SUPERVISED LEARNING

HR DATASET

1. **Abstract:**

The aim of this report is to demonstrate how to apply a supervised learning algorithm on a given dataset, using different classification models, comparing their accuracy, and determining which explanatory variables impact the response variable the most.

A thorough exploratory data analysis has been performed at first in order to understand the dataset and the different relationships the independent variables might have with the dependent variable.

This process was processed without setting a specific route, meaning no variables have been hand-picked in order to understand how the dataset is performing, instead the more data exploration analysis is performed the more we start asking questions, which leads into having interesting insights or intuitions about the dataset.

The classification models that have been applied to our dataset are logistic regression, linear discriminant analysis, classification tree and finally random forest.

The selected dataset is a Human Resources one, where we aim to observe the ‘left job’ as the response variable and analyze what might be the factors that could mainly lead to an employee deciding to leave the company.

2. **Data Description:**

The dataset is composed of 10 variables which are the following:

1) Satisfaction level: 0-1, the closer to 0 the less satisfied the employee is.
2) Last evaluation: 0-1, the closer to 0 the lower the company's evaluation of the employee.
3) Number project (the number of projects a given employee is/was handling)
4) Average monthly hours
5) Time spend company (how many years a given employee has been working for the company)
6) Work accidents (0 for no and 1 for yes)
7) promotions last 5 years (0 for no and 1 for yes)
8) dept (Sales, Technical, Support, IT, product management)
9) Salary (Low, Medium, High)

10) Left (To leave or not to leave) → Response variable

Firstly, the dataset has been checked to see if it suffers from any na/null values and there was none. Then some adjustments have been made on the variables to make the dataset more presentable and easier to understand. For instance, the variable ‘left’ has been renamed ‘left job’. Also, some variables have been changed to factor type with values Yes & No, such as ‘left job’, ‘promotion last 5 years’ and ‘work accident’.

3. Exploratory Data Analysis:

This process was divided into four main parts. First a general data exploration has been performed, where we observe the variables and mainly focus on exploring the response variable.

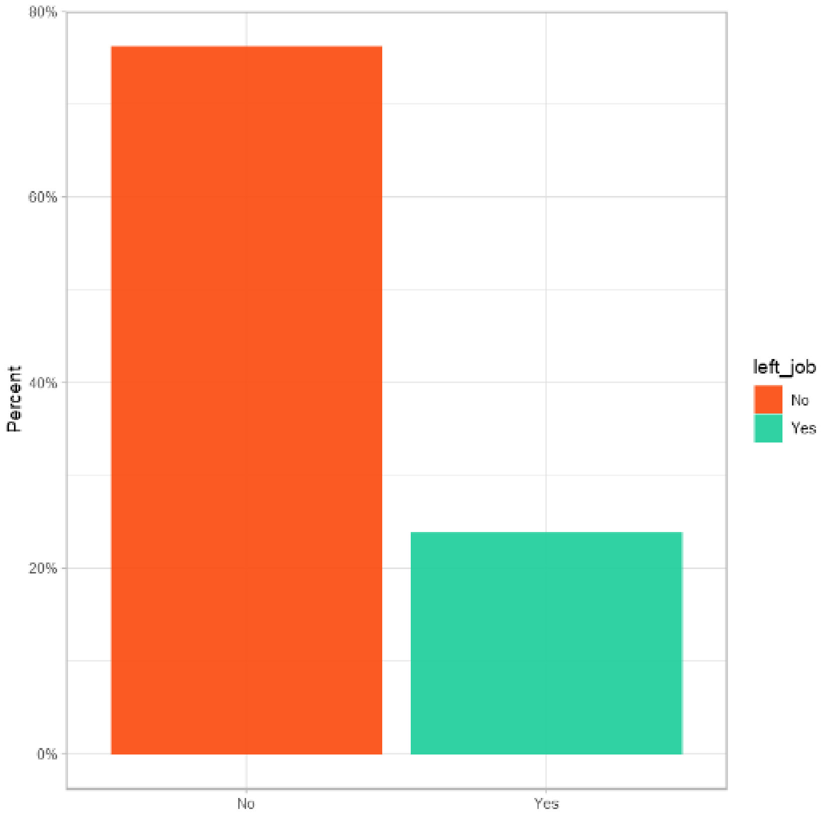


Figure 1

Figure 1 shows how the dataset is mainly distributed, where around 77% of the employees didn’t leave their job and around 23% decided to. This distribution displays a moderate imbalance in the dataset; however, this will be discussed more when applying the classification models.

