**GRoup - 01**

**Laxmikanta Gauda**

**Ambili Jayachandran**

**Maya Vasudevan K**

**Tarun Kamal Khiani**

**M Sarath**

**M Sarath tarun kamal khiani**

Data Mining

Assignment – HR Employee Attrition Prediction

Contents

[Goal 2](#_Toc485704465)

[Solution 2](#_Toc485704466)

[Exploratory Data Analysis 2](#_Toc485704467)

[Hypothesis 5](#_Toc485704468)

[Random forest Model 6](#_Toc485704469)

[Neural Network model 8](#_Toc485704470)

[ensemble model 9](#_Toc485704471)

[Conclusion 10](#_Toc485704472)

# Goal

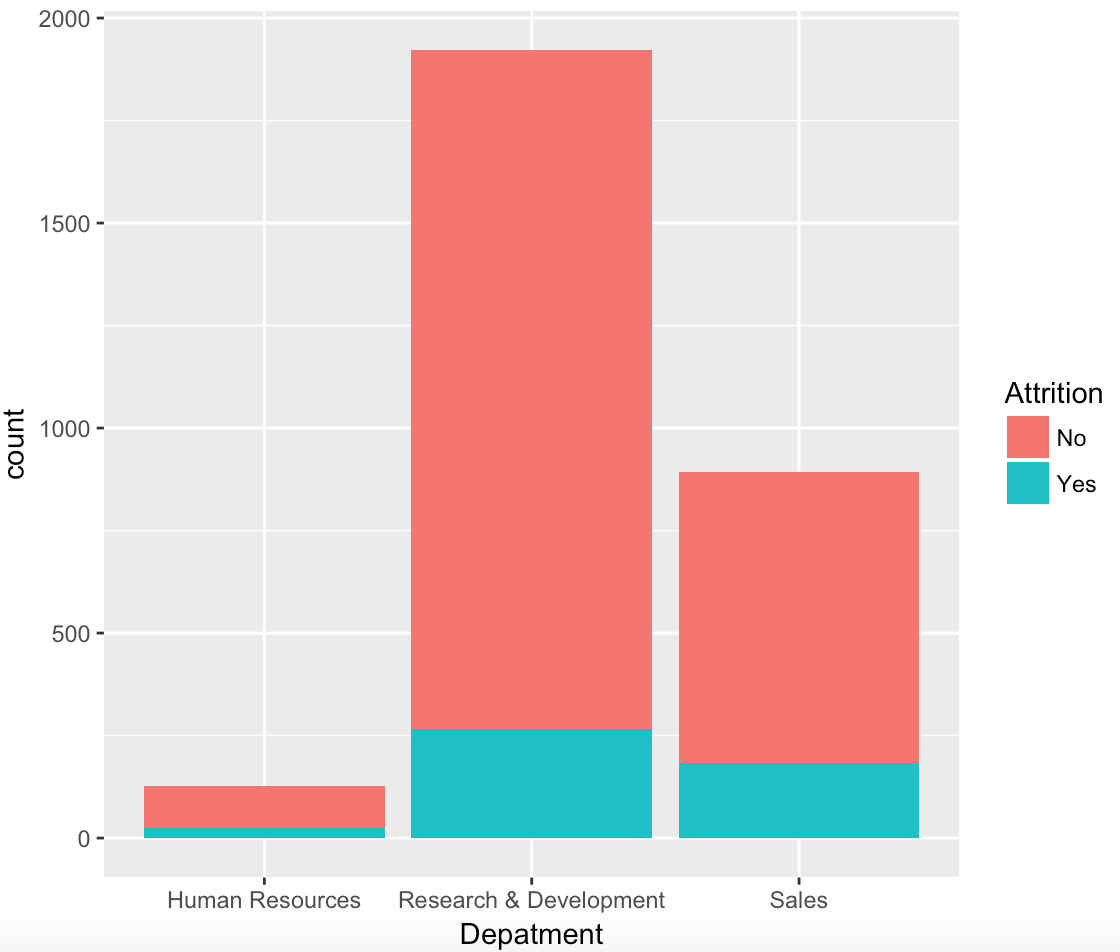
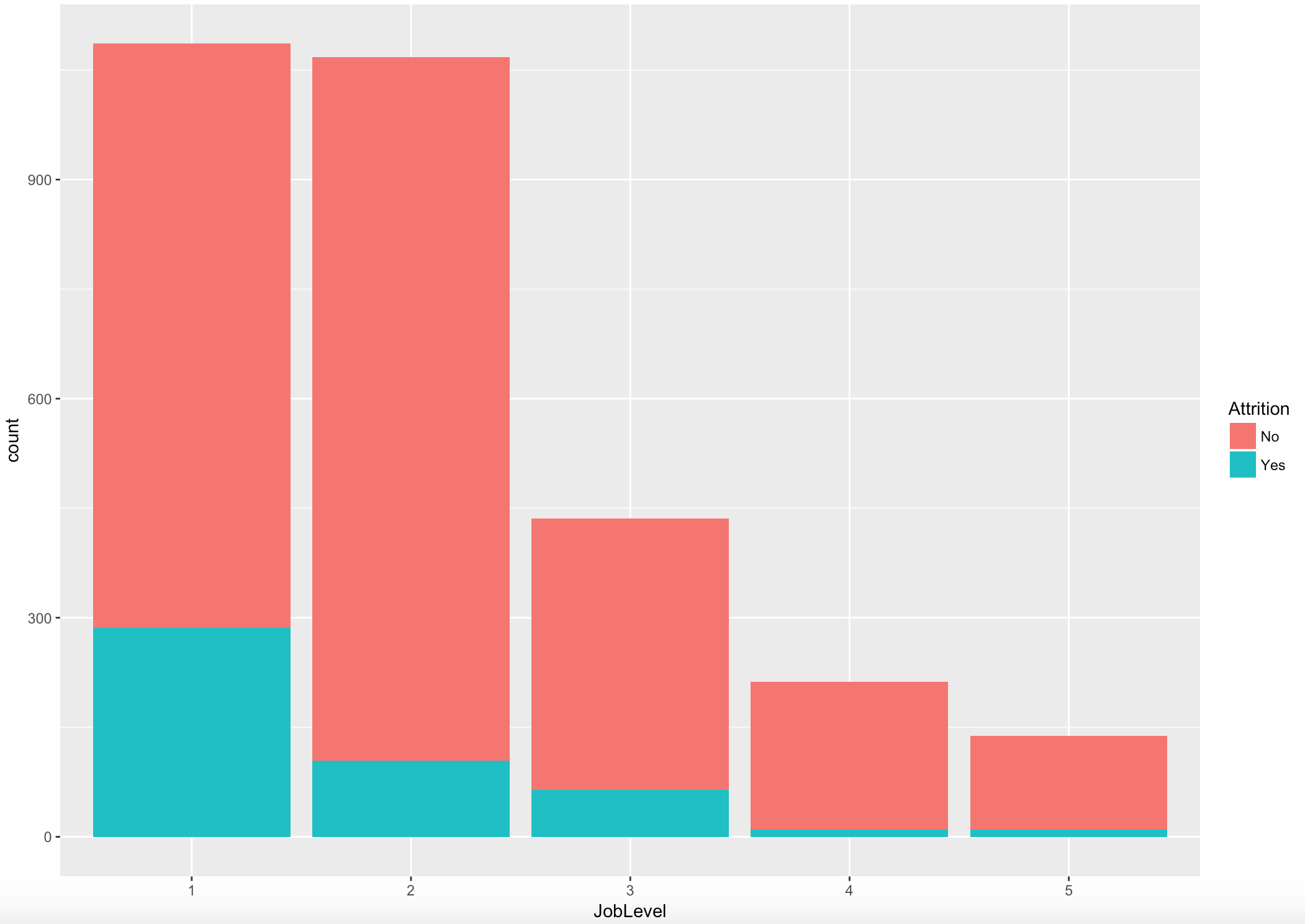
The goal of this assignment is to build a model to predict if an employee will leave the company or not.

