Introduction to dataset

Dataset: IBM HR Employee Attrition dataset

Problem description: Employee attrition is a critical issue faced by organizations globally, impacting productivity, morale, and the overall cost of recruitment and training. High attrition rates can lead to skill gaps, reduced team cohesion, and increased operational expenses. Understanding the factors contributing to attrition, such as job dissatisfaction, lack of growth opportunities, and work-life balance challenges, is essential for developing effective retention strategies.

The objective of the project would be identifying methods to increase employee retention, to directly address the attrition problem by identifying the reasons for attrition. Therefore identifying employees that are likely to quit and taking related actions to entice employees to stay.

Charts describing the dataset's makeup of data in relation to attrition column value

Logistic Regression

Preprocessing steps taken to prepare for logistic regression:

1)Identify highly skewed variables like Monthly income and normalize through scaling

2)Creating bins for Age variable

3)Deriving useful features, Eg: Work satisfaction, a feature based on Job satisfaction and Work life balance.

4)Performing binary encoding of Gender, and Overtime

5)Dimension reduction: Remove columns having only one unique value and source columns of transformations, bins, and encoding.

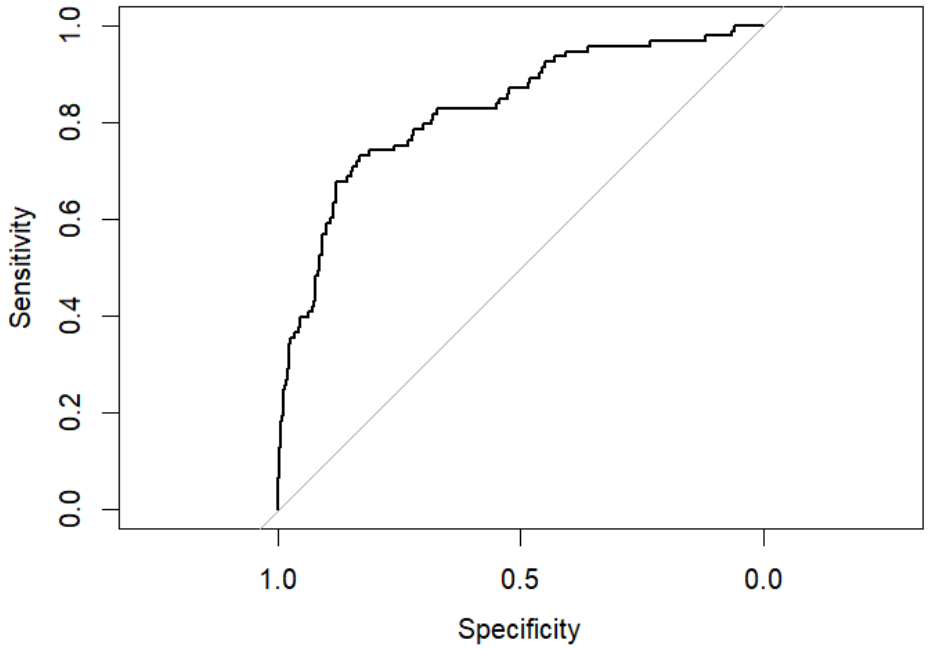
Results obtained from the logistic regression model are as follows:

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -4.6478     0.8848  -5.253 1.50e-07 ***
BusinessTravelTravel_Frequently   2.6079     0.6121   4.261 2.04e-05 ***
BusinessTravelTravel_Rarely       1.7241     0.5815   2.965 0.003030 **
DepartmentResearch & Development -0.2616     0.5512  -0.475 0.635081
DepartmentSales                   0.7802     0.5658   1.379 0.167877
DistanceFromHome                  0.3653     0.1096   3.332 0.000862 ***
EnvironmentSatisfaction          -0.4182     0.1127  -3.710 0.000207 ***
GenderMale                        0.3994     0.2313   1.727 0.084224 .
JobInvolvement                   -0.4546     0.1116  -4.076 4.59e-05 ***
JobLevel                         -0.6995     0.2308  -3.031 0.002436 **
MaritalStatusMarried              0.3106     0.3244   0.957 0.338316
MaritalStatusSingle               1.2120     0.3245   3.735 0.000187 ***
NumCompaniesWorked                0.4759     0.1193   3.988 6.67e-05 ***
TotalWorkingYears                -0.4091     0.2611  -1.567 0.117049
YearsAtCompany                    0.9002     0.2650   3.398 0.000679 ***
YearsInCurrentRole               -0.5498     0.1964  -2.800 0.005111 **
YearsSinceLastPromotion           0.2623     0.1634   1.606 0.108333
YearsWithCurrManager             -0.4644     0.1996  -2.327 0.019974 *
Age_Group31-40                   -0.7685     0.2795  -2.749 0.005978 **
Age_Group41-50                   -0.6334     0.3697  -1.713 0.086632 .
Age_Group51-60                    0.2235     0.5101   0.438 0.661211
WorkSatisfaction                 -0.5914     0.1213  -4.876 1.08e-06 ***
Overtime_Encoded                  0.7327     0.1064   6.886 5.75e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 770.12  on 876  degrees of freedom
Residual deviance: 537.12  on 854  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 583.12

Number of Fisher Scoring iterations: 6
```

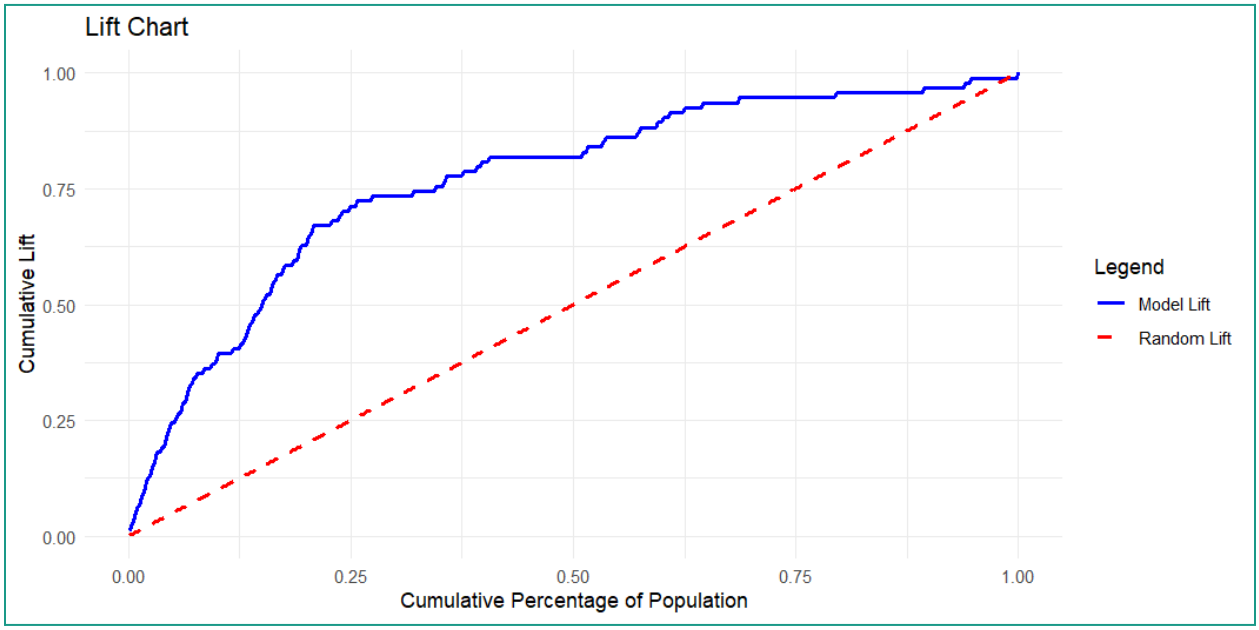
Summary of model

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  470  57
       Yes  22  36

               Accuracy : 0.865
                 95% CI : (0.8346, 0.8916)
    No Information Rate : 0.841
    P-Value [Acc > NIR] : 0.0611500

                  Kappa : 0.404

 Mcnemar's Test P-Value : 0.0001306

            Sensitivity : 0.9553
            Specificity : 0.3871
         Pos Pred Value : 0.8918
         Neg Pred Value : 0.6207
             Prevalence : 0.8410
         Detection Rate : 0.8034
   Detection Prevalence : 0.9009
      Balanced Accuracy : 0.6712

       'Positive' Class : No
```