The second part of the data exploration process is observing the categorical variables which are the following, salary, department, promotion last 5 years and work accident.

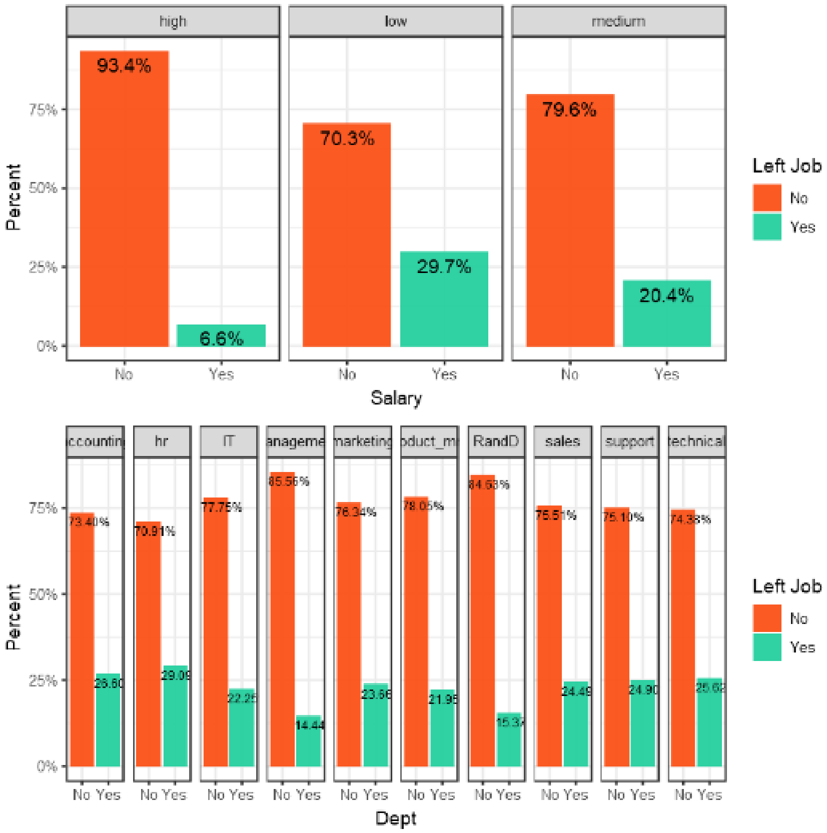


Figure 2

Figure 2 is split into two main plots; the first plot displays the relation between the different types of salaries employees earn and their decision on staying/leaving the company. The output of this plot doesn't show anything out of the ordinary, since normally we would expect employees who earn less to decide to leave the company. The second plot displays the relation between the different types of departments employees work in and their decision on staying/leaving the company. Again, nothing stands out the most from this plot, but it gives an understanding that there's no specific department that has stand-out percentages.