# Solution

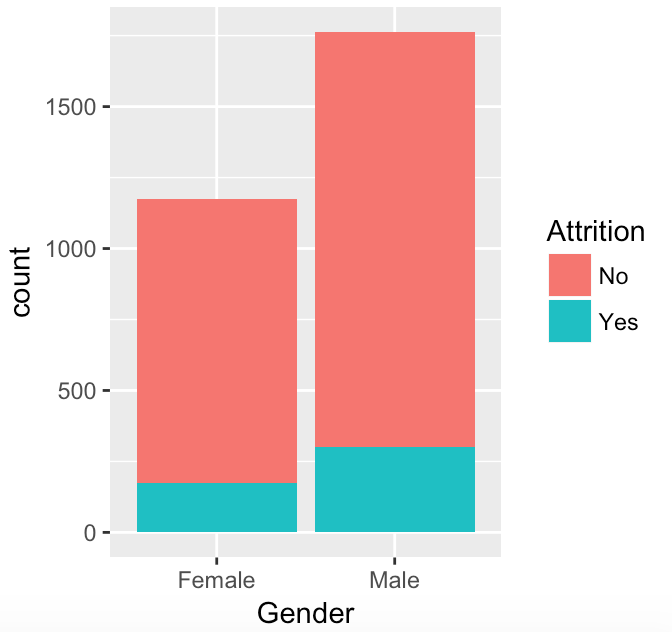
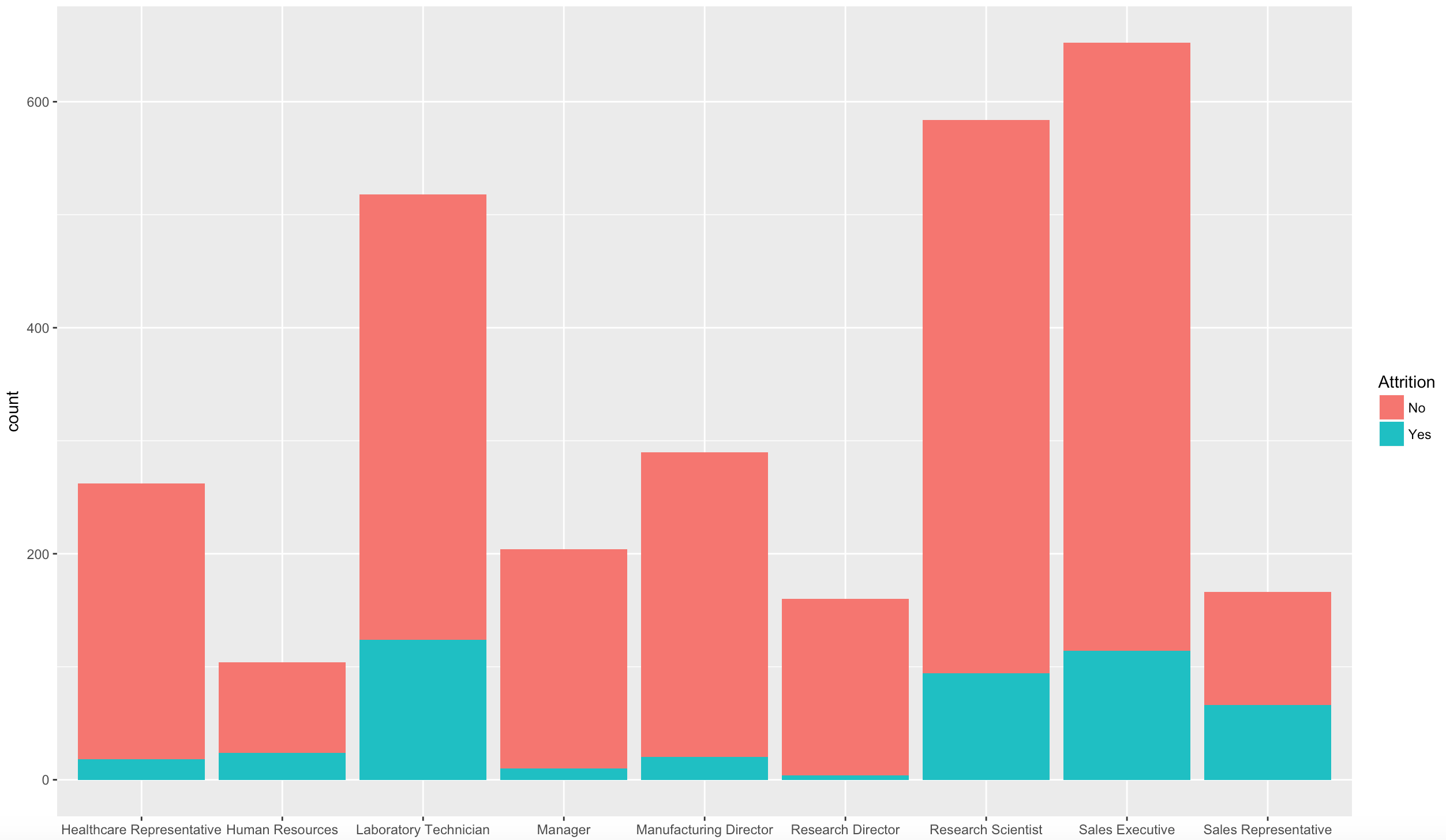
## Exploratory Data Analysis

Important Observations:

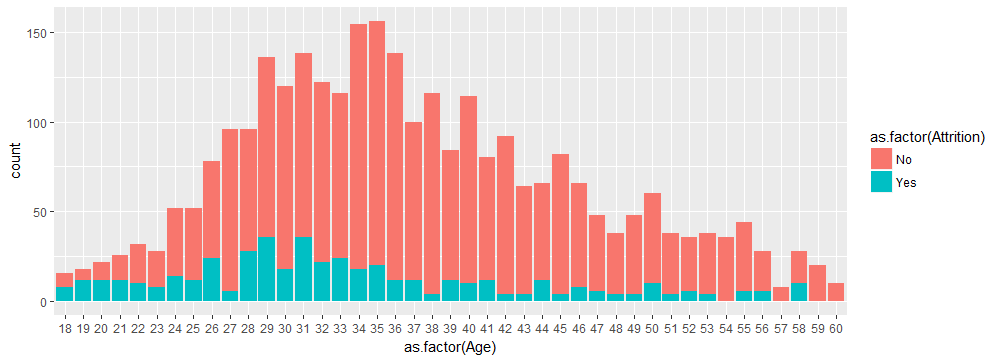
* The data quality is good as there are no null or empty fields.
* EmployeeCount, StandardHours, Over18 are same across the data set hence not needed for classification.
* Employee Number is unique for every employee and hence not needed for classification.
* Attrition Rate is 16.12% which is good for classification hence there is no need to oversample or under sample.
* Attrition is higher for Job Level – 1, Department – Research & Development



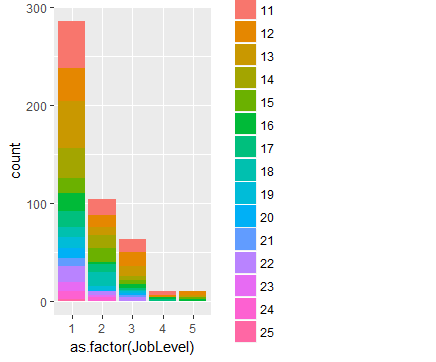
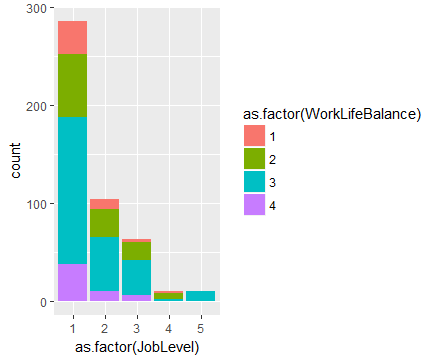
* Attrition is higher for Job Role – Laboratory Technician, Research Scientist, Sales Executive and Sales Representative & is not dependent on Gender



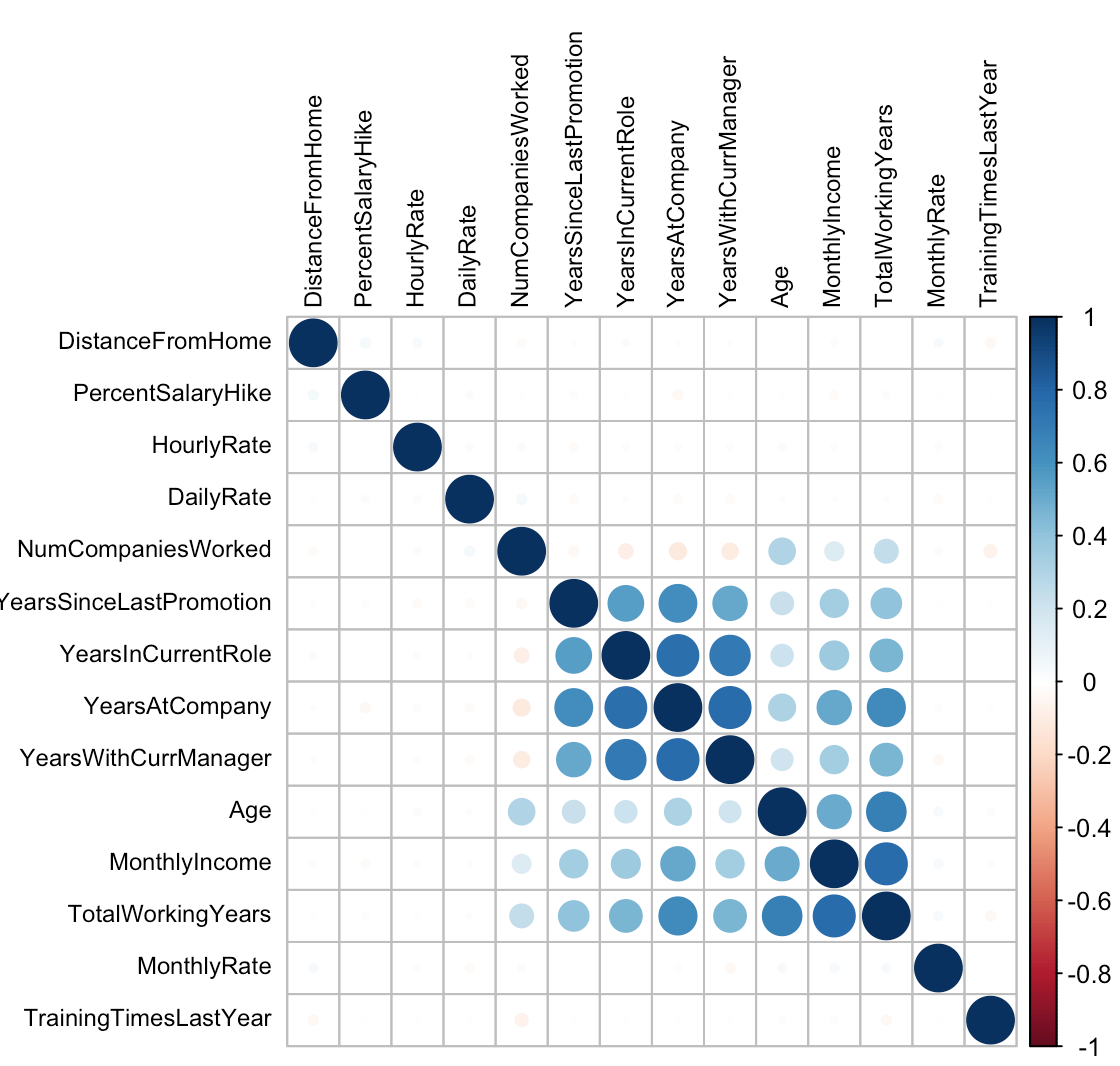
* Attrition is higher for Ages group 20 – 30



