rows for each employee and 35 columns describing their attributes, which includes age, gender, department, education, job level, job involvement, etc.

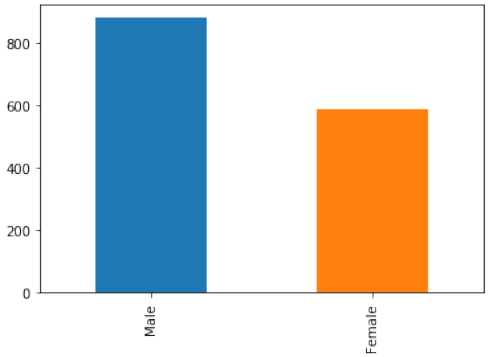
The categorical attributes include: 'Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18', 'OverTime'

And the numberical attributes include: 'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'

Summary information about the employees:



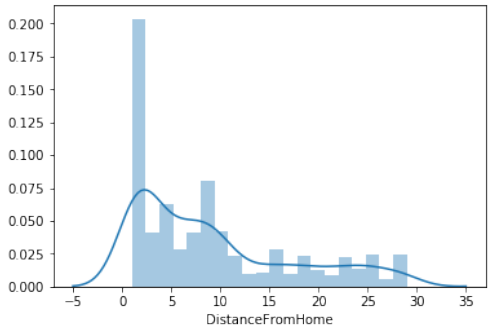
Distribution of employees' age

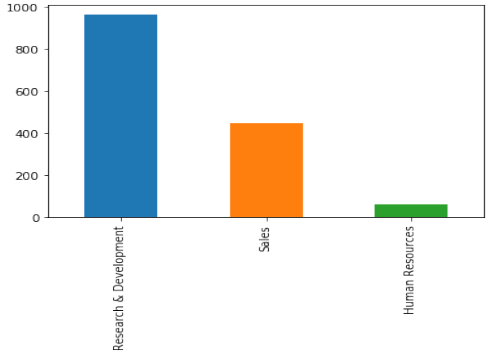


Distribution of employees' gender

The age of employees is close to normally distributed with median age around 35. Most of the employees are in the age range from 25 to 45.

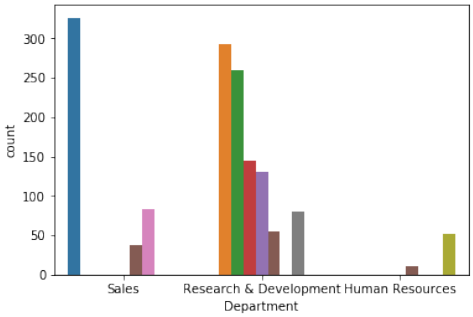
There are more male than female employees, with 60% male and 40% female.





Distribution of employees' distance from home *Distribution of employees' department*

Many employees live very close to their workplace, with around 28% living 1-2 distance units from home and the majority of the employees living within 10 distance units.

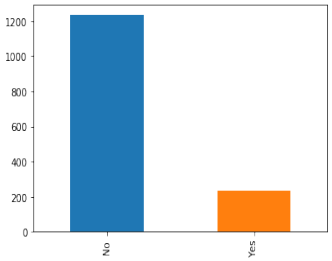


Distribution of employees' job roles within the department

The majority of employees, 65.4%, work in the Research & Development department. Followed by the Sales department at 30.3%. Then, the HR department has 4.3% of employees. For job roles, sales executive is the most popular role and is the majority of role in Sales department. The second and third most popular role are research scientist and laboratory technician within the R&D department.