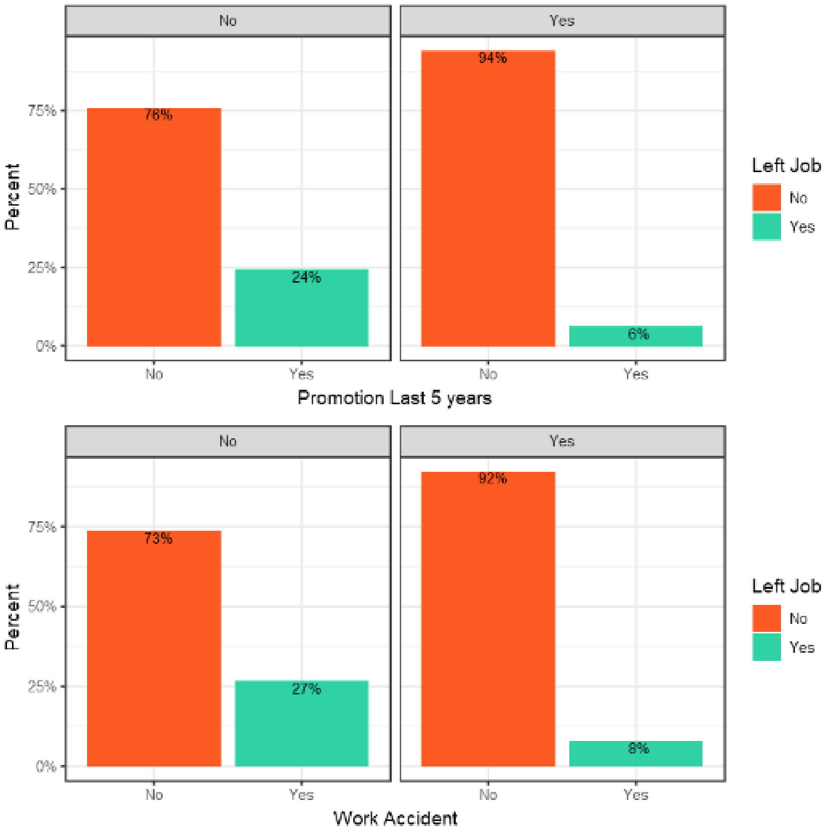


Figure 3

Figure 3 is split into two main plots where the first one displays the relation between employees who got/didn't get promoted in the last 5 years and their decision on staying/leaving the company. As expected, the percentage of resigned employees that didn't get promoted is higher than the percentage of the resigned employees that got promoted. The second plot is somewhat interesting as it shows a higher percentage for employees who left but didn't have any work accidents than the percentage of resigned employees who suffered from an accident in the workplace. An interesting assumption could be that if employees were subject to a work accident, the company would somehow re-imburse them or show them gratitude, which would make the employees more tied to the company as they see how well they've been treated after such incidents.

The third part of the data exploration is observing whether there exist a correlation between the numerical variables which are the following, satisfaction level, last evaluation, number project, average monthly hours, time spend company.

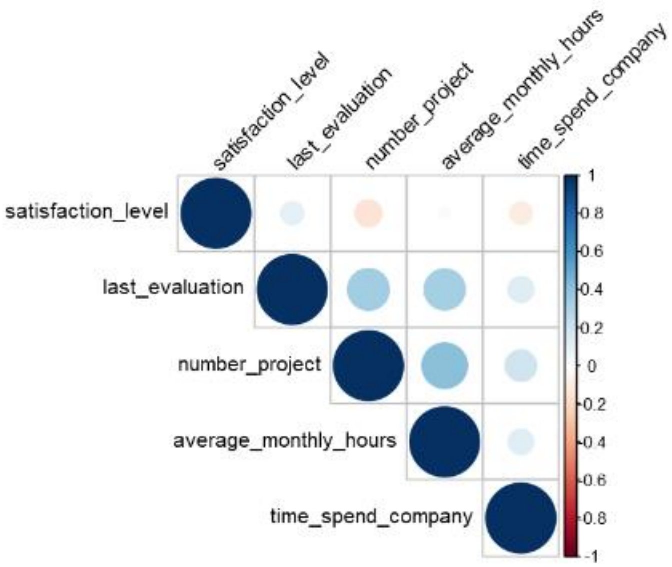


Figure 4

Figure 4 displays a correlogram, we can infer the following from it:

- There is a positive correlation between the last evaluation and the other remaining numerical variables, which is very understandable.
- There is a negative correlation between satisfaction level and (number of projects, average monthly hours and time spend at the company). It's understandable that if the workload increases, the employee will be more dissatisfied (less satisfied in other words). However, what is interesting to note is the negative correlation between time spend at the company and the satisfaction level. This will require a bit more digging to understand why that's the case.

The final part of the exploratory data analysis phase is to observe different relations and start asking questions about the given dataset.