Confusion matrix



ROC Curve: Resulting AUC=0.8282



Lift chart

	Df	Deviance	AIC
<none>		537.12	583.12
+ DailyRate	1	535.37	583.37
- TotalWorkingYears	1	539.68	583.68
- YearsSinceLastPromotion	1	539.70	583.70
+ PercentSalaryHike	1	535.75	583.75
- Gender	1	540.16	584.16
+ TrainingTimesLastYear	1	536.70	584.70
+ Education	1	536.75	584.75
+ PerformanceRating	1	536.83	584.83
+ NormalizedIncome	1	537.09	585.09
+ HourlyRate	1	537.09	585.09
+ StockOptionLevel	1	537.12	585.12
+ MonthlyRate	1	537.12	585.12
- YearsWithCurrManager	1	542.49	586.49
+ JobRole	8	525.40	587.40
- Age_Group	3	548.74	588.74
- YearsInCurrentRole	1	545.01	589.01
+ EducationField	5	533.63	589.63
- JobLevel	1	546.90	590.90
- YearsAtCompany	1	547.92	591.92
- DistanceFromHome	1	548.11	592.11
- EnvironmentSatisfaction	1	551.24	595.24
- Department	2	554.29	596.29
- NumCompaniesWorked	1	552.64	596.64
- JobInvolvement	1	554.15	598.15
- MaritalStatus	2	557.06	599.06
- BusinessTravel	2	564.16	606.16
- WorkSatisfaction	1	563.22	607.22
- Overtime_Encoded	1	587.01	631.01

Stepwise Regression performed and resulting columns

From the following results, we have concluded that the following circumstances will likely lead to an employee's attrition:

1)Frequent Business Travel

2)Poor Environment Satisfaction

3)Low Job Involvement

4)Low Work Satisfaction

5)High number of companies worked at

6)High overtime hours

These points also affect employee attrition to a certain extent, but there is no instant solution to the issues:

1)High distance from home to workplace

2)Marital status: Single

3)Low years with current manager

4)Low years in current role

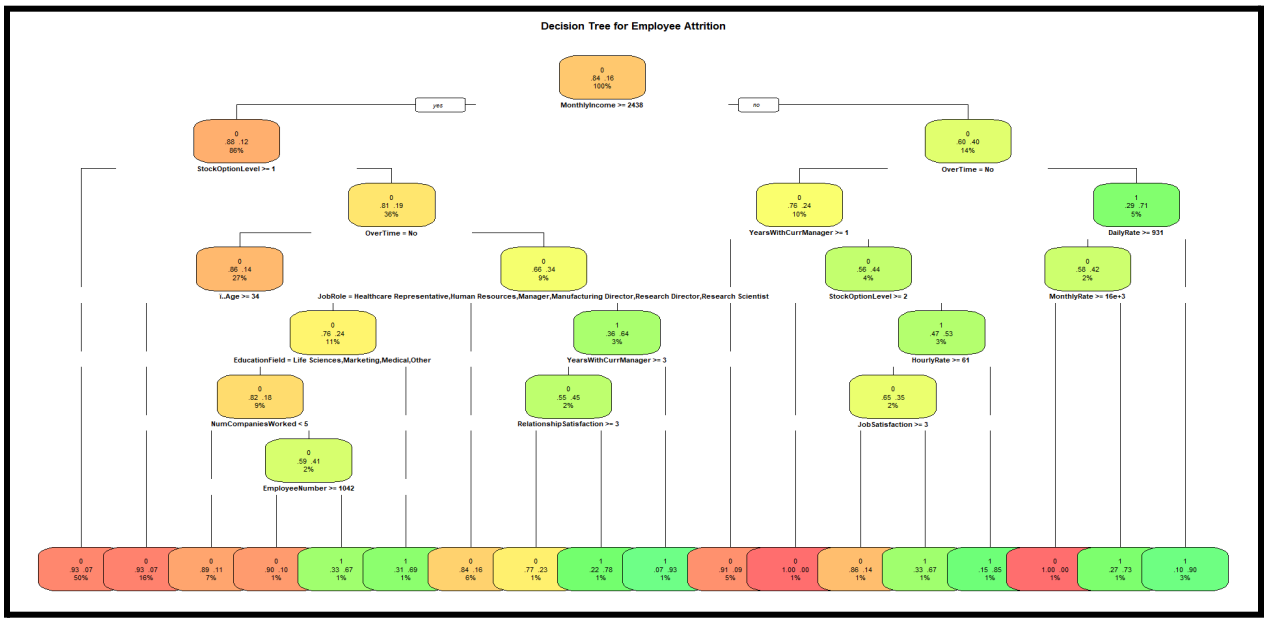
Decision Tree

Preprocessing Steps:

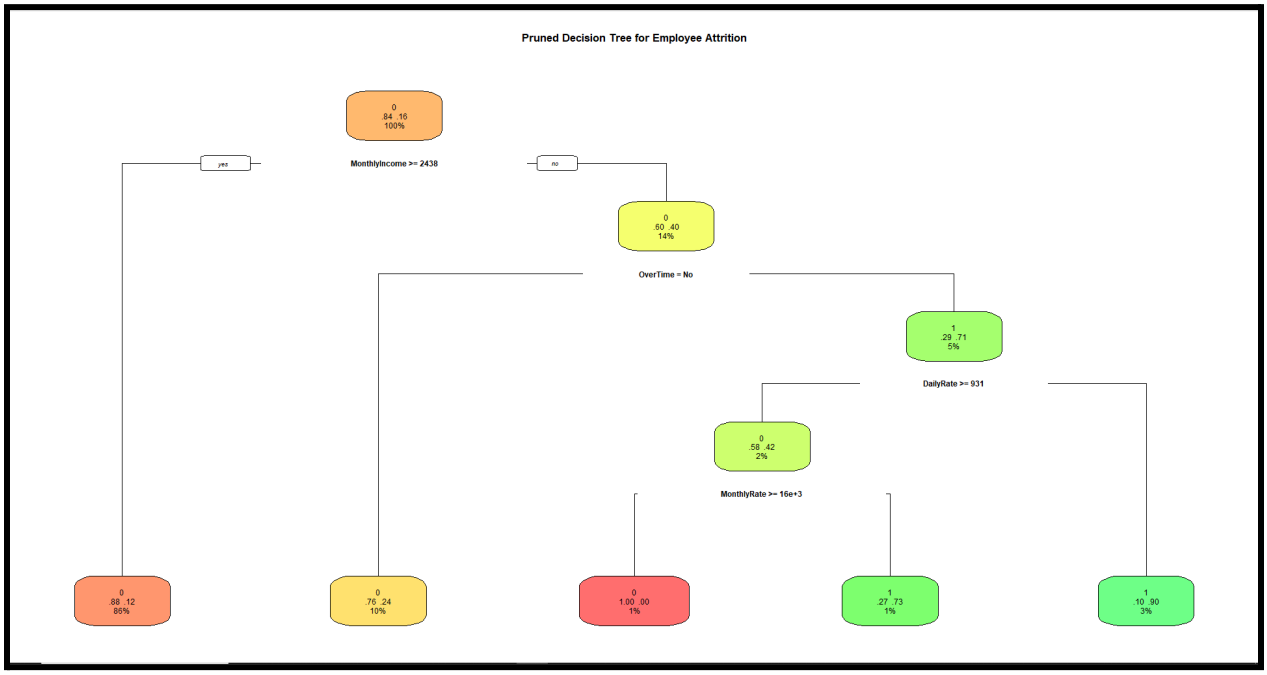
1. **Conversion of Target Variable:**
    - Transformed the Attrition column from categorical ("Yes"/"No") to binary (1 for "Yes," 0 for "No").
2. **Column Removal:**
    - Dropped irrelevant columns:
        - **Identifiers**: EmployeeNumber, ID (non-predictive, unique identifiers).
        - **Unimportant Variables**: YearsAtCompany, Hourly Rate (non-predictive or low impact).
        - **Correlated Variables**: Job Level (high correlation with Monthly Income).
3. **Correlation Analysis:**
    - Generated a **correlation heatmap** to identify relationships between numeric variables for further feature selection.
    - Dropped few columns which have high correlation with other columns to avoid bias.
4. **Data Splitting:**
    - Split the dataset into **training** (70%) and **validation** (30%) sets for modeling.

Complete Decision Tree:

Max depth=12 Min bucket =5

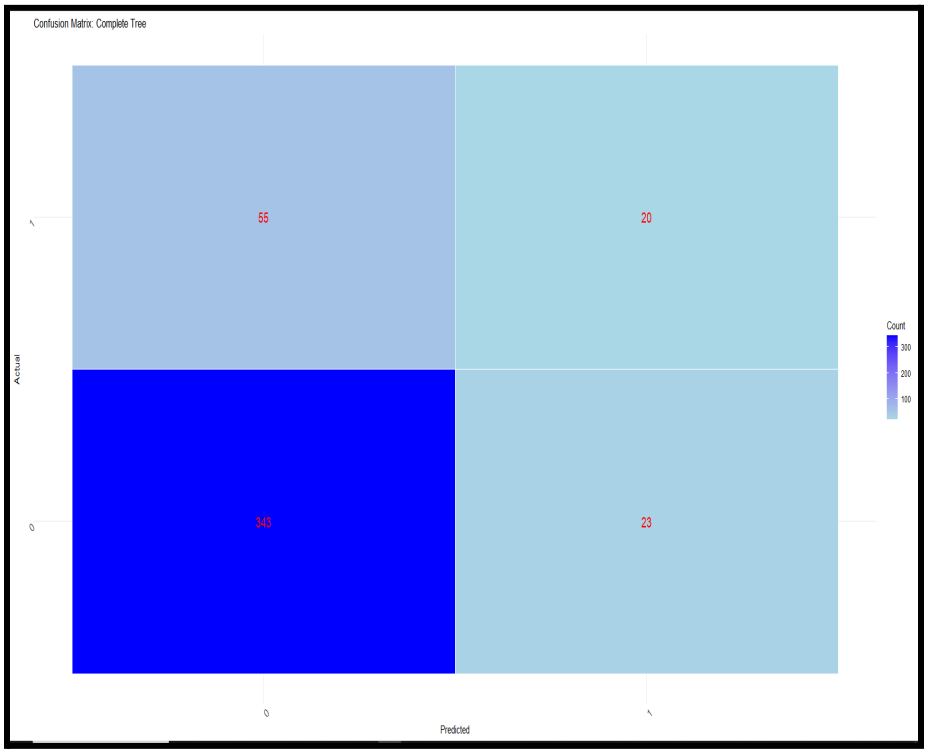


Pruned Tree using the complexity parameter from complete Tree :

