BT2101 Final Project

Names & Email:

1) Wong Yong De Victor (e0406670@u.nus.edu)
2) Ching Zheng Ing (e0415812@u.nus.edu)
3) Leung Hoi Kit Alvin (e0412893@u.nus.edu)
4) Darrell Lai Wei An (e0415893@u.nus.edu)

Introduction of Dataset:

Our data set is obtained from Kaggle. It contains data on IBM HR Analytics Employee Attrition & Performance.

Motivation:

Our underlying motivation is to uncover factors that lead to employee attrition. To accomplish this, we will be analyzing factors that trigger attrition and discuss if they have a significant effect through data analysis and visualisation. Finally, we will predict attrition in employees by using our selected features.

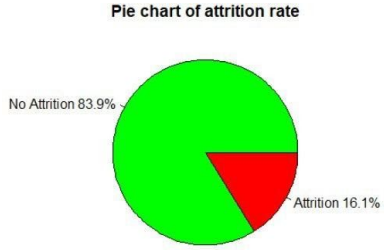
Description:

The data comprises 1470 observations of employee data and 35 types of employee characteristics. We applied head() and summary() functions to observe the data and return the mean, count, minimum, maximum values and quantiles of data. There are no null values.

```
> summary(data) #min, max, quantiles of data, mean, count of data
      Age        Attrition           BusinessTravel   DailyRate                  Department  DistanceFromHome   Education            EducationField EmployeeCount
 Min.   :18.00   No :1233   Non-Travel        : 150   Min.   : 102.0   Human Resources       : 63   Min.   : 1.000   Min.   :1.000   Human Resources : 27    Min.   :1
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0   Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000   Life Sciences   :606    1st Qu.:1
 Median :36.00              Travel_Rarely    :1043   Median : 802.0   Sales                 :446   Median : 7.000   Median :3.000   Marketing       :159    Median :1
 Mean   :36.92                                      Mean   : 802.5                                  Mean   : 9.193   Mean   :2.913   Medical         :464    Mean   :1
 3rd Qu.:43.00                                      3rd Qu.:1157.0                                  3rd Qu.:14.000   3rd Qu.:4.000   Other           : 82    3rd Qu.:1
 Max.   :60.00                                      Max.   :1499.0                                  Max.   :29.000   Max.   :5.000   Technical Degree:132    Max.   :1

 EmployeeNumber   EnvironmentSatisfaction    Gender      HourlyRate     JobInvolvement    JobLevel                      JobRole    JobSatisfaction  MaritalStatus
 Min.   :   1.0   Min.   :1.000          Female:588   Min.   : 30.00   Min.   :1.00   Min.   :1.000   Sales Executive          :326   Min.   :1.000   Divorced:327
 1st Qu.: 491.2   1st Qu.:2.000          Male  :882   1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Research Scientist       :292   1st Qu.:2.000   Married :673
 Median :1020.5   Median :3.000                       Median : 66.00   Median :3.00   Median :2.000   Laboratory Technician    :259   Median :3.000   Single  :470
 Mean   :1024.9   Mean   :2.722                       Mean   : 65.89   Mean   :2.73   Mean   :2.064   Manufacturing Director   :145   Mean   :2.729
 3rd Qu.:1555.8   3rd Qu.:4.000                       3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000   Healthcare Representative:131   3rd Qu.:4.000
 Max.   :2068.0   Max.   :4.000                       Max.   :100.00   Max.   :4.00   Max.   :5.000   Manager                  :102   Max.   :4.000
                                                                                                     (Other)                  :215

 MonthlyIncome    MonthlyRate    NumCompaniesWorked Over18   OverTime   PercentSalaryHike PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
 Min.   : 1009   Min.   : 2094   Min.   :0.000      Y:1470   No :1054   Min.   :11.00     Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
 1st Qu.: 2911   1st Qu.: 8047   1st Qu.:1.000               Yes: 416   1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
 Median : 4919   Median :14236   Median :2.000                          Median :14.00     Median :3.000     Median :3.000            Median :80    Median :1.0000
 Mean   : 6503   Mean   :14313   Mean   :2.693                          Mean   :15.21     Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
 3rd Qu.: 8379   3rd Qu.:20462   3rd Qu.:4.000                          3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
 Max.   :19999   Max.   :26999   Max.   :9.000                          Max.   :25.00     Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000

 TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000   Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median :10.00     Median :3.000         Median :3.000   Median : 5.000   Median : 3.000     Median : 1.000          Median : 3.000
 Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008   Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000   Max.   :18.000     Max.   :15.000          Max.   :17.000
```

In the dataset, the dependent variable is a binary variable, and the 35 independent variables are a mix of categorical, binary and continuous types.

We also noted that the distribution of our dependent variable, Employee Attrition, is extremely imbalanced with 83.9% staying and 16.1% leaving.