* Attrition is higher when the employees have given a lower rating for work life balance (3 & 4) & lower Percent Salary Hike



* Variances of the indices are widely different. Hence scaling of variables is necessary for further analysis.
* From the correlation plot, it is evident that some of the variables are strongly correlated. Several pairwise correlations are over 75%.
  + Total Working Years - Monthly Income, Age, Years at Company
  + Years at Company - Years with Current Manager, Years in Current Role
  + Years in Current Role - Years with Current Manager



* Now the split the data into development and hold out to build and test our classification models.

## Hypothesis

The study was designed as an exploratory project to understand the employee’s data and to evaluate existing methods for analysing employee attrition. The primary objective is to build a model that can predict whether an employee will leave the company or not.

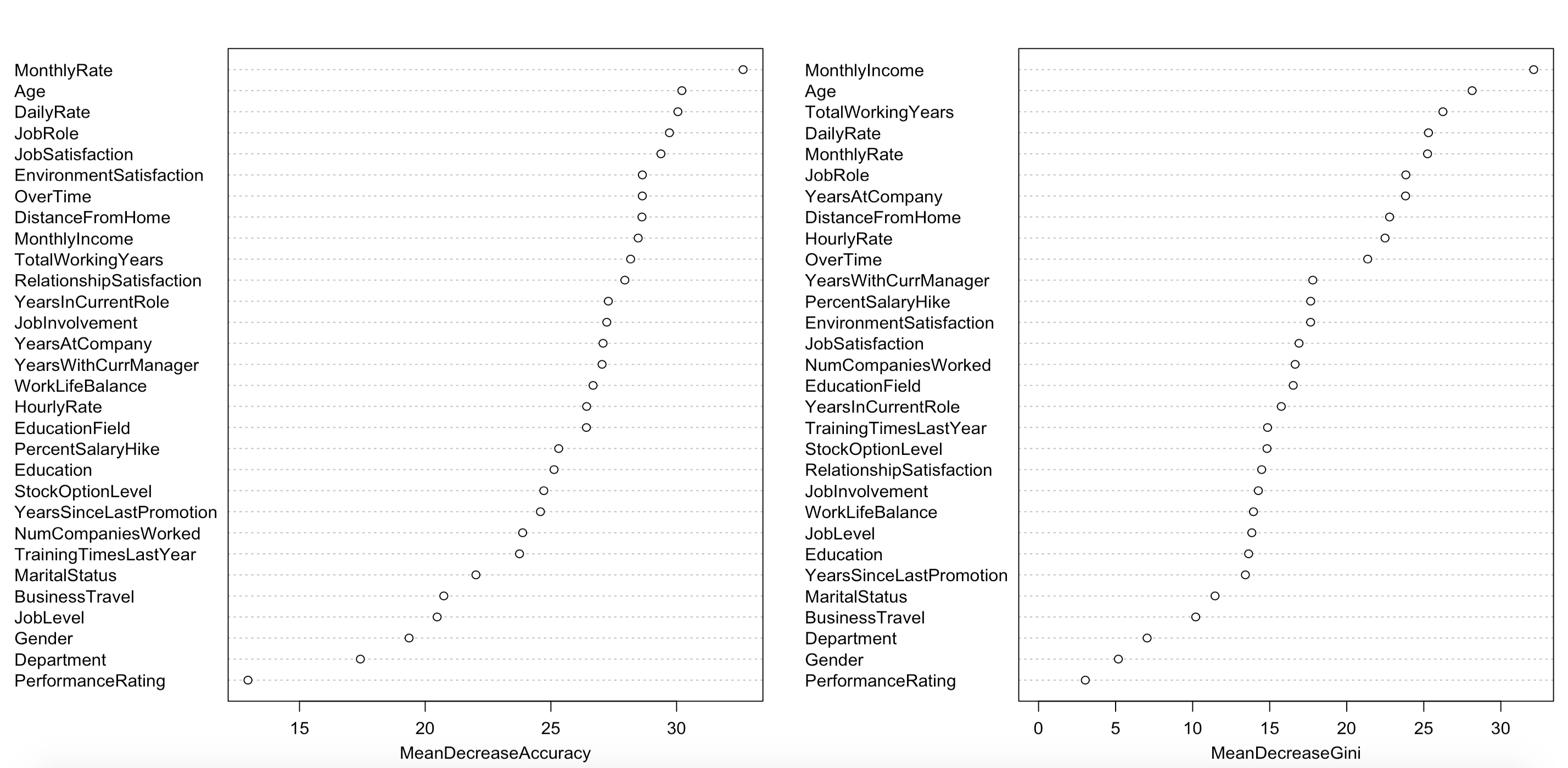
The case study is guided by the following hypotheses:

1. The variables included in the study such as monthly rate, age, job role, daily rate etc are significant measures of employee attrition.
2. The external factors such as market conditions and peer pressure does not affect employee attrition significantly.