Employees and Attrition

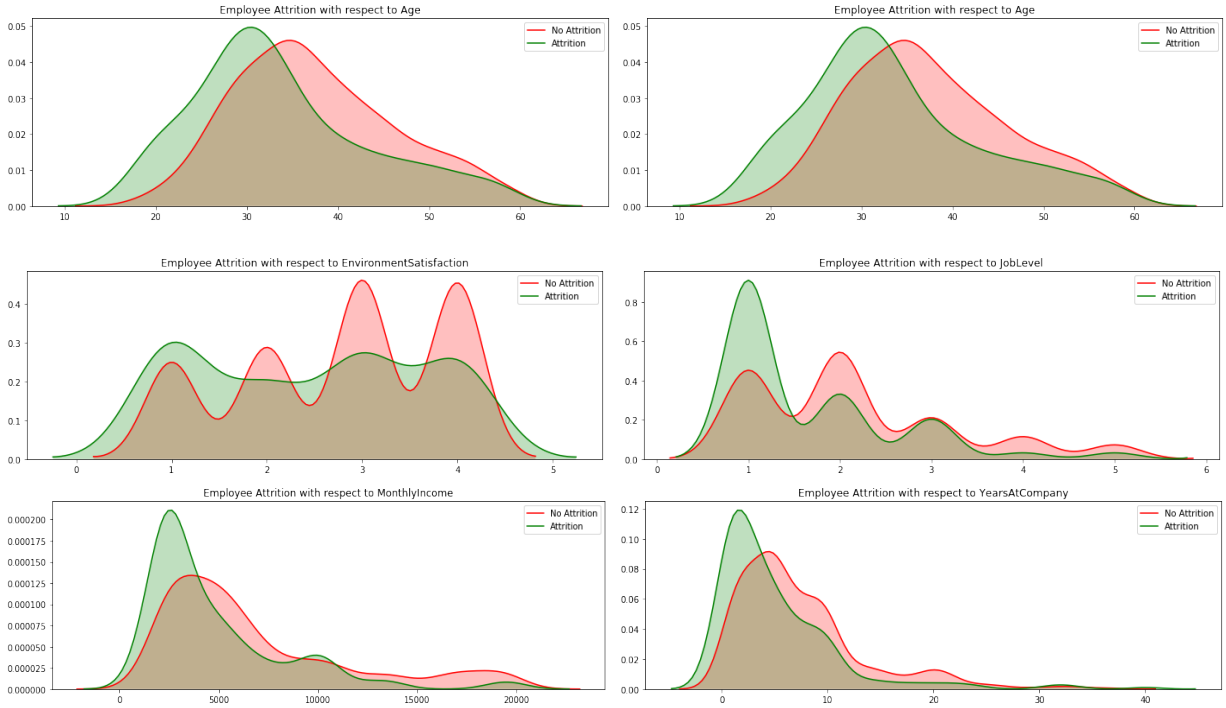
These data contains employees who left the company and also those who stay. Around 83.8% of employees left the company, while 16.1% stay with IBM. We can see that the dataset is imbalanced. Employee attrition can be related to many factors. I will explore some of them with significant implications here.



Distribution of Employee Attrition

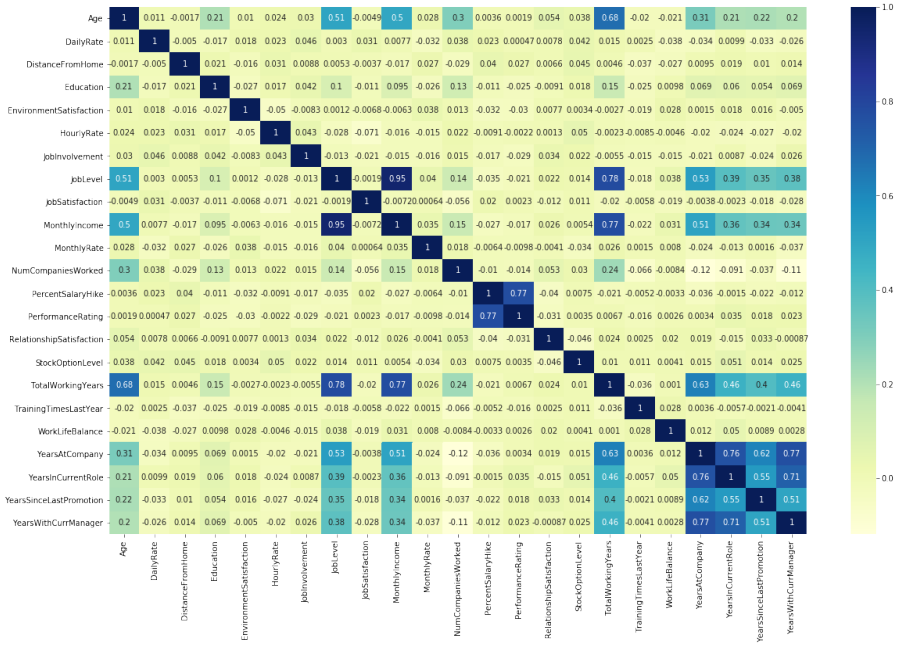
Attrition with respect to other attributes:

Some observations are that attrition rate is higher for younger employees. Surprisingly, employees with low environment satisfaction still stays with the job. Employees with low job level, low income, and less number of years at the company tend to have high attrition.