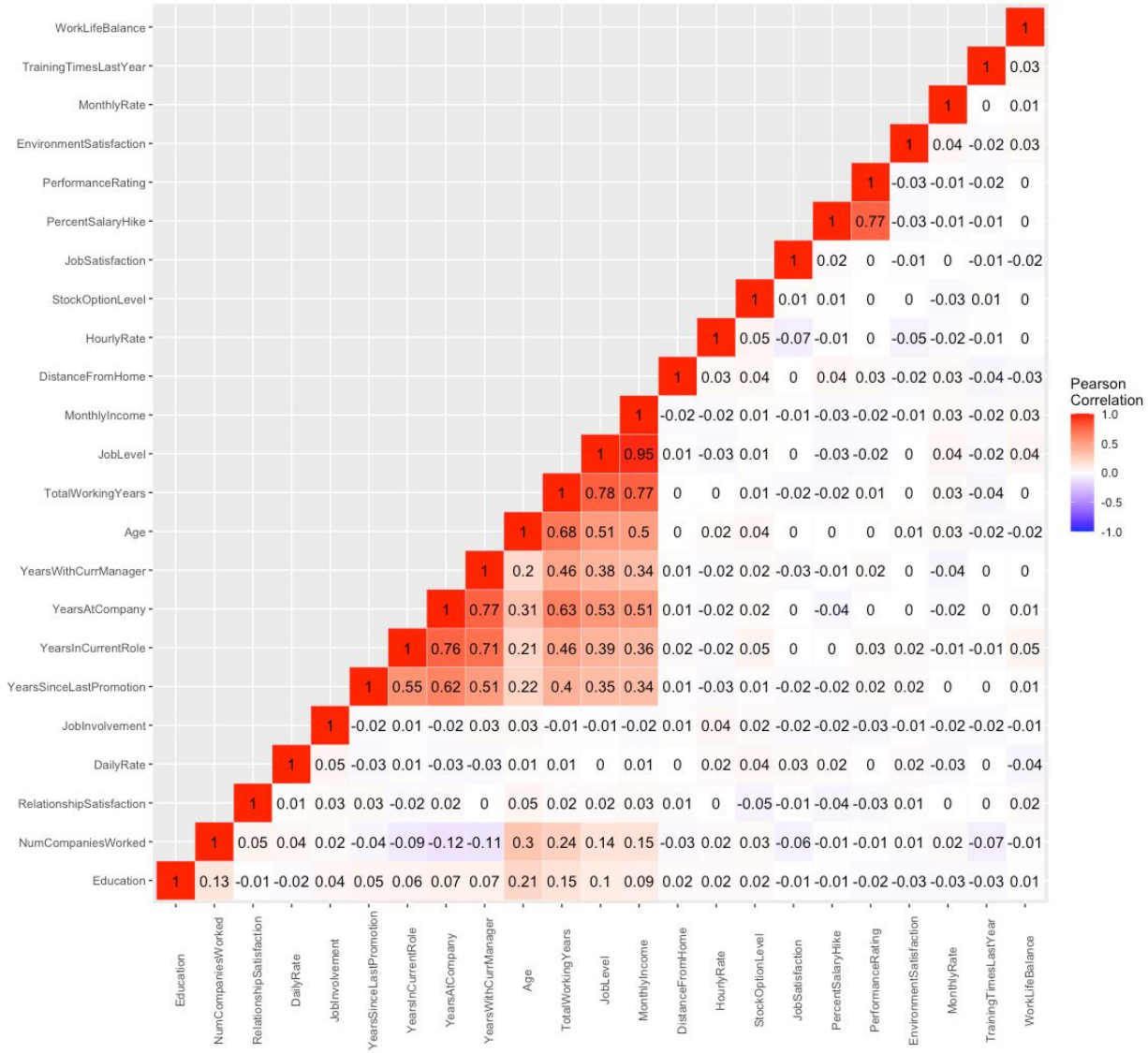


Exploratory Data Analysis (EDA):

To better understand the dataset and create accurate Machine Learning models, we want to examine:

1. Whether there are independent variables that are highly correlated (too similar and are redundant).
2. Whether we require all variables (simplifying the model).
3. Can we combine independent variables or transform them to better represent the data.

To determine which variables we should focus our exploratory data analysis on, we decided to first create a correlation matrix.



From the correlation matrix, we can see that there are 10 relationships between independent variables that are highly correlated (Pearson Correlation Value, r > 0.6).

1. JobLevel and MonthlyIncome (0.95)
2. TotalWorkingYears and JobLevel (0.78)
3. YearsAtCompany and YearsWithCurrManager (0.77)
4. TotalWorkingYears and MonthlyIncome (0.77)
5. PercentSalaryHike and PerformanceRating (0.77)
6. YearsInCurrentRole and YearsAtCompany (0.76)
7. YearsInCurrentRole and YearsWithCurrManager (0.71)
8. Age and TotalWorkingYears (0.68)
9. YearsAtCompany and TotalWorkingYears (0.63)
10. YearsSinceLastPromotion and YearsAtCompany (0.62)

Due to the high correlation between these independent variables, we need to remove one variable from each pair to prevent multicollinearity. We decided to remove MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager and PercentSalaryHike for the following reasons:

MonthlyIncome: MonthlyIncome will be high for employees that also have a high JobLevel. A higher JobLevel will also mean that the employee is paid a higher wage for their work. Therefore, the information gained regarding an employee's MonthlyIncome can be explained by their JobLevel.

TotalWorkingYears: Employees that have been working at the company for a long period of time will also have a high Age and JobLevel. To attain a higher JobLevel, the employee will have to work in the company for a longer period of time to gain the required experience. Similarly, an employee with a higher Age will have worked in the company for a longer period of time. Therefore, the information gained regarding an employee's TotalWorkingYears can be explained by Age and JobLevel.

PercentSalaryHike: Employees that have a high PerformanceRating will also have a high PercentSalaryHike. When an employee is given a high PerformanceRating, they will be awarded a high PercentSalaryHike to reward them for their good work. Therefore, the information gained regarding an employee's PercentSalaryHike can be explained by PerformanceRating.

YearsAtCompany and **YearsWithCurrManager:** YearsAtCompany has a high correlation with 4 other variables: YearsWithCurrManager, YearsInCurrentRole, YearsSinceLastPromotion and TotalWorkingYears. An employee that stays at the company for a longer period of time will be more willing to wait for each subsequent promotion for a longer period of time, and thus stay longer in each role. Therefore, as an employee rises through the ranks of the company, they acquire a more specialized skill set and oversee a larger team. Since there is a smaller talent pool at higher levels, this often means that senior employees tend to remain in the same position for longer periods of time than junior employees. Therefore, we have decided to remove YearsAtCompany and YearsWithCurrManager as they are closely tied to these other variables.

Other variables:

In addition, DailyRate, HourlyRate, MonthlyRate and MonthlyIncome all represent the remuneration amount an employee receives for their work, hence we can simplify our data by only choosing to keep MonthlyRate. Similarly, JobRole and JobLevel represent the seniority of the employee, hence to simplify our data, we will just choose to keep JobLevel.

Also, EmployeeCount, EmployeeNumber, Over18 and StandardHours are redundant independent variables throughout all 1470 employee data observations (they have the exact same value or are just an unique identification key), thus since there is no additional information gained across observations, we will remove all of them.

Therefore, for our subsequent analysis, we will be removing MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager, PercentSalaryHike, DailyRate, HourlyRate, JobRole, EmployeeCount, EmployeeNumber, Over18 and StandardHours.