## Random forest Model

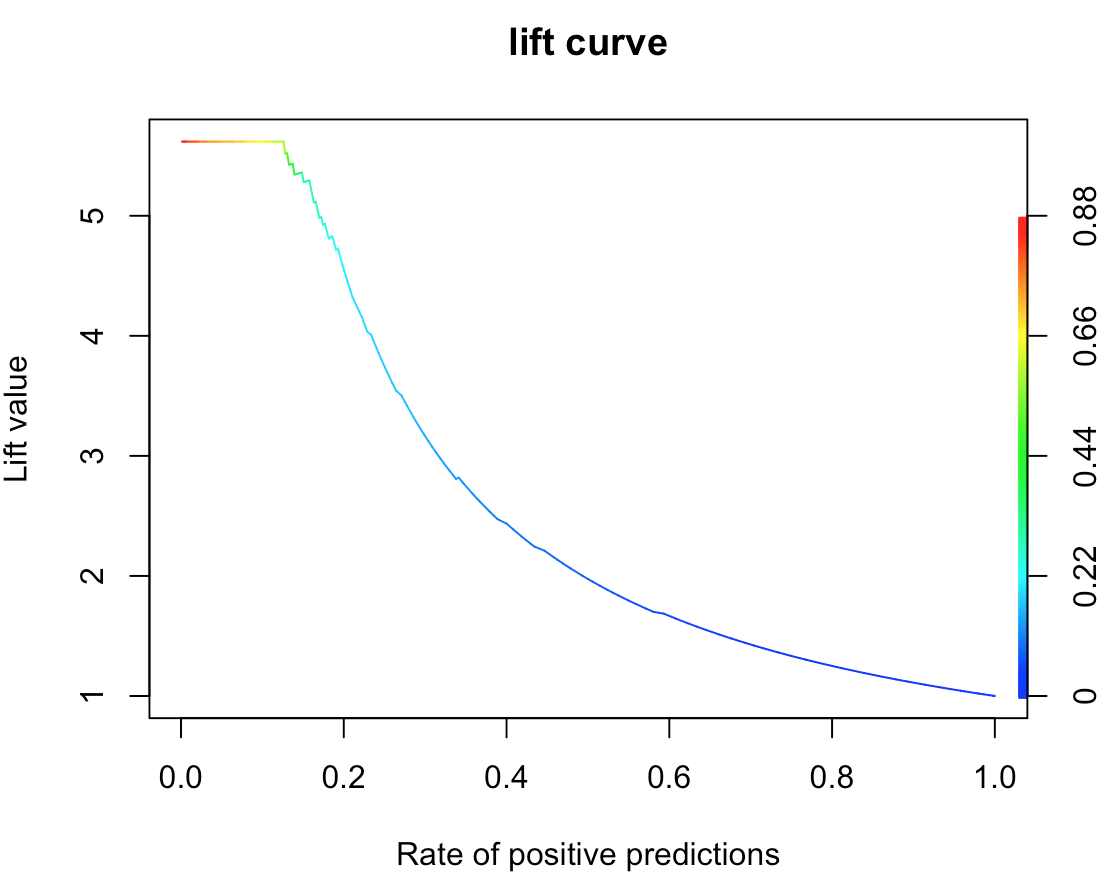
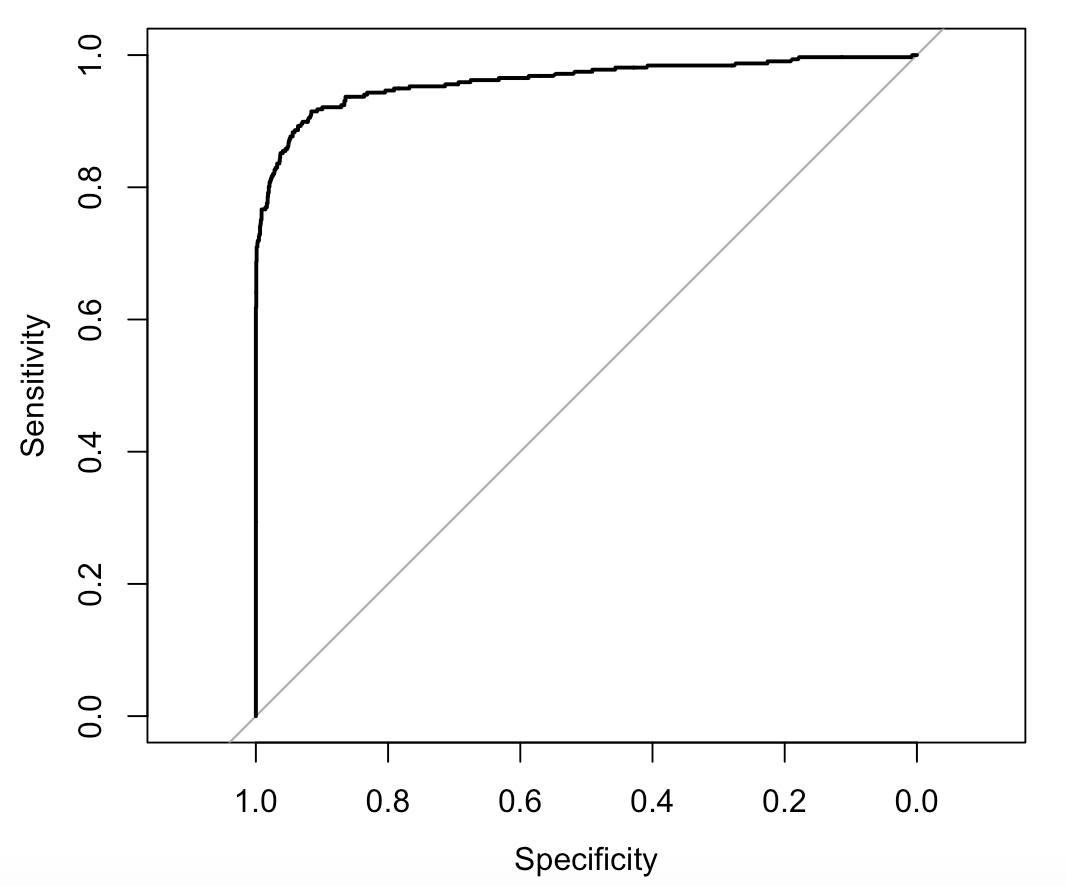
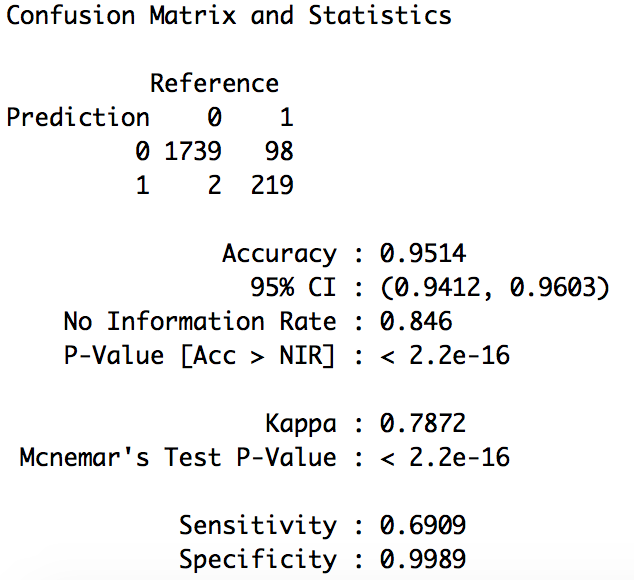
The tuned random forest model forest can be based on 2 randomly selected variables and 300 trees.

It suggests the importance of the variables Monthly Rate, Age, Daily Rate, Job Role are important classifiers. For more details see the image below.