Models to Forecast Attrition

First, I made Attrition numerical by replacing Yes with 1 and No with 0. From looking into the data, I could see that EmployeeCount and StandardHours are the same across all the employees, and EmployeeNumber is just the order of the employees. These columns should not be used in the model. Then, I plotted the heat map to see the correlation between the columns.



There is very high correlation between JobLevel & MonthlyIncome, at 95%. There are also significant correlation between PerformanceRating & PercentSalaryHike, JobLevel & TotalWorkingYears, YearsWithCurrentManager & YearsAtCompany. All these correlations make sense logically. Still, to reduce collinearity, I choose to only exclude JobLevel from the model because I believe it is represented in MonthlyIncome, which represents more information. To fit a model, I do one-hot encoding for the categorical variables. I first tried to fit a simple logistic regression to the training data.

```
----- Simple Logistic Regression -----
Accuracy:  0.8571428571428571
ROC AUC Score:   0.7472527472527473
Confusion Matrix:
 [[239  34]
 [  8  13]]
-----Classification Report-----
              precision    recall  f1-score   support

           0       0.97      0.88      0.92       273
           1       0.28      0.62      0.38        21

    accuracy                           0.86       294
   macro avg       0.62      0.75      0.65       294
weighted avg       0.92      0.86      0.88       294
```

Simple Logistic Regression Model Summary

The model has a high accuracy of 85%, but consider that the precision for Attrition 1 is very low and the recall is quite low as well, and this is also reflected in the low f1-score. As I mentioned above, the dataset is imbalanced as there is 83.8% of 0 and 16.1% of 1 for Attrition. It is important to note that the high accuracy may be very misleading. If the model just always predict 0s, then the accuracy of this model is already 83.8%. So, I look into the confusion matrix and calculate precision and recall for each model.

Resampling is a common way to tackle the problem. Here, I choose to use over-sampling, which is adding instances from the under-represented class (1) sampled with replacement. I think over-sampling is better for this case because I do not have a lot of data and would like to keep all the useful information in the instances with Attrition 0s. I fitted the over-sampled data using 3 models: Logistic Regression, Decision Tree, and Random Forest.

```
----- Logistic Regression with over-sampling -----
Accuracy:  0.8380566801619433
ROC AUC Score:   0.8386116994204469
Confusion Matrix:
 [[202  35]
 [ 45 212]]
-----Classification Report-----
              precision    recall  f1-score   support

           0       0.82      0.85      0.83       237
           1       0.86      0.82      0.84       257

    accuracy                           0.84       494
   macro avg       0.84      0.84      0.84       494
weighted avg       0.84      0.84      0.84       494
```

Logistic Regression with Over-Sampling Summary

```
----- Decision Tree with over-sampling -----
Accuracy:  0.8461538461538461
ROC AUC Score:   0.8467221592868049
Confusion Matrix:
 [[204  33]
 [ 43 214]]
-----Classification Report-----
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       237
           1       0.87      0.83      0.85       257

    accuracy                           0.85       494
   macro avg       0.85      0.85      0.85       494
weighted avg       0.85      0.85      0.85       494
```

Decision Tree with Over-Sampling Summary

```
----- Random Forest with over-sampling -----
Accuracy:  0.9311740890688259
ROC AUC Score:   0.931881987883564
Confusion Matrix:
 [[235  22]
 [ 12 225]]
-----Classification Report-----
              precision    recall  f1-score   support

           0       0.95      0.91      0.93       257
           1       0.91      0.95      0.93       237

    accuracy                           0.93       494
   macro avg       0.93      0.93      0.93       494
weighted avg       0.93      0.93      0.93       494
```

Random Forest with Over-Sampling Summary

The over-sampled logistic regression model has a slightly lower accuracy than the simple model, but it has a significantly higher precision and recall for Attrition of 1. The decision tree and random forest models give higher accuracy, precision, and recall, with the random forest model giving the best scores.

Furthermore, I analyzed the coefficients from the logistic regression model and their significance. Some significant negative coefficients are education field, being married, job involvement, etc. And some positive coefficients are doing overtimes, age, number of companies they've worked at, performance rating, and percent of salary hike.

The HR team could use these models to predict attrition of the employees, and devise future plans to hire more people in certain roles. Or, the team can learn useful information from the characteristics of employees with attrition and the significance of each factor in determining the attrition in the model, so they could develop plans to incentivize employees in the aspects that would make them stay with IBM or know what to look for in hiring people for certain roles.