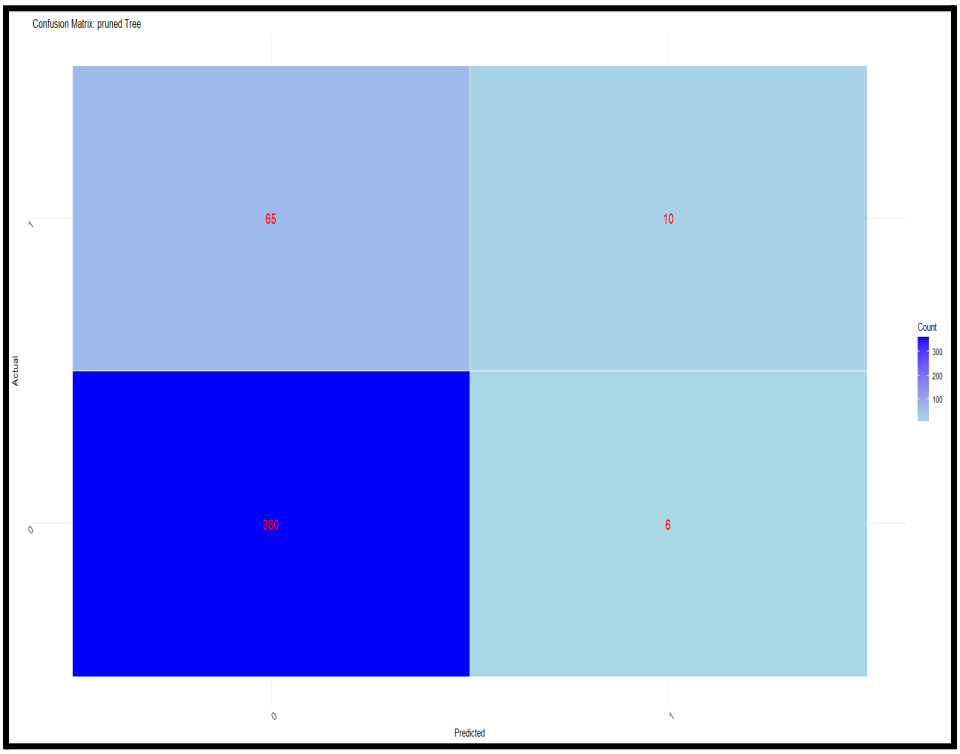


It is clearly seen that monthly income and overtime has more weightage compared to other columns. Compared to the complete tree pruned tree has taken the structure well. IBut the confusion matrix says different stories

Confusion Matrix of complete Tree:



Confusion Matrix of Pruned tree:



The Confusion matrix says that False positives have increased. False Negatives decreased and accuracy of predicting actual attrition employees also decreased. Which means that this model may not suit the data.

```
> print(conf.matrix)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 360  65
         1   6  10

               Accuracy : 0.839
                 95% CI : (0.8013, 0.8721)
    No Information Rate : 0.8299
    P-Value [Acc > NIR] : 0.3328

                  Kappa : 0.1702

 Mcnemar's Test P-Value : 5.847e-12

            Sensitivity : 0.9836
            Specificity : 0.1333
         Pos Pred Value : 0.8471
         Neg Pred Value : 0.6250
             Prevalence : 0.8299
         Detection Rate : 0.8163
   Detection Prevalence : 0.9637
      Balanced Accuracy : 0.5585

       'Positive' Class : 0
```

Based on the Accuracy stats, we can see that accuracy is high, the model is very good at predicting the values for the data. But it has very low kapa score. Which says otherwise about the model. This is one of the main reason we discarded this model.

Variable Importance for this Pruned tree is:

```
Variable importance
         MonthlyIncome              OverTime       TotalWorkingYears       StockOptionLevel               JobRole
                    16                    12                       6                       5                     5
            Department         MaritalStatus  YearsWithCurrManager                     Age             DailyRate
                     5                     5                       4                       4                     4
        EmployeeNumber             Education     YearsInCurrentRole          YearsAtCompany       PercentSalaryHike
                     4                     3                       3                       3                     3
            HourlyRate              JobLevel     NumCompaniesWorked         EducationField EnvironmentSatisfaction
                     3                     3                       2                       2                     2
YearsSinceLastPromotion         MonthlyRate          JobInvolvement      DistanceFromHome          BusinessTravel
                     2                     1                       1                       1                     1
       WorkLifeBalance
                     1
```

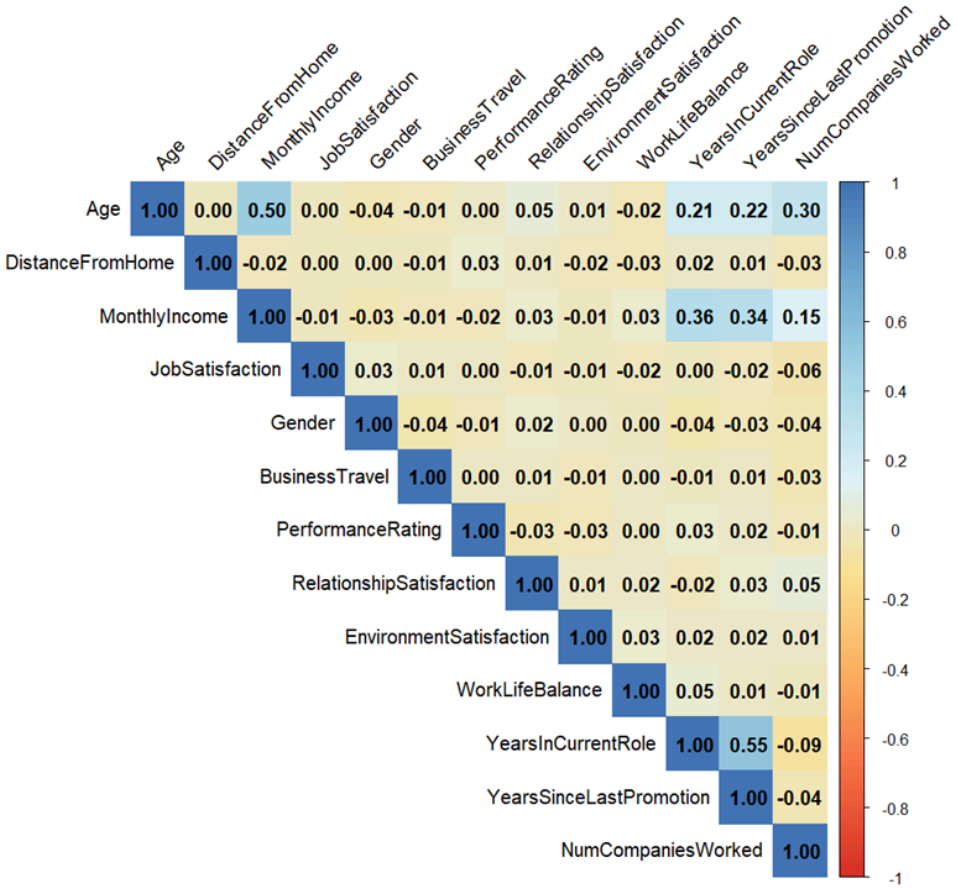
Neural Network

Preprocessing steps taken:

1. **Conversion of Variables:**
    - Transformed the Attrition column from categorical ("Yes"/"No") to factor(1 for "Yes," 0 for "No").
    - Transformed the Gender column from categorical ("Male"/"Female") to numeric(1 for "Yes," 0 for "No").
    - Transformed the Business Travel column from categorical ("Non-Travel", "Travel_Rarely", "Travel_Frequently") to numeric

2. **Column Removal based on Correlation and Relevance:**

    Dropped irrelevant columns: EmployeeNumber, ID (non-predictive, unique identifiers) and attributes with high correlation (above 0.6)



Correlation Chart of chosen variables (variables with correlation < ±0.6)

3. **Scaling of chosen attributes:**

    - Scaling is performed on the chosen attributes to make sure the large scale attributes like monthly income does not dominate during training and all features contribute equally to the model learning process
    - Interaction terms (like YearsInCurrentRole * PerformanceRating) can introduce large values when multiplying unscaled features, exacerbating the scale difference. Scaling ensures these terms are brought to the same level as other features.