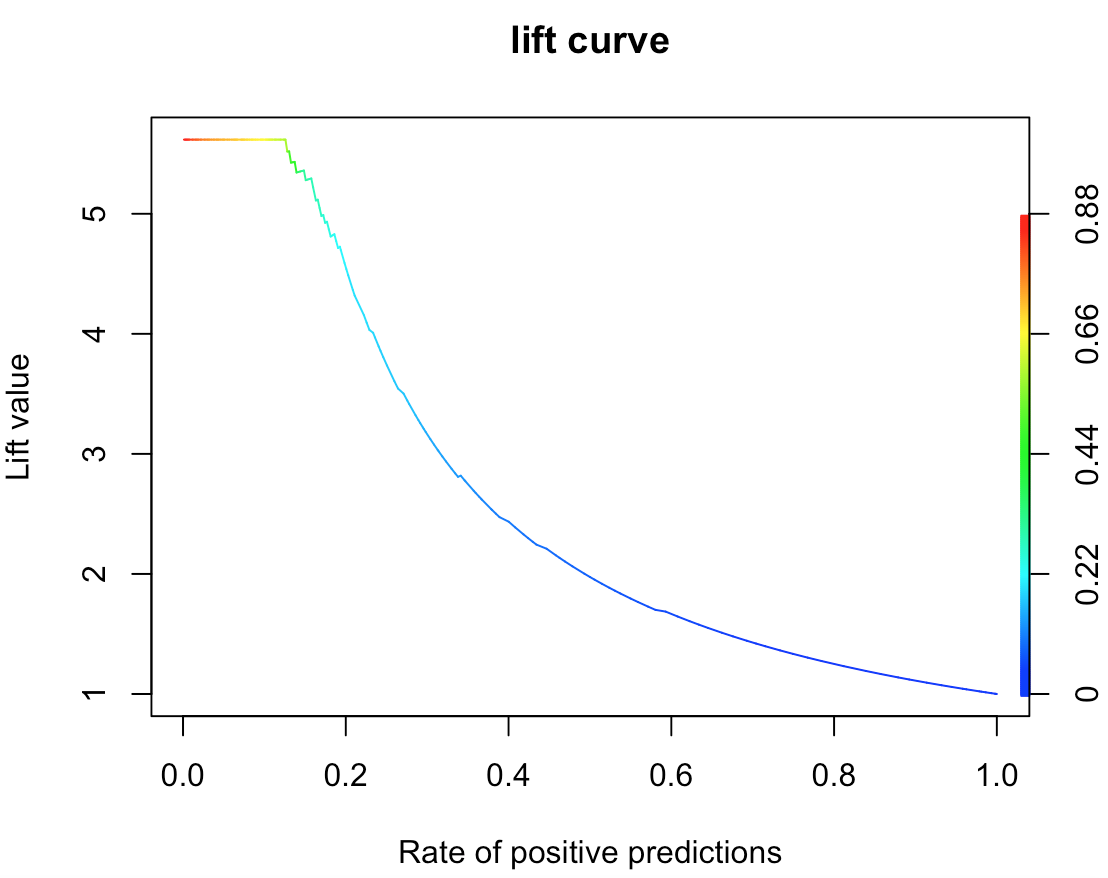
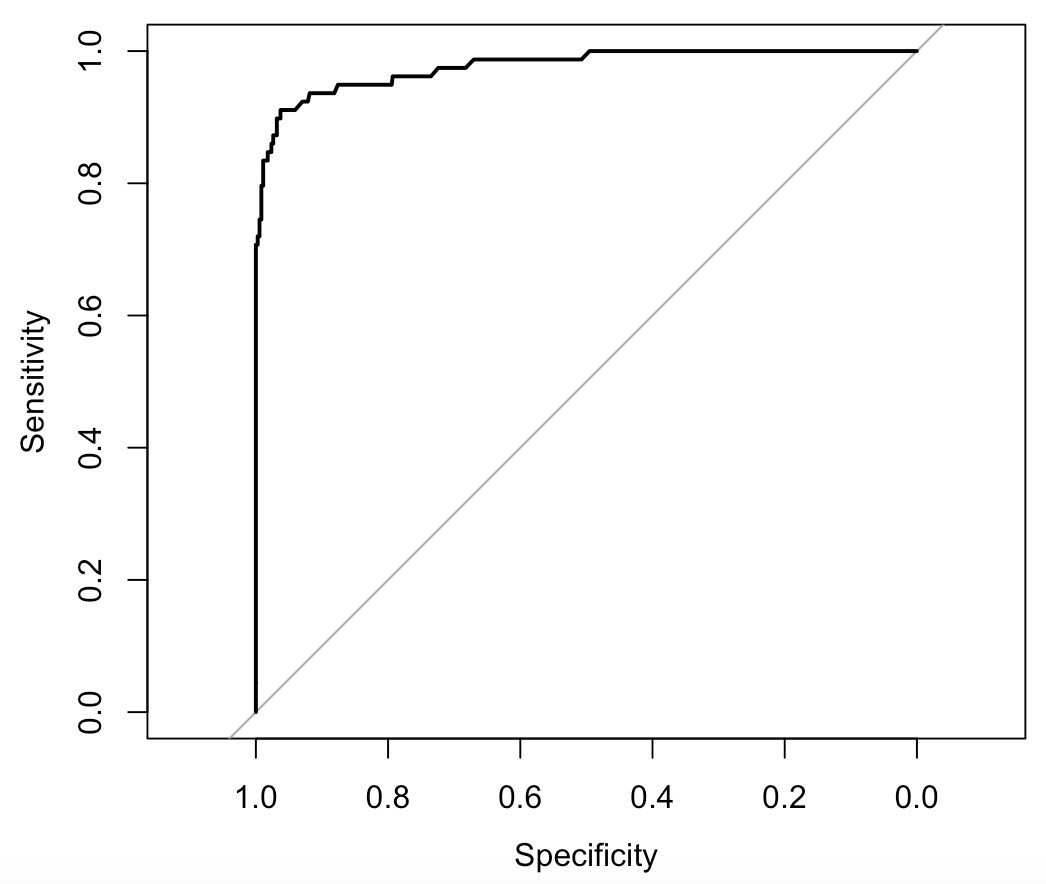
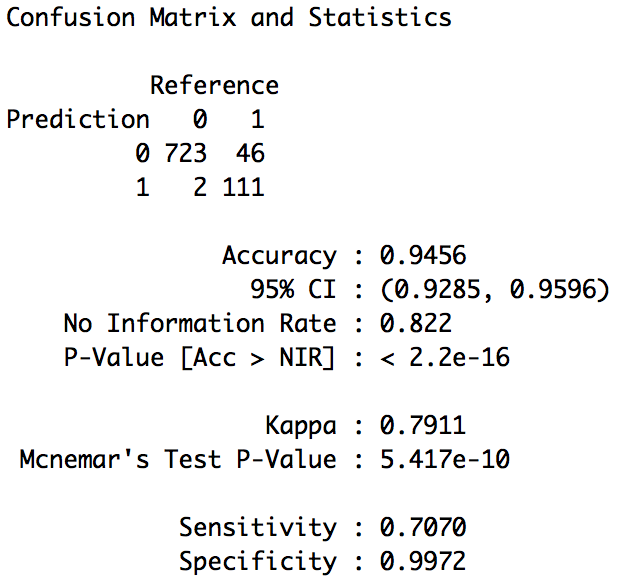
**Classification Accuracy on the development sample**

The Classification Accuracy is 95.14%, Sensitivity is 69.09%, Specificity is 99.89%, AUC is 96.13%, K-S is 82.66% and the first two deciles captures 100% of the candidates. This can be a concern of overfitting.



**Classification Accuracy on the hold-out sample**

The Classification Accuracy is 94.56%, Sensitivity is 70.70%, Specificity is 99.72%, AUC is 97.69%.

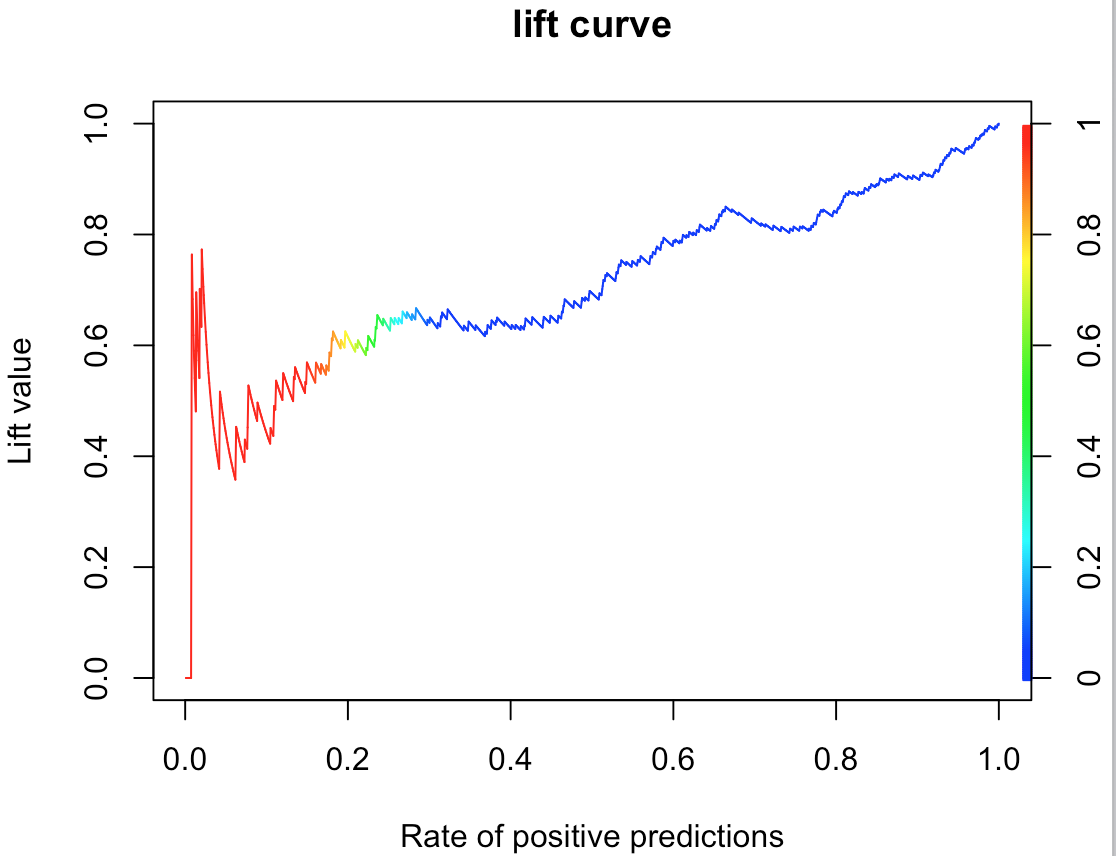
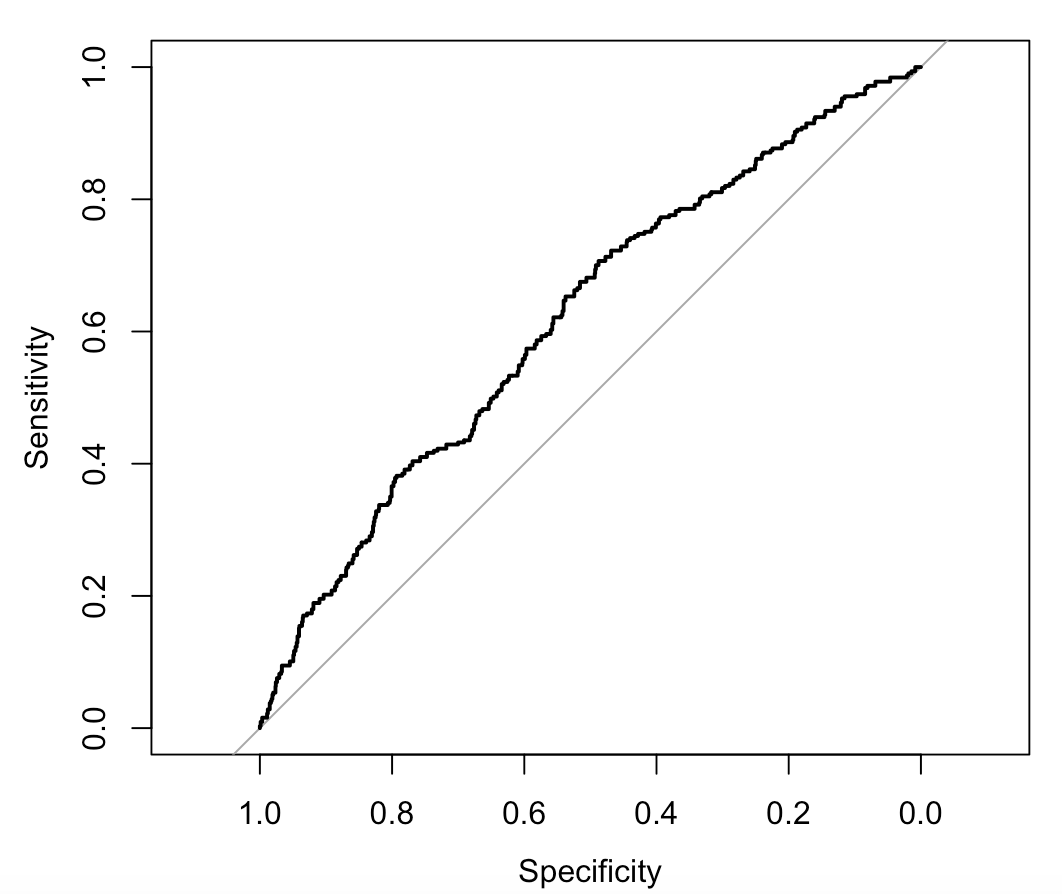
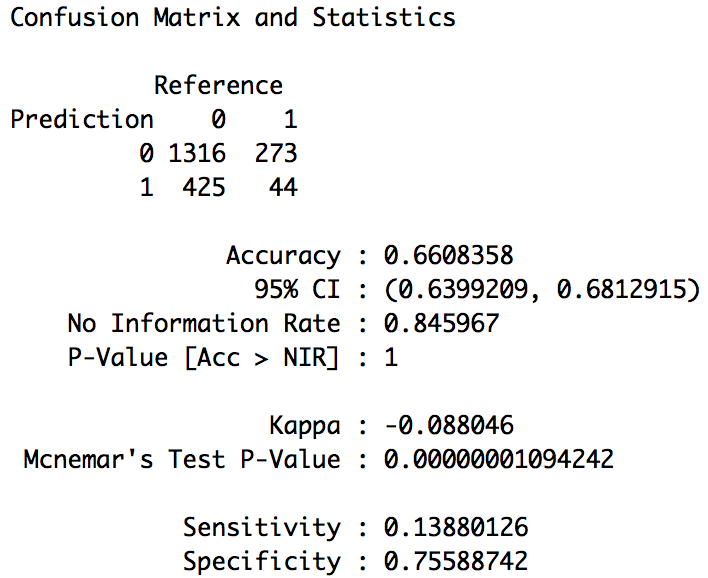