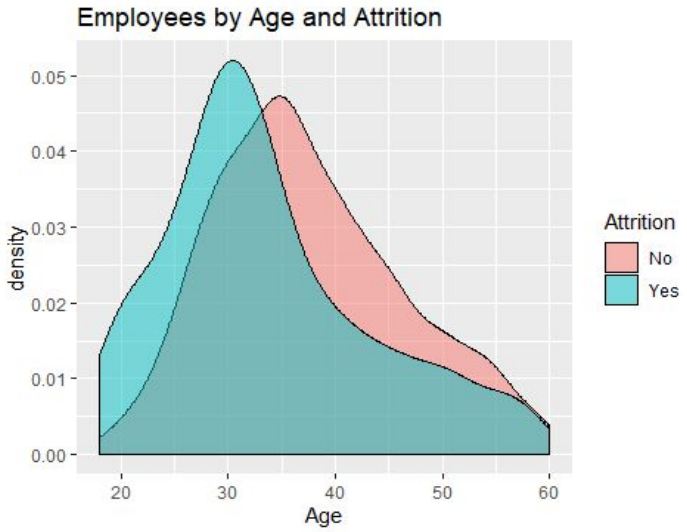
Focusing on:

For the remaining variables, we will focus our data exploration on Age, YearsSinceLastPromotion, YearsInCurrentRole, JobSatisfcation, JobHop and JobLevel to determine if any inferences can be made regarding:

1. Effects of Age on Attrition
2. Employee Frustration (YearsSinceLastPromotion, YearsInCurrentRole, JobSatisfaction)
3. Propensity of Employee to Job Hop
4. Effects of JobLevel on Attrition

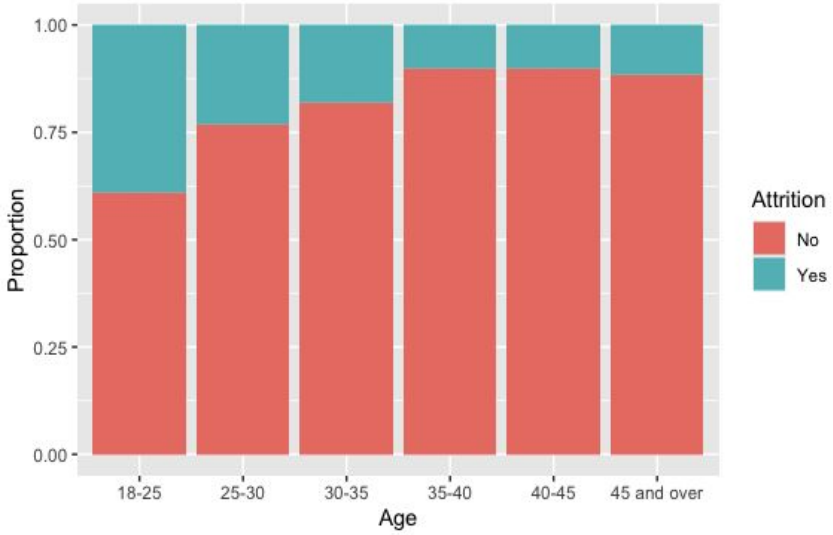
Effects of Age on Attrition:

To better understand the effects of age on employees who attrite, we analyse the age breakdown of both attrited and un-attrited employees.



From the graph above, we can clearly see that there is a difference in attrition rate across different ages. The largest contributors to attrition are in their early-thirties while the bulk of the non-attrited employees comes from people in their mid-thirties. With this knowledge, companies should work on how to engage their employees that are in their early-thirties before they attrite.

However, because the above is a density analysis, we cannot accurately identify the attrition rates of individuals from different age groups. Therefore, we split the employees into different age groups and calculated their respective proportion of attrition.

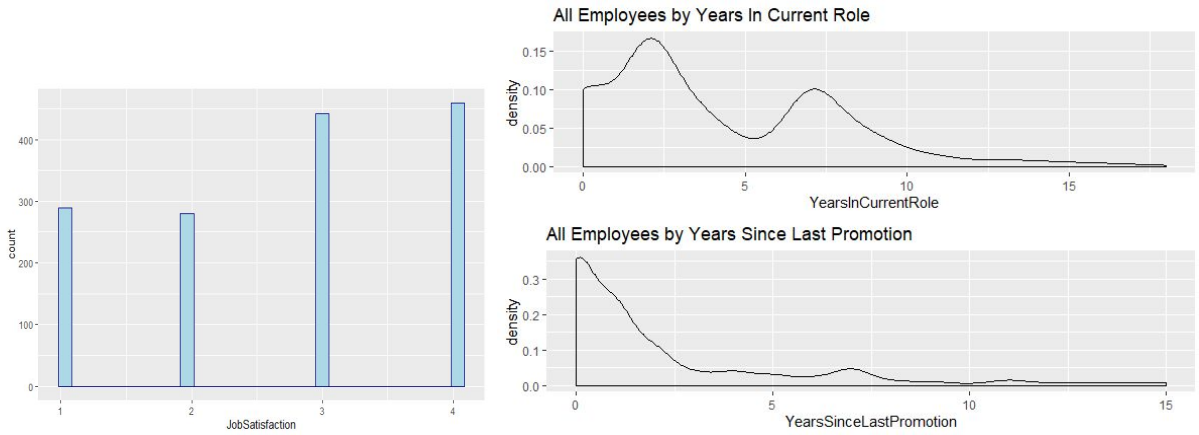


From the graph on the left, we observe that the proportion of attrition is higher for employees in the smaller age groups and are lower for employees in higher age groups. This strengthens our initial observation that there is a difference in attrition rate across different ages.

However, a disadvantage of this visualisation is that it does not account for factors that could have changed over time, that also affected the attrition rate. For example, a major overhaul of management or shift in company culture could affect the motivation levels of the employees, or perhaps affect the employees' work-life balance. Thus, employees of different ages, who have been at the company for different lengths of time could have gone through very different experiences and been exposed to a very different style of work. This difference in experience could be an omitted variable that influences the attrition rate, and is also correlated with an employee's Age.