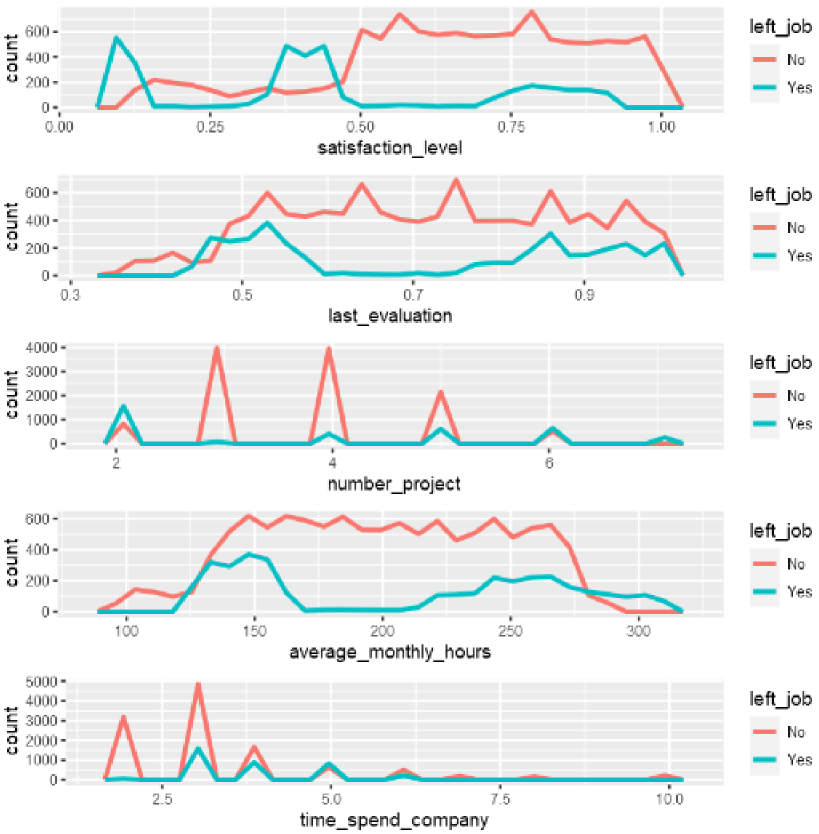


Figure 5

As per our last interpretation, it seems that the more the employee has been working for the company, the less likely they would leave. In the fifth plot in Figure 5, we can notice that the number of employees who quit their job decrease, starting from around year 3, but the decrease is significant after the 5th year. Hence, we can infer that employees are more likely to leave within their first 5 years.

The question that we would be interested in answering would be if employees are more dissatisfied at the beginning of their journey at the company (within 5 years, as per what we have determined from the last point.)

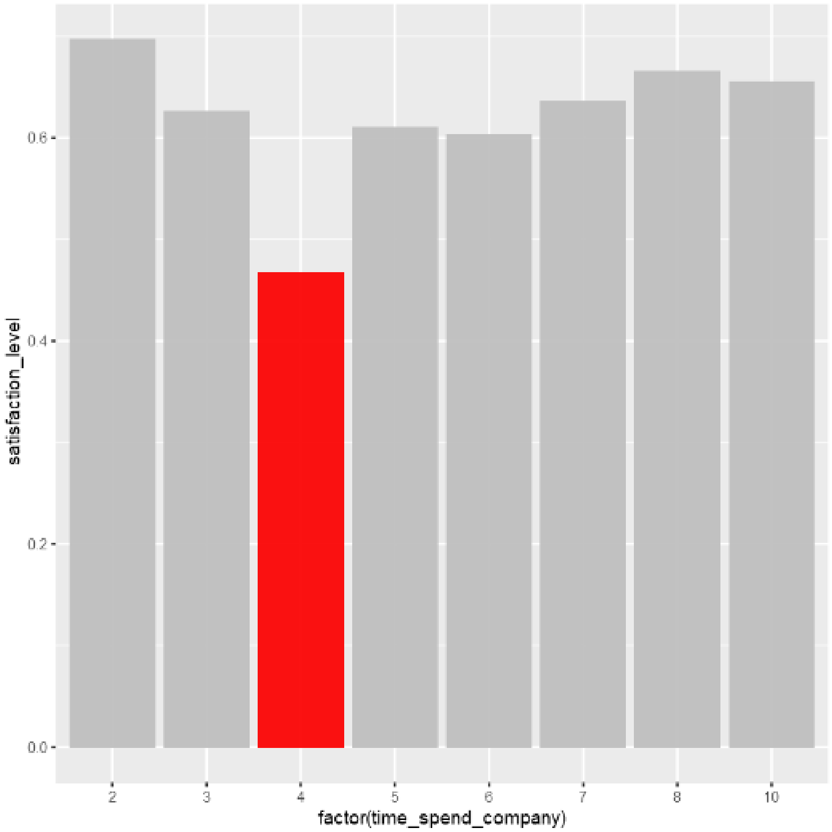


Figure 6

Looking at Figure 6, we can interpret that the satisfaction level does decrease when the employee is at the beginning of his journey in the company, lowest point is around year 4. This again confirms our interpretation that the unsatisfied employees leave during their first 4 to 5 years in the company. Satisfaction levels seem to increase during the 5th year and comparing it with the last plot (Figure 5), we can see that the number of employees who leave their jobs does in fact decrease. Meaning both interpretations go hand in hand with each other.

So now that we have an idea or an interpretation that employees tend to leave during the first 4-5 years of their journey in the company, we need to do more research in order to understand why that's the case.