As the RF models scales extremely well on the hold out sample we do not need to worry about model overfitting.

## Neural Network model

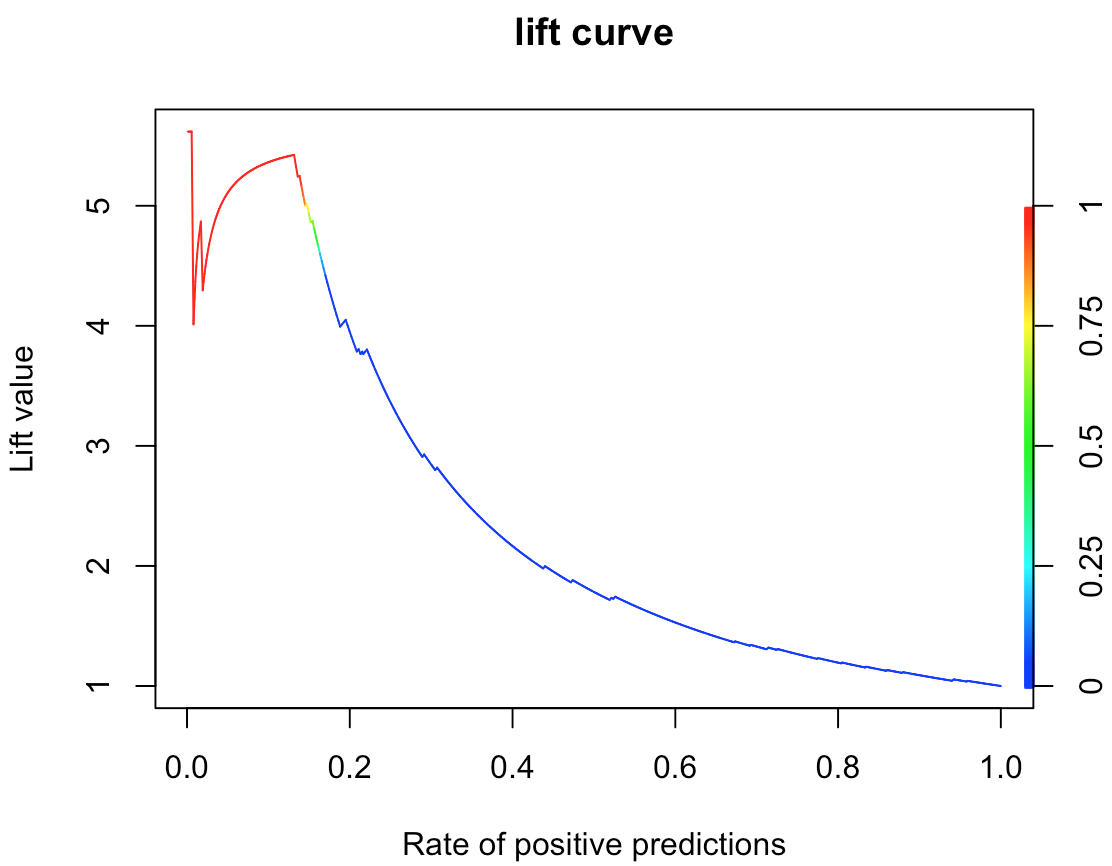
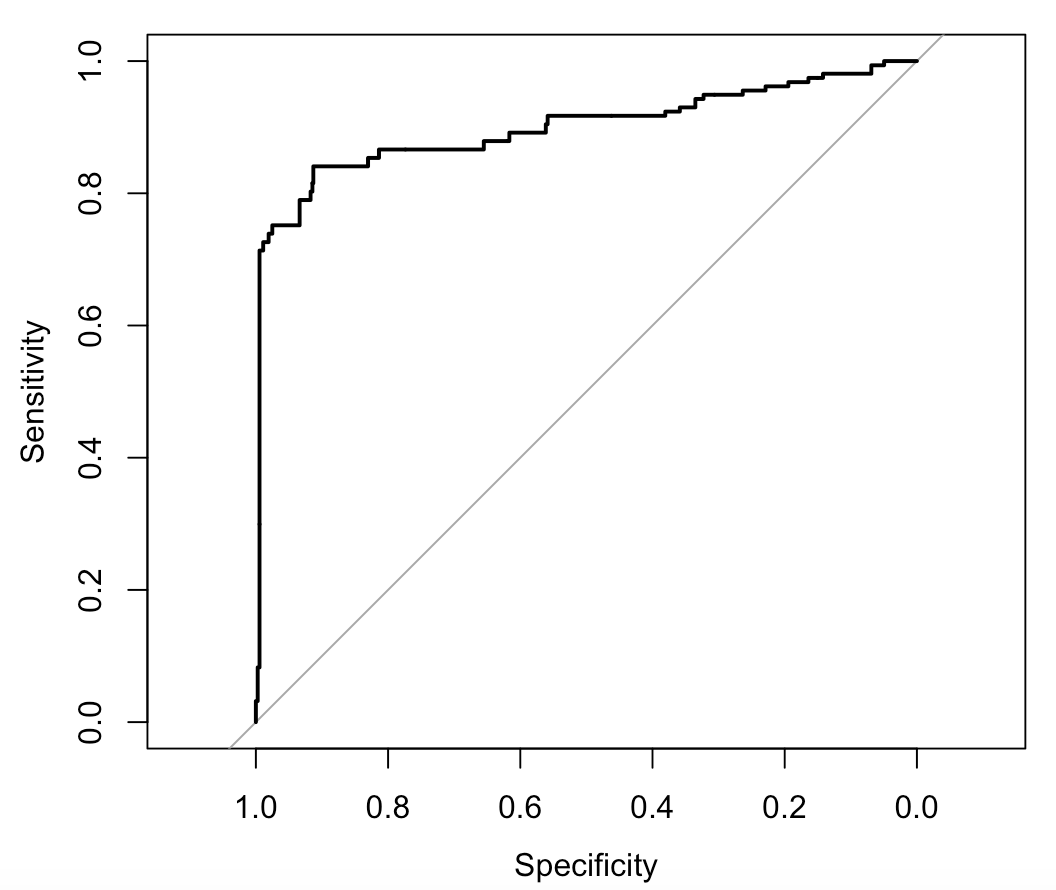
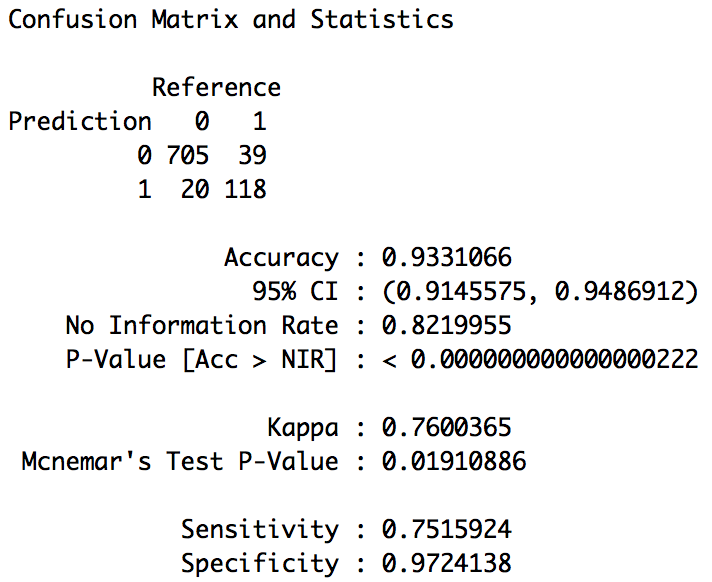
**Classification Accuracy on the development sample**

The Classification Accuracy is 65.50%, Sensitivity is 18.92%, Specificity is 73.98%, AUC is 59.95%. This is an extremely poor model.