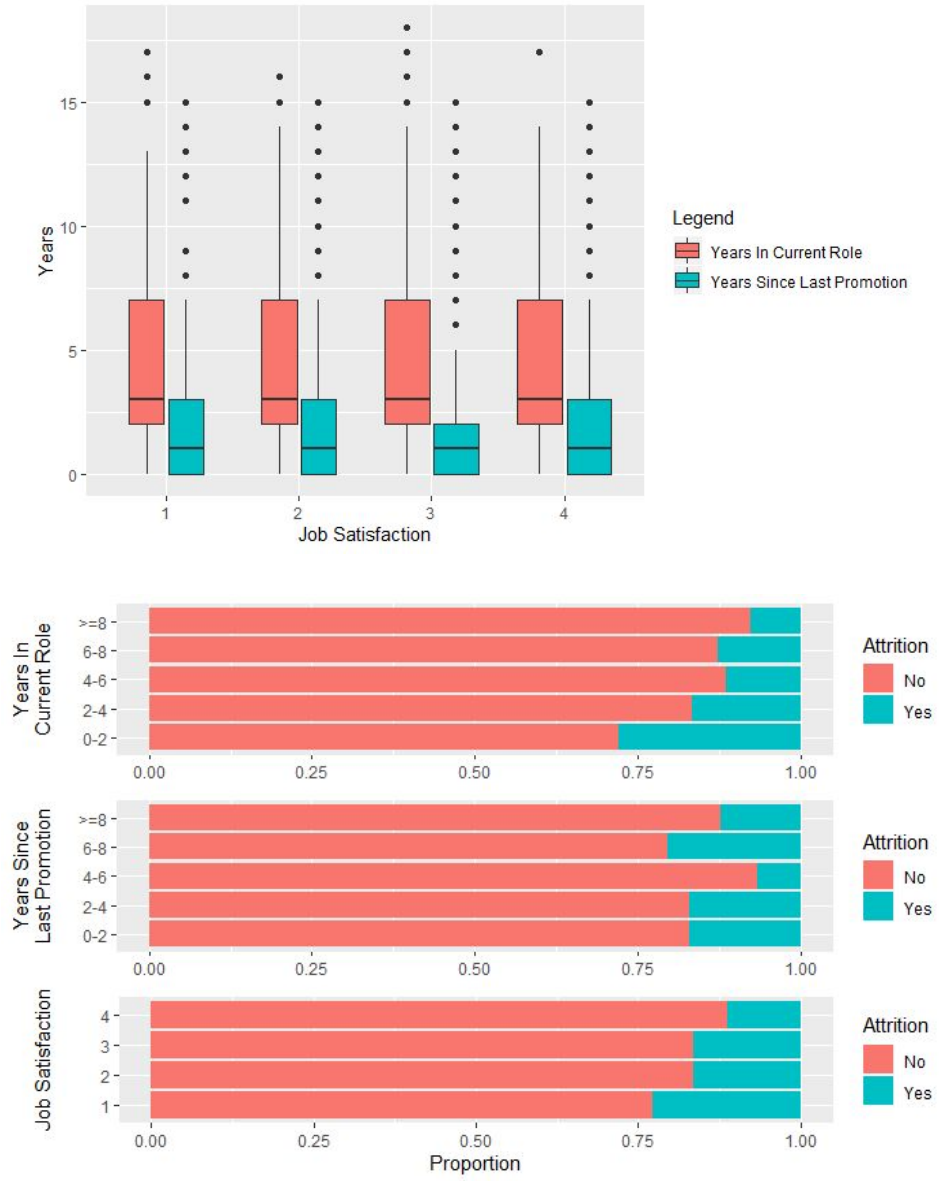
Employee Frustration (YearsSinceLastPromotion, YearsInCurrentRole, JobSatisfaction):

In our data set, 62.6% of employees have Years In Current Role exceeding their Years Since Last Promotion, 10.8% have Years In Current Role smaller than Years Since Last Promotion, and 26.6% have spent an equal number in either. This shows that most employees spend more time between roles than between promotions. In other words, it is more common for employees to get promoted while remaining in the same role, than to change roles without getting promoted. A question we asked was, could employee attrition be due to frustration at slow career advancement or lack of variety in their jobs? And between the two, which has a larger effect on attrition rates? Employees who have had promotions few and far between, or have been stuck working in the same role for many years, could be demoralised and disillusioned with their careers and the company.



To investigate this claim, we will look in more detail at the following variables, YearsInCurrentRole, YearsSinceLastPromotion and JobSatisfaction. From the histogram, we can see that there are more employees with higher levels of JobSatisfaction (3 and 4) than with lower levels (1 and 2). For Years In Current Role, there seems to be two peaks, with most people having stayed in their current roles for around 2 years and 7 years. As for Years Since Last Promotion, however, there is an overall decreasing trend.

Now, we take a look at attrition as a proportion across the different levels of these three variables. We also check if the distribution of Years Since Last Promotion and Years In Current Role is heterogeneous across different levels of satisfaction by plotting a grouped boxplot. However, there seems to be little difference.

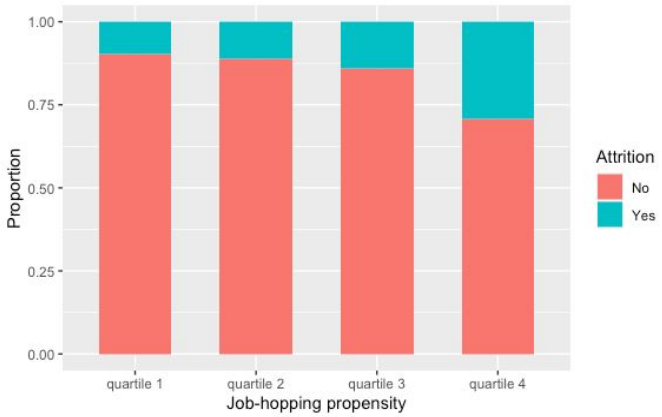


From the bar graphs, it seems that employees who have spent fewer years in their current roles are actually *even more likely* to attrite than those who have spent more years. For those who have spent up to 2 years in their current role, more than 25% end up quitting. This defies our expectations that frustration with lack of variety in their jobs is leading to attrition. For Years Since Last Promotion, there seems to be no clear trend, and it cannot be said that this has any bearing on Attrition Rate at all. The only of these three variables that seems to have clear correlation with attrition is Job Satisfaction, where those with lower Job Satisfaction have much higher attrition rates.

Propensity of Employee to Job Hop:

Next, we want to discover whether the past behaviour of employees (tendency to job-hop) has any effect on attrition rate. Thus to understand it, we created a new variable, Jobhop, that is equal to the employee's NumCompaniesWorked/TotalWorkingYears. This variable aims to show the propensity of an employee to Job Hop.

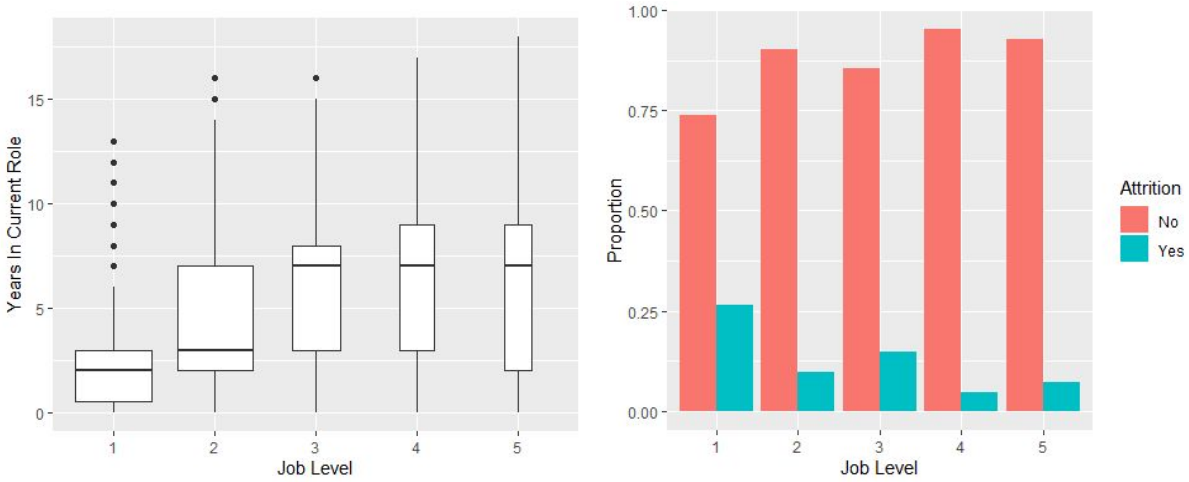
We created a bar graph that showed the proportion of attrition in each category of propensity to Job Hop, where quartile 1 is the lowest propensity and quartile 4 the highest propensity.