The correlogram (Figure 4), gave us an idea of the features that are correlated together. This will be useful to understand and answer the previous point mentioned.

We're going to focus mainly on the first 5 years, by measuring the last evaluation vs the number of projects, we can also use average number of hours, but we know that number of projects and average number of hours are highly correlated with each other, hence no need to do both.

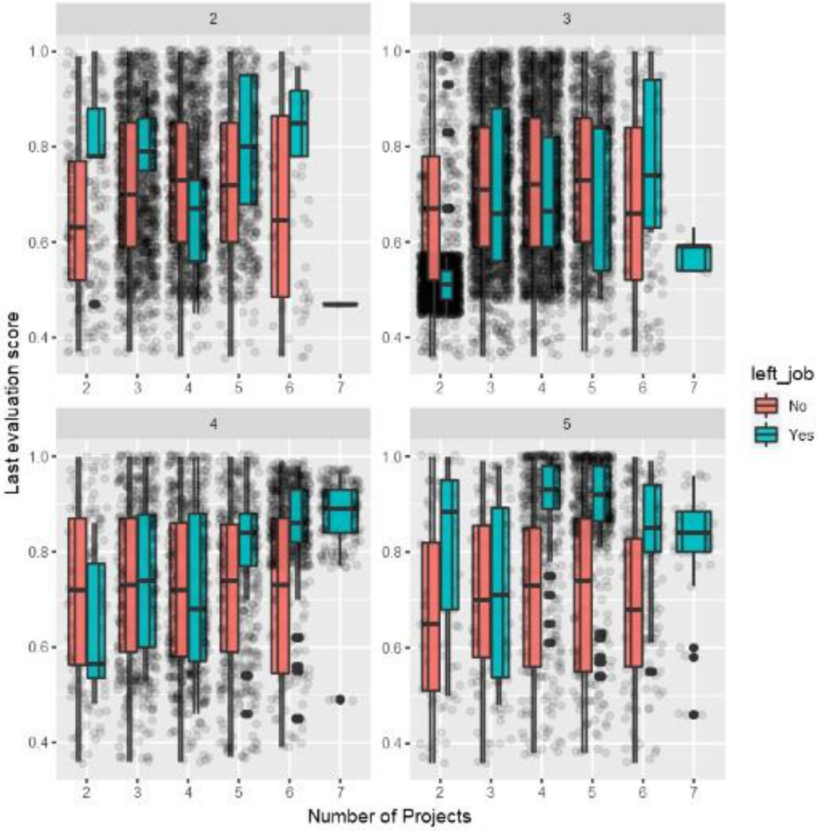


Figure 7

First graph (2nd year), we can notice that employees are handling between 3 to 5 projects (largest group), their evaluation score is around 0.7.

Second graph (3rd year) shows two interesting observations: the first one is we can notice a huge portion of employees handling 2 projects and they're getting low evaluation scores. This means they're not performing as what they company is expecting from them. Second observation is also we have a big portion of employees handling 3 to 5 projects.

We can notice in three of the plotted graphs that employees who are handling 7 projects end up leaving their job, even though we can observe them getting high evaluation scores in the 4 and 5th year. Hence, they are probably leaving because they cannot handle this workload and might be too much for them. We can say also the same thing about employees who are handling 6 projects, although a large portion of these employees decide to continue with the company.

Another observation is that in the 4th and 5th graphs, employees who left the company (3 to 6 projects) are getting high evaluation scores, but they still decided to leave. Salary could be a factor in this case.