4. **Data Splitting:**

    - Split the dataset into **training** (60%) and **validation** (40%) sets for modeling.

Hyper-Parameter selection and Feature Engineering

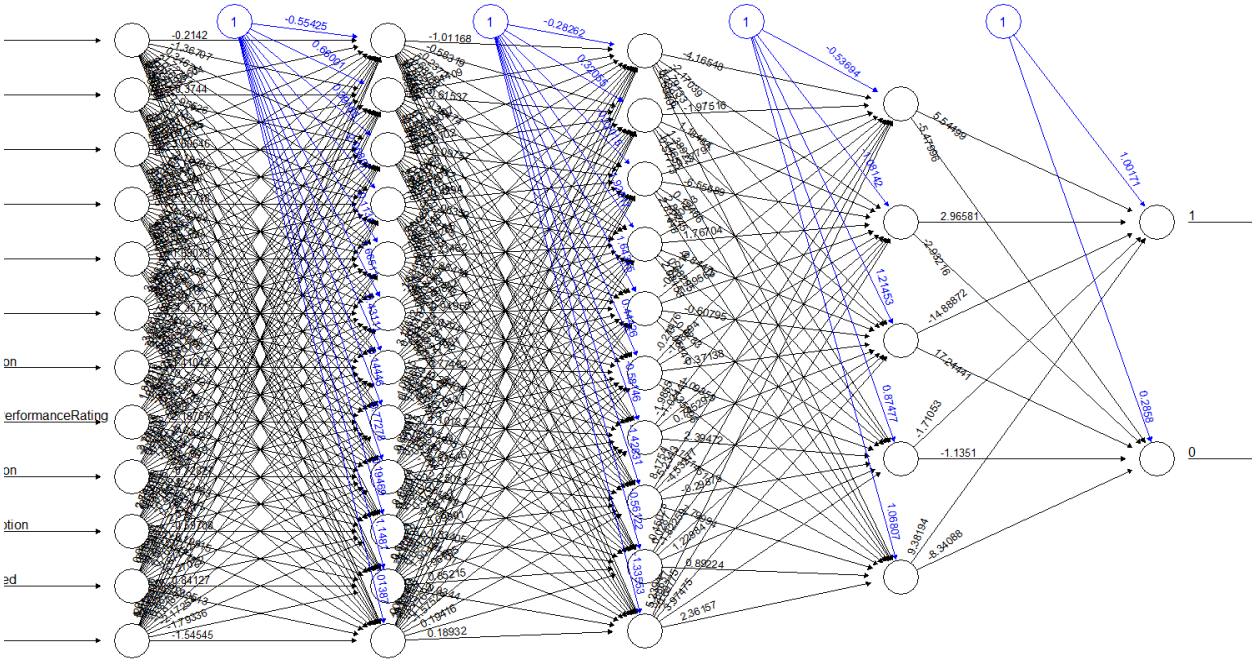
The hyperparameters have been tuned to improve the accuracy. The following are the parameters of the model:

- 3 hidden layers with 12, 10 and 5 nodes in the respective layers
- Maximum steps: 10^8
- Learning rate : 0.005
- linear.output : False

Feature Engineering

- Established an interaction term between the attributes (**YearsInCurrentRole * PerformanceRating)** to improve the model accuracy.
- Scaling the resulting data to ensure accurate contribution from each variable