From the graph, we can observe that past job-hopping tendencies do indicate some relationship with the employees decision to attrite at IBM. The higher propensity to Job Hop (quartile 4) has the highest attrition while the lowest propensity to Job Hop (quartile 1) has the lowest attrition.

However, because the Job Hop variable is made from NumCompaniesWorked/TotalWorkingYears, the Pearson Correlation Value between Job Hop and NumCompaniesWorked is 0.70. Since we are more interested in the effects of Job Hop, we will remove NumCompaniesWorked from subsequent analysis to prevent the presence of multicollinearity.

Effects of JobLevel on Attrition:



We wanted to understand if JobLevel could offer some insight into Attrition Rates, so we plotted a graph of proportion of attrited employees against each Job Levels.

From the bar chart on the right, we see that at lower Job Levels, employees are more likely to quit, especially at JobLevel 1, the lowest, where more than 25% attrite. We have established on page 7 that employees who have spent fewer years in their current role are actually more likely to attrite. It is obvious from the boxplot on the left that there is some variance of Years In Current Role across Job Levels. Employees at lower Job Levels tend to have spent less time in their current role, and this is supported by the fact that the two variables have a small positive correlation of 0.34. Thus, we are unable to say through visualisation alone which of these variables has higher explanatory power with respect to Attrition. We will leave this to our model selection in the next section.

Hypotheses:

After completing our exploratory data analysis, we have decided to focus on the following hypothesis.

1. Employees that are younger are more likely to quit
2. Employees that are frustrated in their jobs are more likely to quit
3. Employees that have a tendency to hop between jobs are more likely to quit
4. Employees that have a lower job level are more likely to quit.

Model Selection

First, we split the dataset into train and test sets. (70% train, 30% test)

In order to evaluate the chosen factors, we decided to use a logistic regression model and evaluate it using accuracy, sensitivity, specificity, and the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. The performance of our model will be compared against other predictive models trained using the same data including a neural network, random forest model as well as bagging.

We decided to use logistic regression as the main model as it not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative). Logistic Regression can also check the significance of independent variables in predicting the dependent variable, helping us answer our hypothesis. However it should be noted that logistic regression assumes linearity between the dependent variable and the independent variables.

Bootstrap aggregation was included as a method which improves the stability and accuracy of the model while reducing variance by combining the predictions from multiple simpler models together. This helps reduce the problem of overfitting and results in more accurate predictions compared to any single model. However this comes at the cost of loss of interpretability of the model, since bagging allows us to check the importance of factors but not how they affect the final prediction, which in our case does not help us answer our hypotheses. Moreover there is an issue of high bias if modelling is not done properly. Finally, as the number of procedures increases it requires exponentially larger amounts of computing power, making it hard to scale up, for example if our data set were to be expanded significantly to include all of IBM's employees numbering in the hundreds of thousands in a real-world scenario running a similar model would require much more computing power.

We included random forest as a model as it is robust to outliers, as splits of variables depend on order and not magnitude of variables. This addresses the varied magnitude in our variables such as Age and Jobhop which can differ significantly. As compared to regression, random forest is non-linear and non-parametric, allowing for a wide variety of relationships between variables. As compared to bagging, random forest avoids overfitting as not all variables are chosen for every tree. However, random forest omits relationships between individual predictors, which regressions can account for using interactions

The neural network model works well for our purposes because they can learn from past examples of attrited employees. Furthermore, they are more fault tolerant compared to other models because they are able to respond to significant changes in input without being too sensitive to insignificant changes. However, similar to bagging models there is a loss of interpretability of the model along with issues of high bias if modelling is not done properly.

We first analysed the significance of each variable in predicting the likelihood of attrition. We do so by removing correlated variables identified during exploratory analysis, namely MonthlyIncome, TotalWorkingYears, NumCompaniesWorked, YearsAtCompany, JobRole, YearsWithCurrManager, DailyRate, HourlyRate and PercentSalaryHike to prevent multicollinearity.

```
Call:
glm(formula = Attrition ~ ., family = "binomial", data = train_test)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7667  -0.4720  -0.2433  -0.0879   3.2757

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     3.007e+00  1.596e+00   1.884 0.059566 .
Age                            -1.408e-02  1.368e-02  -1.030 0.303146
BusinessTravelTravel_Frequently 2.160e+00  5.099e-01   4.237 2.27e-05 ***
BusinessTravelTravel_Rarely     1.339e+00  4.699e-01   2.849 0.004390 **
DepartmentResearch & Development 1.748e-01  8.622e-01   0.203 0.839328
DepartmentSales                 1.208e+00  8.882e-01   1.360 0.173870
DistanceFromHome                4.287e-02  1.331e-02   3.221 0.001279 **
Education                      -3.628e-03  1.072e-01  -0.034 0.973008
EducationFieldLife Sciences    -1.800e+00  1.125e+00  -1.600 0.109598
EducationFieldMarketing        -1.357e+00  1.173e+00  -1.157 0.247215
EducationFieldMedical          -1.971e+00  1.134e+00  -1.738 0.082139 .
EducationFieldOther            -1.510e+00  1.184e+00  -1.276 0.201972
EducationFieldTechnical Degree -1.079e+00  1.152e+00  -0.937 0.348947
EnvironmentSatisfaction        -5.298e-01  1.043e-01  -5.080 3.77e-07 ***
GenderMale                      4.821e-01  2.281e-01   2.114 0.034545 *
JobInvolvement                 -7.672e-01  1.526e-01  -5.029 4.93e-07 ***
JobLevel2                      -1.653e+00  3.108e-01  -5.319 1.04e-07 ***
JobLevel3                      -6.467e-01  3.584e-01  -1.805 0.071147 .
JobLevel4                      -2.633e+00  7.789e-01  -3.380 0.000725 ***
JobLevel5                      -1.767e+00  7.680e-01  -2.300 0.021433 *
JobSatisfaction                -3.172e-01  9.729e-02  -3.261 0.001111 **
MaritalStatusMarried            1.284e-01  3.287e-01   0.390 0.696183
MaritalStatusSingle             8.052e-01  4.232e-01   1.903 0.057105 .
MonthlyRate                     2.803e-05  1.560e-05   1.797 0.072297 .
OverTimeYes                     1.936e+00  2.344e-01   8.258  < 2e-16 ***
PerformanceRating               3.624e-02  2.921e-01   0.124 0.901259
RelationshipSatisfaction       -2.675e-01  1.012e-01  -2.643 0.008217 **
StockOptionLevel               -2.785e-01  2.045e-01  -1.362 0.173243
TrainingTimesLastYear          -1.270e-01  8.610e-02  -1.475 0.140238
WorkLifeBalance                -3.411e-01  1.480e-01  -2.305 0.021193 *
YearsInCurrentRole             -1.278e-01  4.477e-02  -2.854 0.004316 **
YearsSinceLastPromotion         1.246e-01  4.813e-02   2.588 0.009645 **
jobhop                          1.652e+00  3.736e-01   4.421 9.82e-06 ***
```