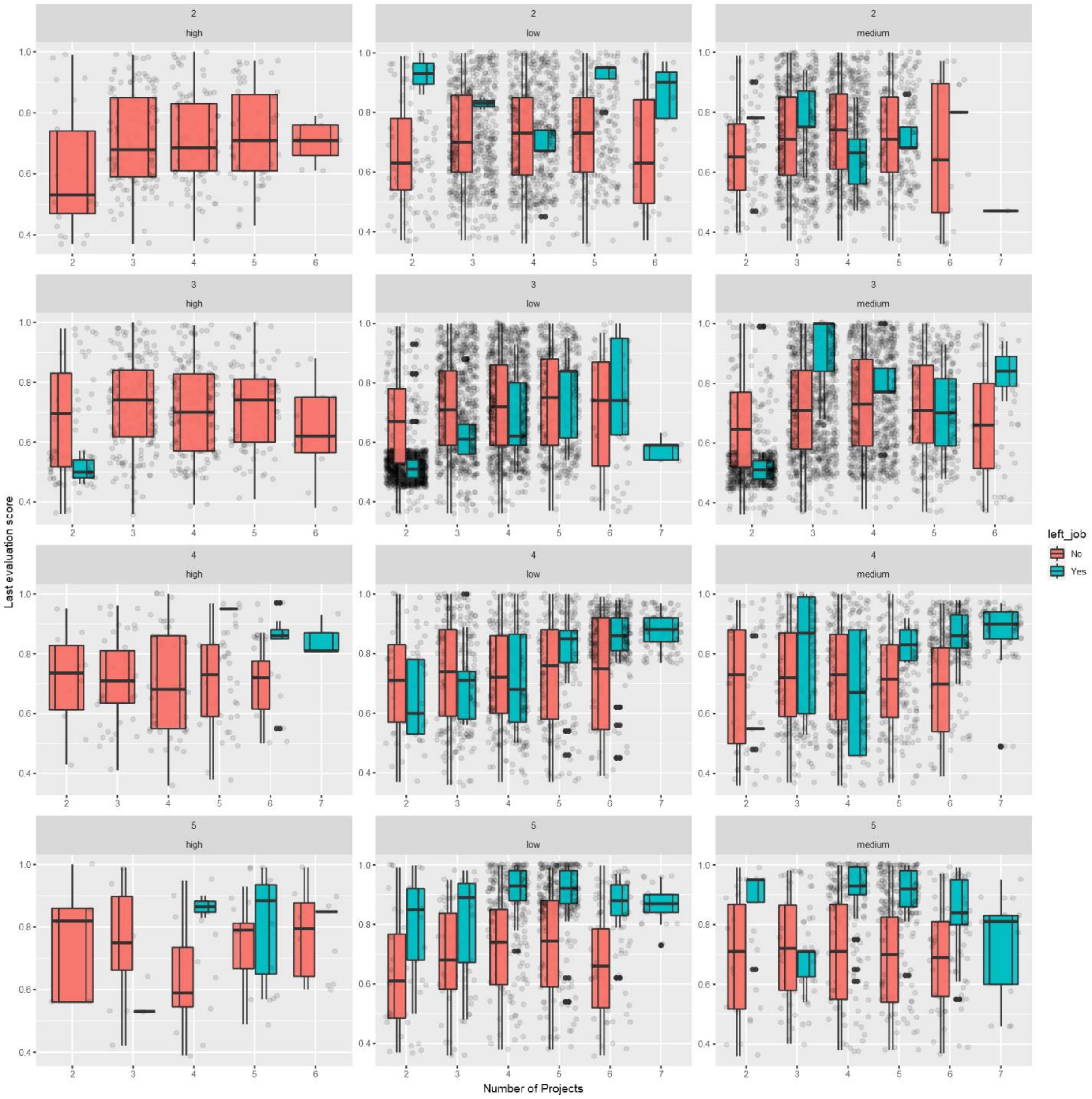


Figure 8

We can infer from Figure 8 an interesting point, employees with high salaries mostly decide to continue with the company except for few cases in year 4 and 5, one reason might still be the heavy workload (5 & 6 projects). The other reason might simply be they decided to pursue a different opportunity in a different company.

Another interesting observation is regarding medium and low salary employees who decided to leave as the number of years increase employees the evaluation rate also increases. Meaning even though employees are getting high evaluation rates (year 4 and 5), they still decided to leave, probably because they feel they have been overworked but not getting enough compensation for their time and effort.

So, what could be interesting to explore is to replace the last evaluation level with the satisfaction level, since we are noticing that the evaluation rate isn't something employees do consider when they decide to leave the company; at least during the 4th and 5th year.