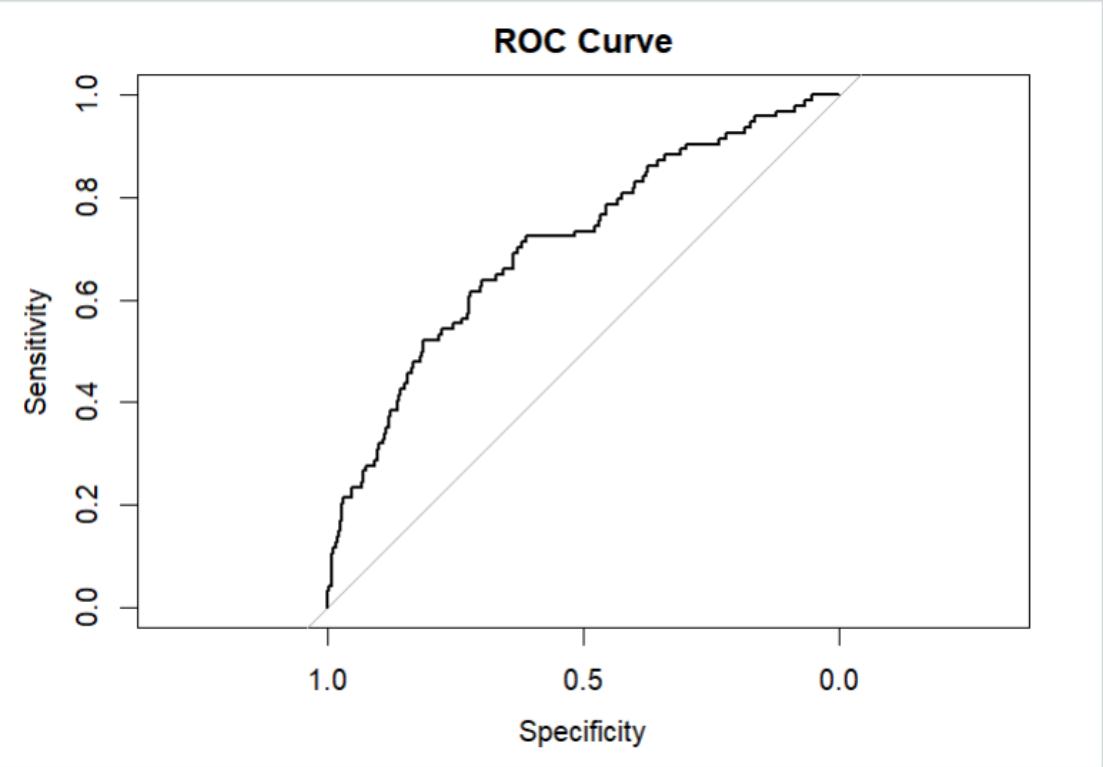
Neural Network Results



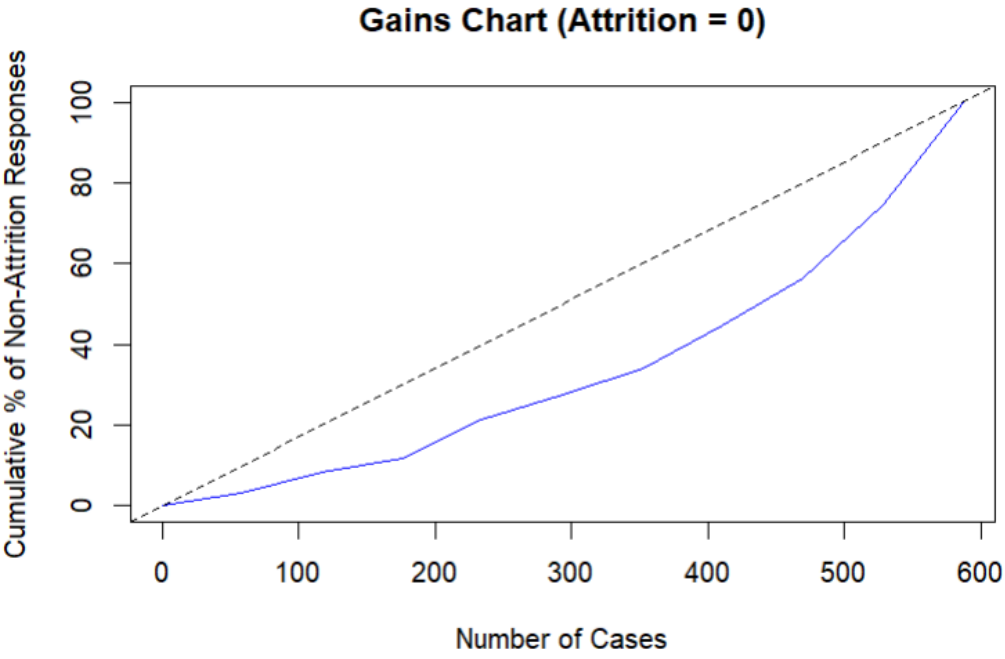
Neural Network Model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 461  72
         1  32  22

               Accuracy : 0.8228
                 95% CI : (0.7895, 0.8529)
    No Information Rate : 0.8399
    P-Value [Acc > NIR] : 0.8804433

                  Kappa : 0.2043

 Mcnemar's Test P-Value : 0.0001312

            Sensitivity : 0.9351
            Specificity : 0.2340
         Pos Pred Value : 0.8649
         Neg Pred Value : 0.4074
             Prevalence : 0.8399
         Detection Rate : 0.7853
   Detection Prevalence : 0.9080
      Balanced Accuracy : 0.5846

       'Positive' Class : 0
```

Confusion Matrix denoting Accuracy, Sensitivity and Specificity



Accuracy Result and ROC curve with area 0.71



Gains chart for Attribution = 0

Model evaluation and findings

- The model performance is better than random but is not ideal at predicting non-attrition cases early on based on the Gains chart.
- The model is **moderately effective** at predicting employee attrition based on the ROC curve with an area under curve 0.71.
- The accuracy of the model based on the training and validation dataset is 82.28%, with a sensitivity of 93.51% and specificity of 23.4%

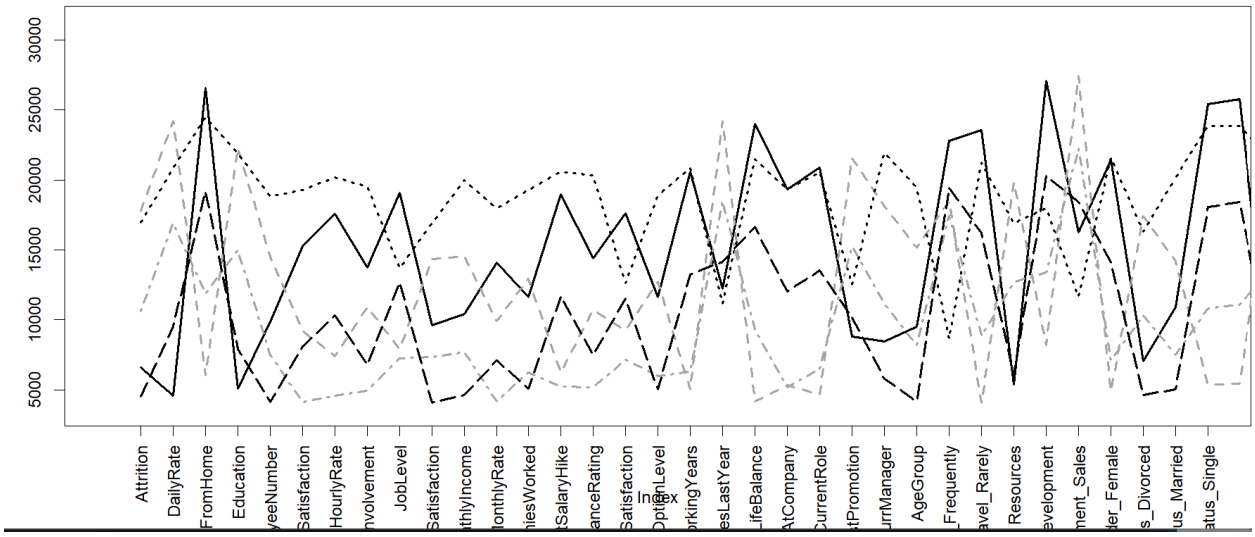
Clustering and association rules

Preprocessing steps taken

An agile approach was taken, and the preprocessing steps reflect the final model that was used.

The data was normalized with min/max scales, to change the values to ranges of large values such as pay(monthly rate, daily rate, monthly pay), distance from home. This helps to balance the values to prevent large values from creating a cluster on its own.
Removing columns that are irrelevant with the same data across rows, or numerical values that hold no meaning.
Data was binned into groups where appropriate, mainly Age data.
Splitting up columns using fastdummies into dummy columns,(department,education field,business travel etc), then removal of original columns and redundant columns after split(ie:male/female).
There was also further conversion of binary variables to a yes/no format to allow for categorization of output in the association rules portion.

Clustering results



The clusters generated could help to predict relations between factors such as employee's satisfaction, employee's pay, work-life balance, to other factors such as years at company, distance from home, and age group. But did not provide an accurate measure to predict attrition.

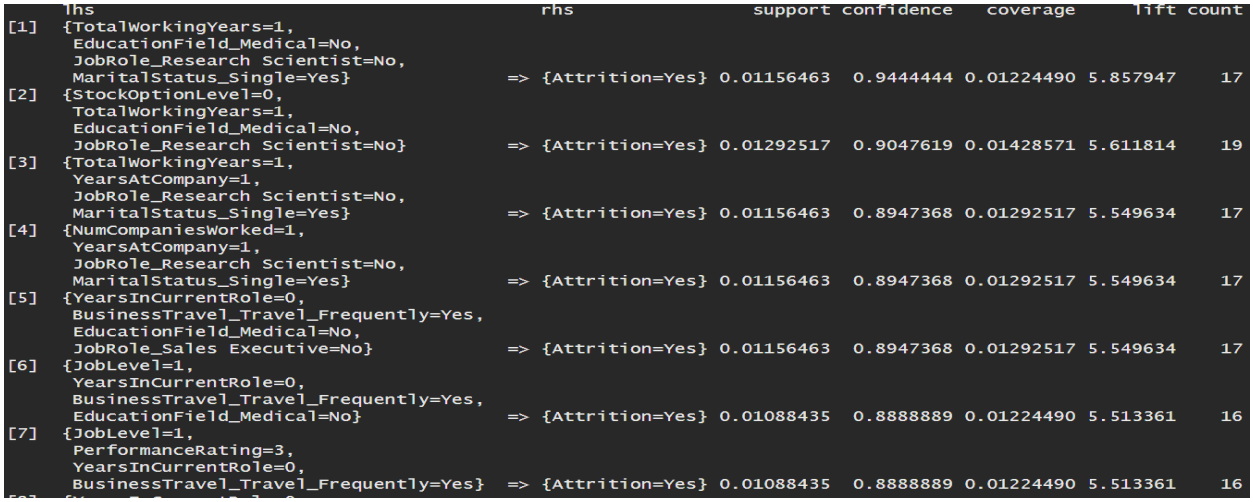
Judging from the results from the clusters generated with our dataset, it seemed that the data was unsuitable to generate usable information for our scenario, which was to generate insights from the data to find predictors that contributed to employee attrition.

Association rules