As seen from the logistic regression above, not every variable has a p-value of less than 0.05, indicating that the coefficient of some variables at the 5% level of significance are not significantly different from zero and have no use in the prediction of likelihood of attrition among employees. Certain categorical variables are only significant for certain levels. For example, MartialStatus which has 3 levels, “Married”, “Divorced” and “Single” but only “Single” level shows significance in predicting attrition.

For such variables, dummy coding is applied based on a certain condition.

Variable 1: JobLevelHigh
Original variable = JobLevel (1-5)
JobLevelHigh = 1 when JobLevel is greater 3
JobLevelHigh = 0 when JobLevel is 3 and below

Variable 2: BusinessTravelFrequent
Original variable = BusinessTravel (“Travel Frequently”, “Travel Rarely”, “Non-travel”)

BusinessTravelFrequent = 1 when BusinessTravel is “Travel Frequently”
BusinessTravelFrequent = 0 when BusinessTravel is “Travel Rarely” or “Non-travel”

Variable 3: MarriedBefore
Original variable = MartialStatus (“Married”, “Divorced”, “Single”)
MarriedBefore = 1 when MartialStatus is “Married” or “Divorced”
MarriedBefore = 0 when MartialStatus is “Single”

After selecting the significant variables, we end up with Age, BusinessTravelFrequent, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, marriedBefore, OverTime, YearsSinceLastPromotion, YearsInCurrentRole, jobhop and RelationshipSatisfaction. It should be noted the p value of JobLevelHigh is greater than 0.05 and we excluded it from the following models as an independent variable.

```
Call:
glm(formula = Attrition ~ Age + BusinessTravelFrequent + DistanceFromHome +
    EnvironmentSatisfaction + JobInvolvement + JobSatisfaction +
    marriedBefore + OverTime + YearsSinceLastPromotion + YearsInCurrentRole +
    jobhop + RelationshipSatisfaction, family = "binomial", data = train_test)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9179  -0.5200  -0.3244  -0.1653   3.3026

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.49971    0.71733   4.879 1.07e-06 ***
Age                      -0.05211    0.01235  -4.220 2.45e-05 ***
BusinessTravelFrequent    1.11558    0.23187   4.811 1.50e-06 ***
DistanceFromHome          0.04574    0.01162   3.935 8.30e-05 ***
EnvironmentSatisfaction  -0.38564    0.09253  -4.168 3.08e-05 ***
JobInvolvement           -0.61765    0.13542  -4.561 5.09e-06 ***
JobSatisfaction          -0.42746    0.09037  -4.730 2.24e-06 ***
marriedBefore            -0.87409    0.20725  -4.218 2.47e-05 ***
OverTimeYes               1.77804    0.21190   8.391  < 2e-16 ***
YearsSinceLastPromotion   0.12877    0.03877   3.322 0.000895 ***
YearsInCurrentRole       -0.13645    0.03986  -3.423 0.000620 ***
jobhop                    1.57925    0.32099   4.920 8.66e-07 ***
RelationshipSatisfaction -0.23799    0.09200  -2.587 0.009688 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 909.34  on 1028  degrees of freedom
Residual deviance: 666.72  on 1016  degrees of freedom
AIC: 692.72

Number of Fisher Scoring iterations: 6
```

Using the diagram above, we can answer our hypotheses.

Hypothesis 1: Employees that are younger are more likely to quit.
At 95% level of significance, we have sufficient evidence to conclude that age affects the odds of attrition since its p-value is less than 0.05. Furthermore, the coefficient for age is negative, indicating that older workers are less likely to attrit and thus, we do not have enough evidence to reject our hypothesis.

Hypothesis 2: Employees that are frustrated in their jobs are more likely to quit.

At 95% level of significance, we have sufficient evidence to conclude that YearsSinceLastPromotion, JobSatisfaction, YearsInCurrentRole affects the odds of attrition since its p-value is less than 0.05. The coefficient of YearsSinceLastPromotion is positive while the other two variables are negative. This means that employees who have not been promoted for an extended period of time will be more likely to quit, supporting our initial theory. The negative coefficient of job satisfaction also supports our theory that employees who are more satisfied with their job will have a lower odds of quitting.

However, the negative coefficient of years in the current role is contrary to our initial theory that employees stuck in the same role for an extended period of time will be frustrated and quit. Employees who stayed in the same role are less likely to quit. This suggests that the employees who have kept the same role enjoy what they are doing and is inversely proportional to frustration.

Furthermore, BusinessTravelFrequent, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, marriedBefore, OverTime and RelationshipSatisfaction seem to affect the frustration level of employees as well.

Hypothesis 3: Employees that have a tendency to hop between jobs are more likely to quit.

At a 95% level of significance, we have sufficient evidence to conclude that jobhop (indicating employee tendency to hop between jobs) affects the odds of attrition since its p-value is less than 0.05. The coefficient of jobhop is positive which means that the higher an employees tendency to jobhop, the more likely they are to attrit, which both makes sense and supports our initial theory.

Hypothesis 4: Employees that have a lower job level are more likely to quit.