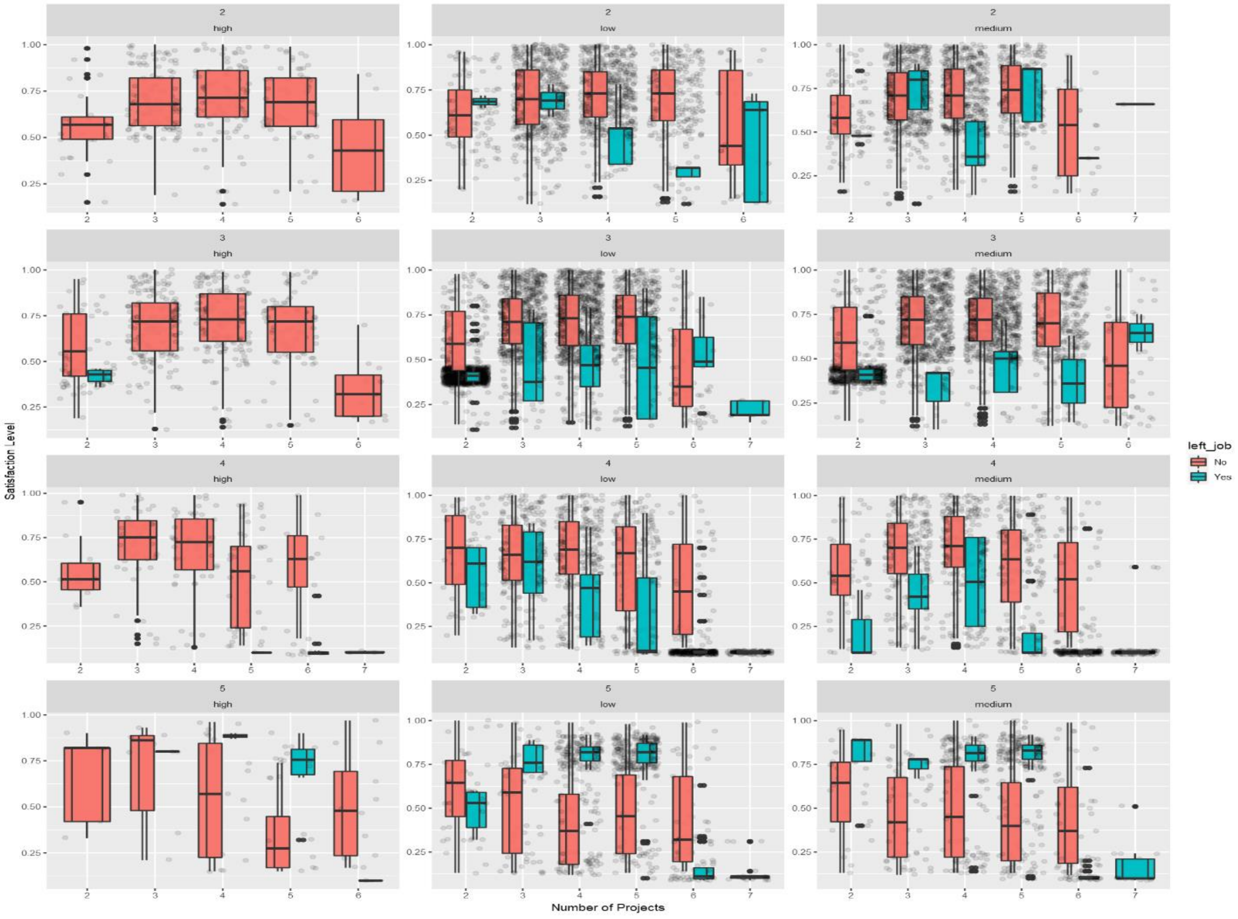


Figure 9

We can observe from Figure 9, that the satisfaction levels of medium and low salary employees tend to be average or even below the average when they have been overworked with projects. This is very noticeable in the 3rd and 4th year.

In the 5th year, we notice an interesting observation that goes hand in hand with the previous inference (high evaluation score), medium and low salary employees left the company even though their satisfaction levels were high, except for employees who have been assigned more than 5 projects. This could mean that they have found another job with a higher salary elsewhere.

Because we know that there exists a correlation between satisfaction level, last evaluation, number of projects and time spent at the company; we can then ask the question whether there exists a significant satisfaction levels difference between the employees who decided to stay and the employees who decided to leave. Same question can be asked regarding the salary.