**Classification Accuracy on the hold-out sample**

The Classification Accuracy is 93.31%, Sensitivity is 75.15%, Specificity is 97.24%, AUC is 89.95%.



The neural network model does not have a good classification accuracy for the development sample but performs well on the hold out sample.

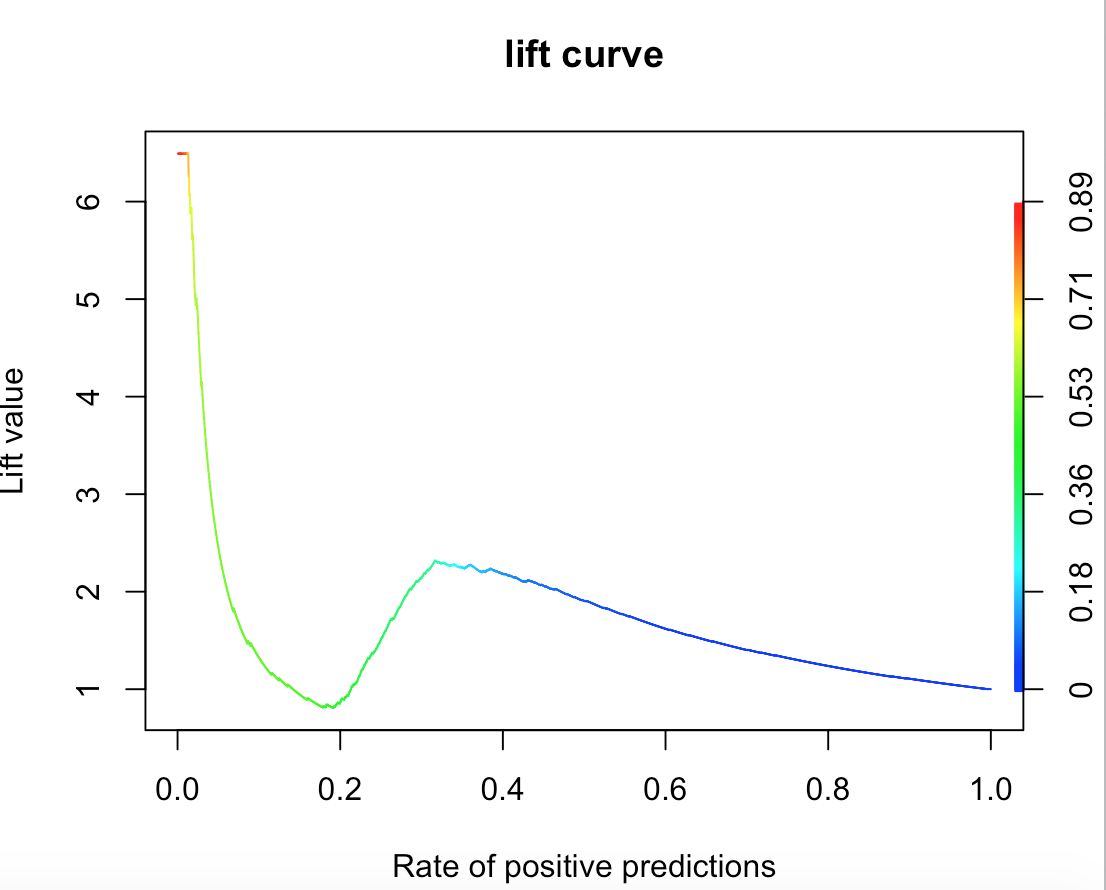
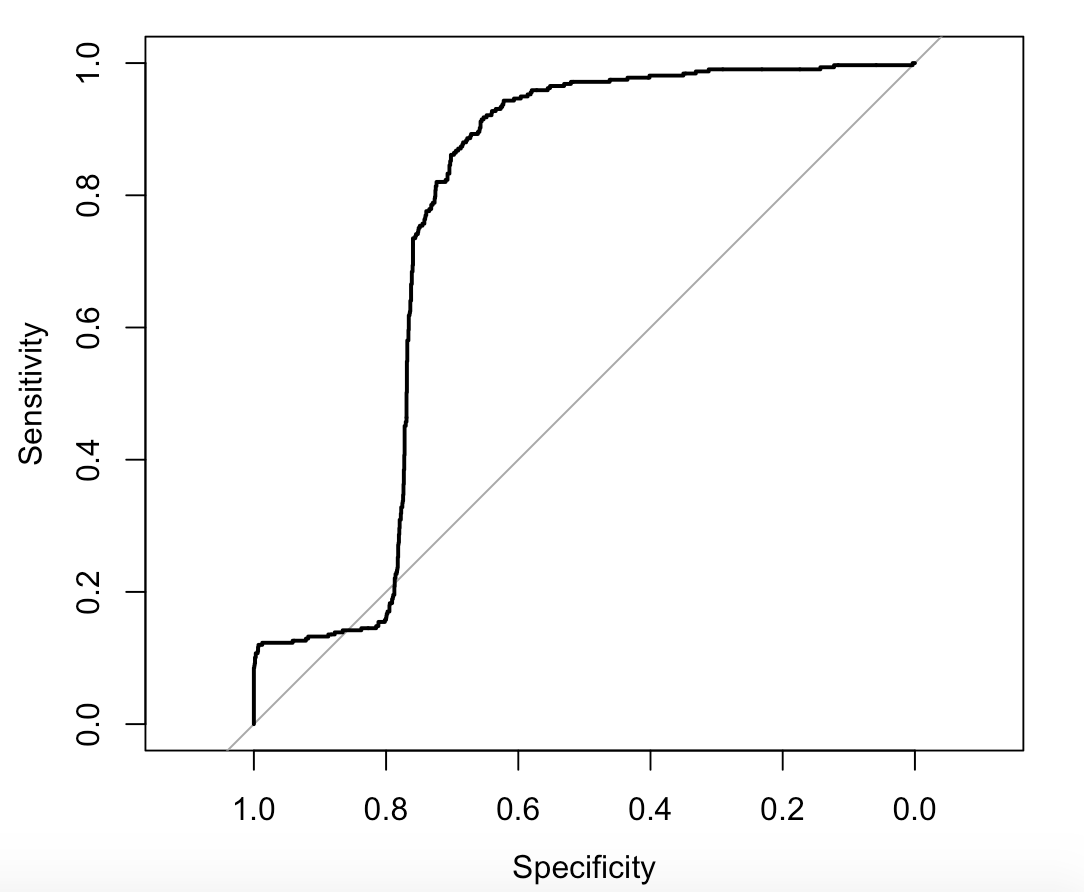
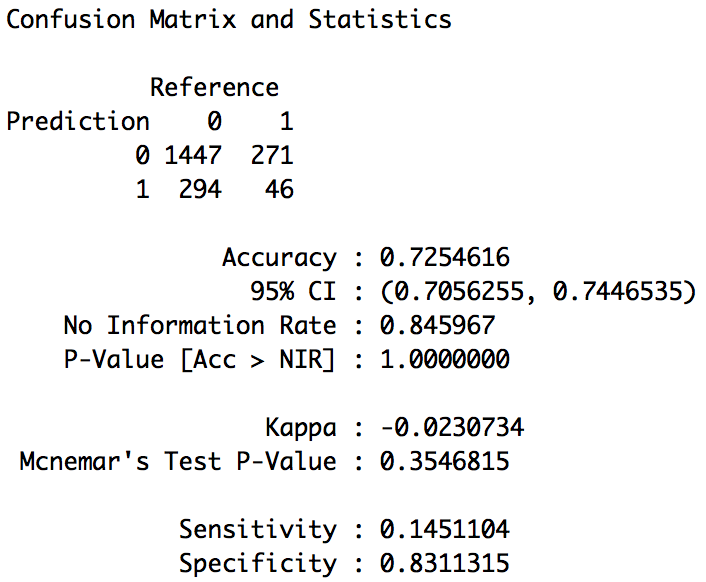
.