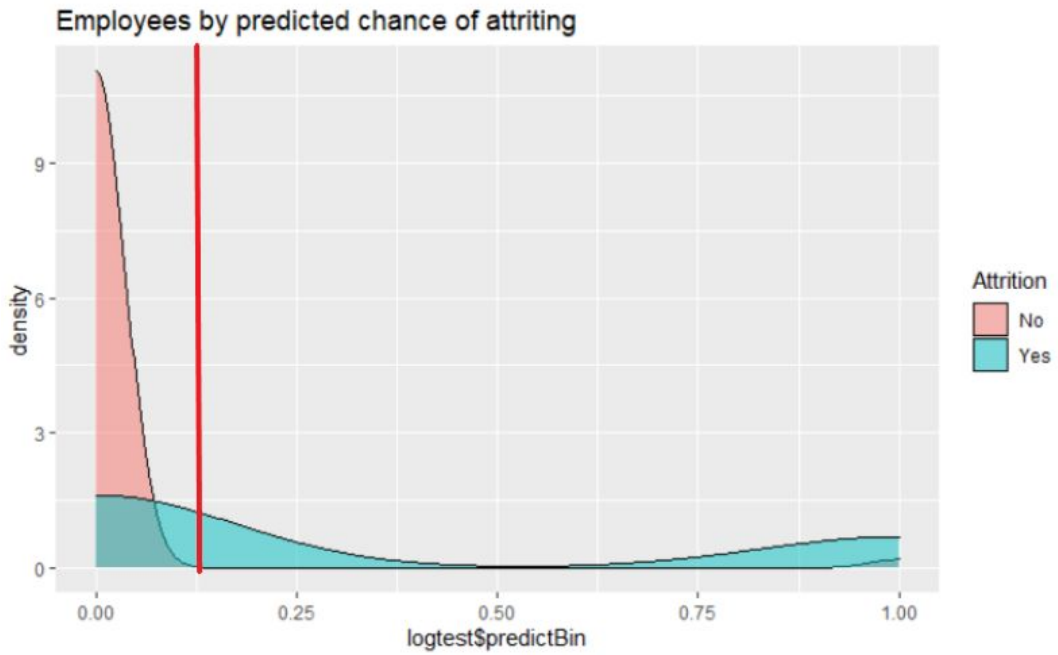
At a 95% level of significance, we did not have sufficient evidence to conclude that JobLevelHigh (indicating whether an employee was in a high job level) affects the odds of attrition since its p-value is more than 0.05. The coefficient of JobLevelHigh was negative which means that the higher an employees job level, the less likely they are to attrit, which supports our initial theory. However, due to the p-value being more than 0.05 we had to conclude that at a 95% level of significance that the coefficient of JobLevelHigh was not significantly different from 0, and hence it had to be removed from our models.

Model 1: Logistic Regression

Using a logistic regression model trained with the train set, we predicted the Attrition outcome of the test set. Since logistic regression predicts the odds ratio, any predicted value above 0.5 will be considered as employees who attrited while any value below that will be employees who stay. The model gives the following result:

Model	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression	0.6417	0.8594	0.3000	0.9834

While the accuracy is high, the AUC and sensitivity is not ideal. The sensitivity is low at 0.3000. Since there is a significantly lower percentage of data with attrition, 83.9% no attrition and 16.1% attrition, a high accuracy can be achieved by simply predicting 'No' to attrition. Upon further examining the model, we find that 0.5 is not a good split to determine whether an employee attrit or not. As seen from the graph below, a better split will be around 0.18, which is quite close to the proportion of attrited employees in the original input data.



As seen in the graph above, splitting at 0.5 results in a large number of attrited personnel being classified as non attrited, resulting in the low sensitivity result. However, since the vast majority of non-attrited personnel are clustered to the left side of the graph between 0.00 and 0.18, we can dramatically increase sensitivity at the cost of a small amount of specificity by setting the classification cutoff at 0.18 instead, where we shall classify the employee in question as attrited should the predicted value be 0.18 instead of 0.50.

After splitting at 0.18, the AUC and sensitivity significantly improves.

Model	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression	0.7655	0.7914	0.7250	0.8061

Model 2: Bagging

Next, we trained a Bagging model with the same training set. When used to predict on the test set, the model performs the following:

Model	AUC	Accuracy	Sensitivity	Specificity
Bagging	0.7418	0.8549	0.3194	0.9593

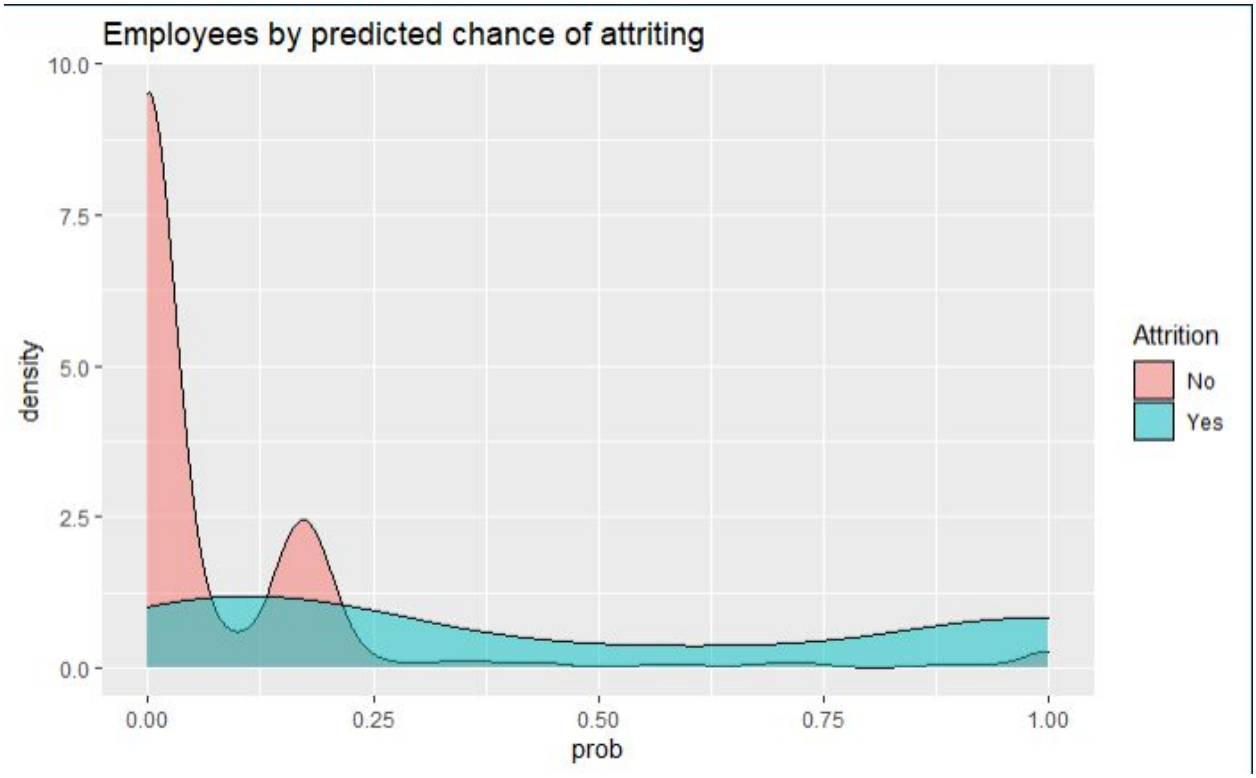
The model performs well compared to the logistic regression model in terms of accuracy and specificity. Despite this, it has a marginally lower AUC and much lower sensitivity. From this we can conclude that this is likely due to the model classifying the majority of the samples as non-attritted personnel (the majority group in our dataset) which results in a high accuracy rate at the cost of specificity. Hence Bagging is not particularly useful for our purposes where we aim to maximise AUC and specificity.