```
     lhs                                                rhs              support confidence  coverage     lift count
[1]  {StockOptionLevel=1,
      WorkLifeBalance=3,
      Department_Human Resources=No}                 => {Attrition=No} 0.2115646  0.9339339 0.2265306 1.113449   311
[2]  {StockOptionLevel=1,
      WorkLifeBalance=3,
      Department_Human Resources=No,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.2115646  0.9339339 0.2265306 1.113449   311
[3]  {StockOptionLevel=1,
      WorkLifeBalance=3}                             => {Attrition=No} 0.2231293  0.9318182 0.2394558 1.110927   328
[4]  {StockOptionLevel=1,
      WorkLifeBalance=3,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.2231293  0.9318182 0.2394558 1.110927   328
[5]  {JobLevel=2,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.2360544  0.9302949 0.2537415 1.109111   347
[6]  {JobLevel=2,
      Department_Human Resources=No,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.2285714  0.9281768 0.2462585 1.106585   336
[7]  {StockOptionLevel=1,
      BusinessTravel_Travel_Frequently=No,
      Department_Human Resources=No}                 => {Attrition=No} 0.2884354  0.9237473 0.3122449 1.101305   424
[8]  {StockOptionLevel=1,
      BusinessTravel_Travel_Frequently=No,
      Department_Human Resources=No,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.2884354  0.9237473 0.3122449 1.101305   424
[9]  {StockOptionLevel=1,
      BusinessTravel_Travel_Frequently=No}           => {Attrition=No} 0.3027211  0.9232365 0.3278912 1.100696   445
[10] {StockOptionLevel=1,
      BusinessTravel_Travel_Frequently=No,
      MaritalStatus_Single=No}                       => {Attrition=No} 0.3027211  0.9232365 0.3278912 1.100696   445
```

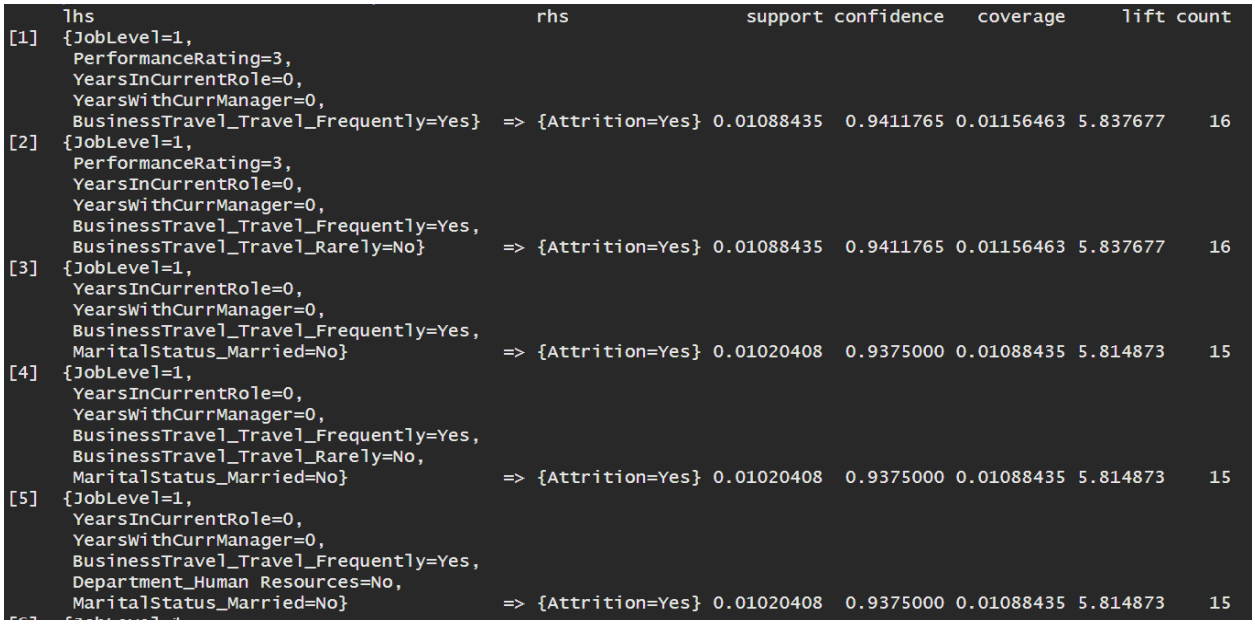
With the first iteration of the model, the results showed that it was heavily skewed towards no employee attrition, due to the fact that the dataset was about 80:20 ratio for attrition=yes : attrition=no. Hence the decision to further transform the model to only allow Attrition=Yes on the RHS(result factor), as well as lowering the required support level, reducing the required length and raising the maximum length in the case that the factors went above 5.