## ensemble model

In the ensemble model, we take the average of the predicted probability from the random forest model and the neural network.

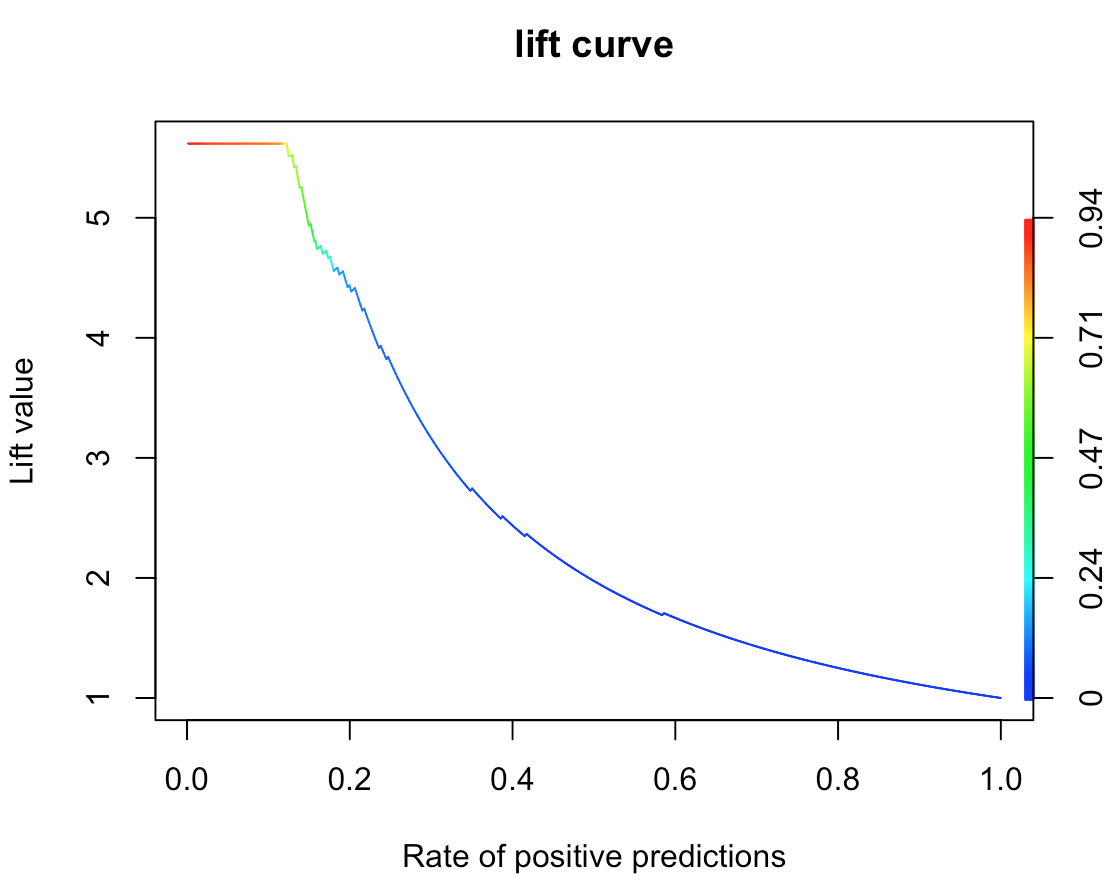
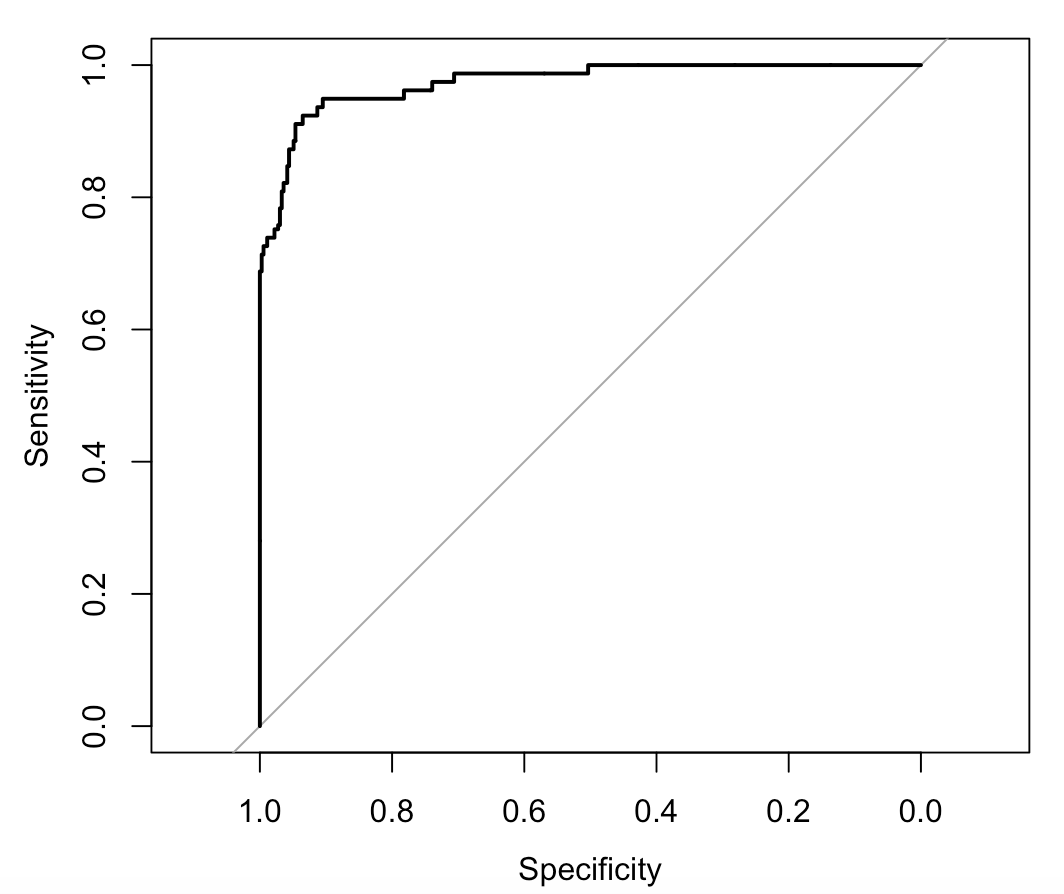
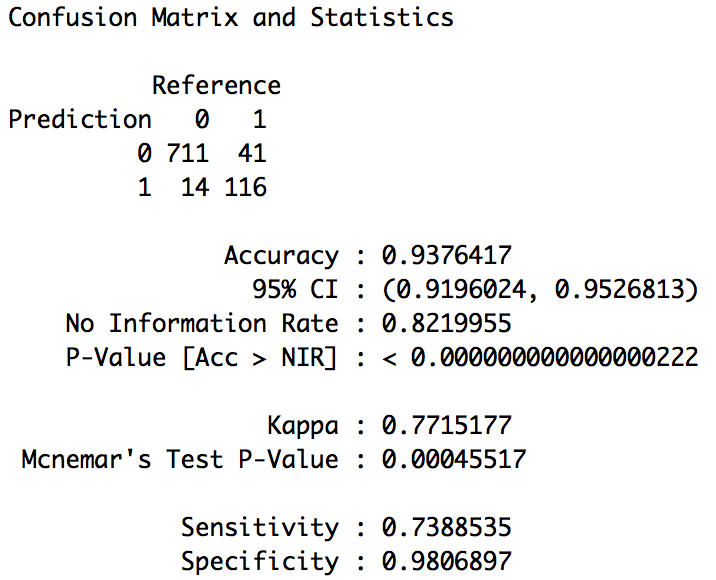
**Classification Accuracy on the development sample**

The Classification Accuracy is 65.50%, Sensitivity is 18.92%, Specificity is 73.98%, AUC is 76.87%. The model performance is average on the development sample.



**Classification Accuracy on the hold out sample**

The Classification Accuracy is 93.76%, Sensitivity is 73.88%, Specificity is 98.06%, AUC is 97.36%.



The ensemble model scales well on the hold out sample.

# Conclusion

**Classification Accuracy on the development sample**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Random Forest | 95.14% | 69.09% | 99.89% | 96.13% |
| Neural Network | 65.50% | 18.92% | 73.98% | 59.95% |
| Ensemble Model | 65.50% | 18.92% | 73.98% | 76.87% |

**Classification Accuracy on the hold out sample**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Random Forest | 94.56% | 70.70% | 99.72% | 97.69% |
| Neural Network | 93.31% | 75.15% | 97.24% | 89.95% |
| Ensemble Model | 93.76% | 73.88% | 98.06% | 97.36% |

The random forest has the best classification performance in the development sample.

All the three models perform well on the hold out sample.

As our intention is to identify attrition and need to ensure a high Sensitivity & AUC we recommend the **ensemble model** over the random forest or the neural network.