Model 3: Neural Network

In addition, we also tried to train a neural network with the same training set. Similar to logistic regression which predicts the odds ratio, any predicted value above 0.5 will be considered as employees who attrited while any value below that will be employees who stay. The model gives the following result:

Model	AUC	Accuracy	Sensitivity	Specificity
Neural Network	0.7934	0.8707	0.4267	0.9617

This model performs very well on the AUC and Specificity front, with higher accuracy compared to the previous models as well. However this comes at the cost of sensitivity, which is not ideal given that we aim to identify as many employees who are likely to attrit as possible and hence require a model with high sensitivity. However, similar to the logistic regression, since the neural network returns predicted values between 0 and 1 for each test case, we can see that 0.5 is not a good split to determine whether an employee quits or not. As seen from the graph below, a better split will be around 0.18, which is quite close to the proportion of attrition in the original input data. With the split, the model classifies more attrited employees correctly while misclassifying fewer non-attrited employees.



After splitting at 0.18, sensitivity is improved at the expense of AUC and specificity.

Model	AUC	Accuracy	Sensitivity	Specificity
Neural Network	0.7274	0.8458	0.5600	0.9044

Model 4: RandomForest

Finally, a random forest model is trained with the same training set. When used to predict on the test set, the model performs as follows:

Model	AUC	Accuracy	Sensitivity	Specificity
Random Forest	0.8660	0.9478	0.7375	0.9945

Overall Performance across all models:

Model	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression (split at 0.5)	0.6417	0.8594	0.3000	0.9834
Logistic Regression (split at 0.18)	0.7655	0.7914	0.7250	0.8061
Bagging	0.7418	0.8549	0.3194	0.9593
Neural Network (split at 0.5)	0.7934	0.8707	0.4267	0.9617
Neural Network (split at 0.18)	0.7274	0.8458	0.5600	0.9044
Random Forest	0.8660	0.9478	0.7375	0.9945

The random forest model performs the best, with all 4 measures being higher than the other models. However, it can be noted that there is still a relatively big gap between specificity and sensitivity.

Out of our 4 indicators of success, we would place the most importance on sensitivity and AUC, as we are aiming to find out through our hypothesis what factors are indicators of an employee likely to attrit, so as to help the human resources department save on retraining and hiring costs. As such, we would like to maximise sensitivity in our models to catch as many attrited employees as possible but balanced against maintaining a high AUC so that we are not just identifying every employee as a potential attrition risk, and potentially wasting resources on engaging these wrongly identified employees.

Initially all models performed to differing degrees of success on the test set, especially with regards to sensitivity (finding the personnel who left the company), which was lower than specificity across the board at first glance. However, this can be attributed to the imbalance in our original dataset with much more samples for non-attrited personnel compared to attrited personnel. Overall the Random Forest model performed the best, attaining the highest AUC, Accuracy, specificity and sensitivity.

Oversampling and Undersampling:

The huge imbalance in the dataset with regards to the ratio of attritioning to remaining employees, with attrition making up a mere 16% of the employees in our dataset, may affect our outcome variable, severely underrepresenting attrited employees and ultimately making it difficult for our models to have high sensitivity.

In order to address this, we decided to try two sampling methods, over and under-sampling so as to make up for the imbalance in our dataset and properly train our models. After training our models on the new sample training sets, performance on the test sets was much improved. (As seen in the table below).

Measure of different models after conducting **oversampling**

Model (Over)	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression	0.7468	0.7687	0.7125	0.7812
Bagging	0.6530	0.8299	0.4426	0.8921
Neural Network	0.7003	0.8254	0.5867	0.8743
Random Forest	0.8868	0.9501	0.7875	0.9861

Measure of different models after conducting **undersampling**

Model (Under)	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression	0.7275	0.7052	0.7625	0.6925
Bagging	0.6325	0.7188	0.7067	0.7213
Neural Network	0.6721	0.7687	0.7600	0.7705
Random Forest	0.8295	0.7846	0.9000	0.7590

After conducting oversampling, we observe that model performance across the board remains similar or has improved, with AUC, accuracy, specificity and sensitivity increasing as a result. However, for the Bagging and Neural network models sensitivity still remains low, albeit improved from the original model. On the other hand, upon trying under sampling, specificity performance is wildly improved, which unfortunately comes at the cost of AUC, accuracy and sensitivity.

Conclusion

Given the context of a company's Human Resources department and our motivation of trying to identify as many likely to attrition employees without misclassifying those who want to stay on, since the Random Forest model trained using undersampling outshined all our other models, it should be the primary model used in decision making. However, should there be limitations to the number of employees that the Human resource department can engage or prepare to replace, it would be better to use the oversampled model which has the highest accuracy.

Areas for Future Improvement

One problem that we found was that for the same model and same variables, model performance could differ moderately simply from just what subset of employees were randomly selected into the training and test set, with the predicted test results indicating the model to be varying from mediocre to good and sometimes completely unusable. We tried to resolve this reproducibility issue by training all our models on the same training dataset and evaluating it on the same testing set. However this might reflect a deeper underlying problem with either the dataset variables not being predictors of attrition, which could be due to our limited dataset. We would like to try and resolve this issue by getting more samples from similar datasets of other companies such as Apple, Hewlett-Packard or Lenovo so as to see if our models are generalizable to other companies and if our problems were in fact caused by having a limited dataset to work with.

Another possible expansion of scope which we thought to include would be to add interaction terms into our logistic regression model, in order to account for the relationship between factors that could interact with each other, similar to how we created the job hopping propensity variable jobhop. It is possible that other such factors such as job satisfaction and monthly income could also be related, and our current analysis emits such possibilities. One such example we can think of is age and the need to make frequent business trips. A younger employee may be more willing to take on jobs that require frequent business trips as compared to an older employee who may have settled down. This could increase the accuracy of our logistic regression model as compared to the rest as they cannot include such interactions.