```
attrition_rules_YN <- apriori(Attrition.trans,
                              parameter = list(supp = 0.2, conf = 0.6, minlen = 3,maxlen=5),
                              appearance = list(rhs = c("Attrition=Yes", "Attrition=No"), default = "lhs"))
#sorted_rules <- sort(attrition_rules, by = "lift")
inspect(sort(attrition_rules_YN, by = "lift")[1:10])  # View top 10 rules by lift

attrition_rules <- apriori(Attrition.trans,
                           parameter = list(supp = 0.01, conf = 0.6, minlen = 2,maxlen=10),
                           appearance = list(rhs = c("Attrition=Yes"), default = "lhs"))

inspect(sort(attrition_rules, by = "lift")[1:10])
```



	lhs		rhs	support	confidence	coverage	lift	count
[1]	{TotalWorkingYears=1, EducationField_Medical=No, JobRole_Research Scientist=No, MaritalStatus_Single=Yes}	=>	{Attrition=Yes}	0.01156463	0.9444444	0.01224490	5.857947	17
[2]	{StockOptionLevel=0, TotalWorkingYears=1, EducationField_Medical=No, JobRole_Research Scientist=No}	=>	{Attrition=Yes}	0.01292517	0.9047619	0.01428571	5.611814	19
[3]	{TotalWorkingYears=1, YearsAtCompany=1, JobRole_Research Scientist=No, MaritalStatus_Single=Yes}	=>	{Attrition=Yes}	0.01156463	0.8947368	0.01292517	5.549634	17
[4]	{NumCompaniesWorked=1, YearsAtCompany=1, JobRole_Research Scientist=No, MaritalStatus_Single=Yes}	=>	{Attrition=Yes}	0.01156463	0.8947368	0.01292517	5.549634	17
[5]	{YearsInCurrentRole=0, BusinessTravel_Travel_Frequently=Yes, EducationField_Medical=No, JobRole_Sales Executive=No}	=>	{Attrition=Yes}	0.01156463	0.8947368	0.01292517	5.549634	17
[6]	{JobLevel=1, YearsInCurrentRole=0, BusinessTravel_Travel_Frequently=Yes, EducationField_Medical=No}	=>	{Attrition=Yes}	0.01088435	0.8888889	0.01224490	5.513361	16
[7]	{JobLevel=1, PerformanceRating=3, YearsInCurrentRole=0, BusinessTravel_Travel_Frequently=Yes}	=>	{Attrition=Yes}	0.01088435	0.8888889	0.01224490	5.513361	16

The second model generated was made to force the result to have the attrition value=yes, but through this result, we realization that there were too many field/role/department values that directly related to each other, and for the association rules, we decided to view the employees purely as employees without association to their department. We then revisited the preprocessing stage to remove these columns to get a better general result.



```
     lhs                                                rhs                support confidence   coverage     lift count
[1]  {JobLevel=1,
      PerformanceRating=3,
      YearsInCurrentRole=0,
      YearsWithCurrManager=0,
      BusinessTravel_Travel_Frequently=Yes}  => {Attrition=Yes} 0.01088435  0.9411765 0.01156463 5.837677    16
[2]  {JobLevel=1,
      PerformanceRating=3,
      YearsInCurrentRole=0,
      YearsWithCurrManager=0,
      BusinessTravel_Travel_Frequently=Yes,
      BusinessTravel_Travel_Rarely=No}       => {Attrition=Yes} 0.01088435  0.9411765 0.01156463 5.837677    16
[3]  {JobLevel=1,
      YearsInCurrentRole=0,
      YearsWithCurrManager=0,
      BusinessTravel_Travel_Frequently=Yes,
      MaritalStatus_Married=No}              => {Attrition=Yes} 0.01020408  0.9375000 0.01088435 5.814873    15
[4]  {JobLevel=1,
      YearsInCurrentRole=0,
      YearsWithCurrManager=0,
      BusinessTravel_Travel_Frequently=Yes,
      BusinessTravel_Travel_Rarely=No,
      MaritalStatus_Married=No}              => {Attrition=Yes} 0.01020408  0.9375000 0.01088435 5.814873    15
[5]  {JobLevel=1,
      YearsInCurrentRole=0,
      YearsWithCurrManager=0,
      BusinessTravel_Travel_Frequently=Yes,
      Department_Human Resources=No,
      MaritalStatus_Married=No}              => {Attrition=Yes} 0.01020408  0.9375000 0.01088435 5.814873    15
```

The final model produced better results that correlated more fields that were not negatively correlated, ie department_sales=yes, department_HR=no. Although business travel was still kept due to the variables being split into 3 groups, and should not be removed as it was a major factor that was identified with the other models. The rules generated were cross referenced to the best performing model(logistic regression), and with the various rules that matched the results from the model, this further reinforced the factors that positively and negatively affected employee attrition.

Model choice and recommendations

Scenario:

The company is trying to implement measures to reduce employee attrition rate, but are unsure on what direction they should take. Based on model performance, what predictors should be looked into?

Criteria for selection:

How well can a model accurately predict the attrition value, taking into account Accuracy, specificity, sensitivity, RMSE, and AUC.

For the chosen model, are the predictors chosen for the model accurate and insightful?

Justifications:

Based on the result of each variable, we are able to determine which model has the best performance. After removing the highly correlated predictors, and cross referencing it to the association rules, should give us an accurate understanding of the business situation and reinforce the choices that we choose to undertake given the findings of the model.

Model	Accuracy	AUC	Specificity	Sensitivity	RMSE	MAE
Logistic Regression	0.865	0.82	0.9553	0.3871	0.215	0.175
Decision Tree	0.8418	-	0.3434	0.9427	0.4012452	0.160997
Neural Network	0.8228	0.71	0.2340	0.9351	0.4209	0.1772

Performance of models & factors affecting choice of model

Justifications: Taking into account accuracy, specificity, and sensitivity, the logistic regression model outperformed the neural network in all aspects. Where the model is important in identifying the cases correctly, it is however unable to give a good enough prediction for the employees that are likely to quit. High false positive rate, on an already small percentage of testing data would lead to irrelevant results when identifying actual employees that may want to quit. The decision tree was also outperformed by the logistic regression model, and hence our final choice was to proceed with the logistic regression model's insights to determine the methods to address employee attrition.

Business Recommendations from the logistic regression model, cross referencing the association rules, further confirmed that we were proceeding in the right direction.

Business recommendations based on chosen model's insights

Improving Environment Satisfaction: Comfortable work environment, hot desking, and recreational spaces.

Lowering Frequency of Business Travel: Evenly allocating business travel across employees.

Improving Job Involvement: Ensuring that all team members are engaged in a project.

Improving Job Satisfaction: Rewarding and recognition of employees through awards, bonuses and incentives.

Lowering overtime hours: Ensuring that employees are not forced to work overtime, and optimally splitting workloads to prevent unnecessary overtime.

Increasing years in current role: Offering promotions to deserving candidates to keep them in similar roles, avoid reassignment of employees to new roles

Increasing years with current manager: Keep employees under the same supervisors/managers if possible, avoid frequent switching of supervisors

Increasing work life balance: Create opportunity to engage the employee's family, or to give employees more personal. Building platforms for employees to present their extra-curricular skills, or pursue a hobby alongside their formal work.

For factors outside of sphere of control:

High number of companies worked at: Ensuring that the candidate does not have high turnover rate in previous companies while recruiting

High distance from home to workplace: Offer relocation stipend to employee

Marital status: Single, employees that are single are more likely to quit if the other factors listed above are not met, to pay more attention to these employees if their retention is important.