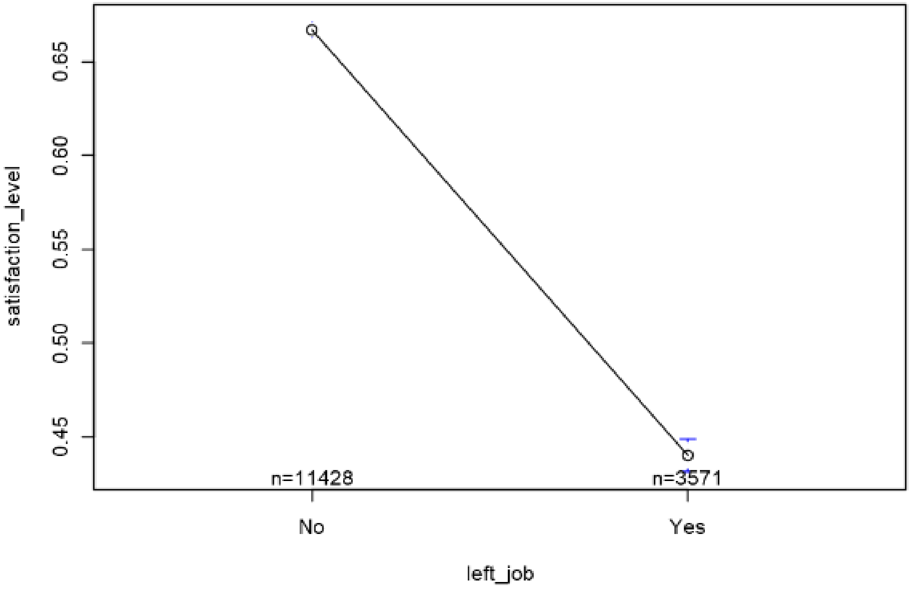


Figure 10

There exists a significant satisfaction level difference between employees who stayed and employees who left, chi-square test was also performed on the salary variable and the result is there exist a significant salary difference between employees who stayed and employees who left.

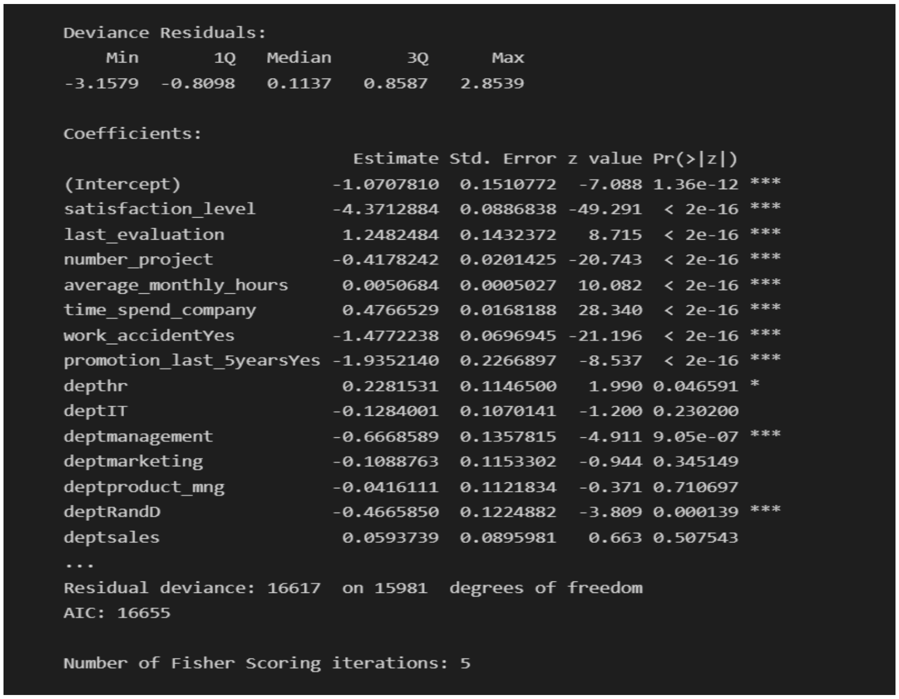
Exploratory data analysis phase has been finished resulting in some meaningful intuitions about the given dataset. Classification models will be performed in order to see if these intuitions lead to important insights and predictions.

4. Classification Models:

Firstly, the dataset has been split into two main parts(70/30), train set and test set. Another point to note is that as mentioned earlier, the dataset suffers from moderate imbalance, which means a given algorithm won't get the proper necessary information about the minority class (in our case it's employees who left) to make an accurate prediction. Therefore a resampling method is needed in order to adjust the number of majority and minority instances we have. One method we're going to use is up sampling, meaning we're going to inject into the dataset synthetically generated data points that correspond to the minority class. However, we need to be aware that this might cause overfitting if applied incorrectly. Since we're going to use 10-fold cross validation in each of our models, we need to apply the resampling method on the training set only, while we're training the model.

A) Logistic Regression:

The first model applied is logistic regression, which is simply a method that predicts a binary outcome, in our case 'yes' or 'no' based on prior observations of the given dataset. It can be considered as an extension of the linear regression model for classification problems. The model was first trained using k-fold cross validation and the accuracy was around 75.9%.



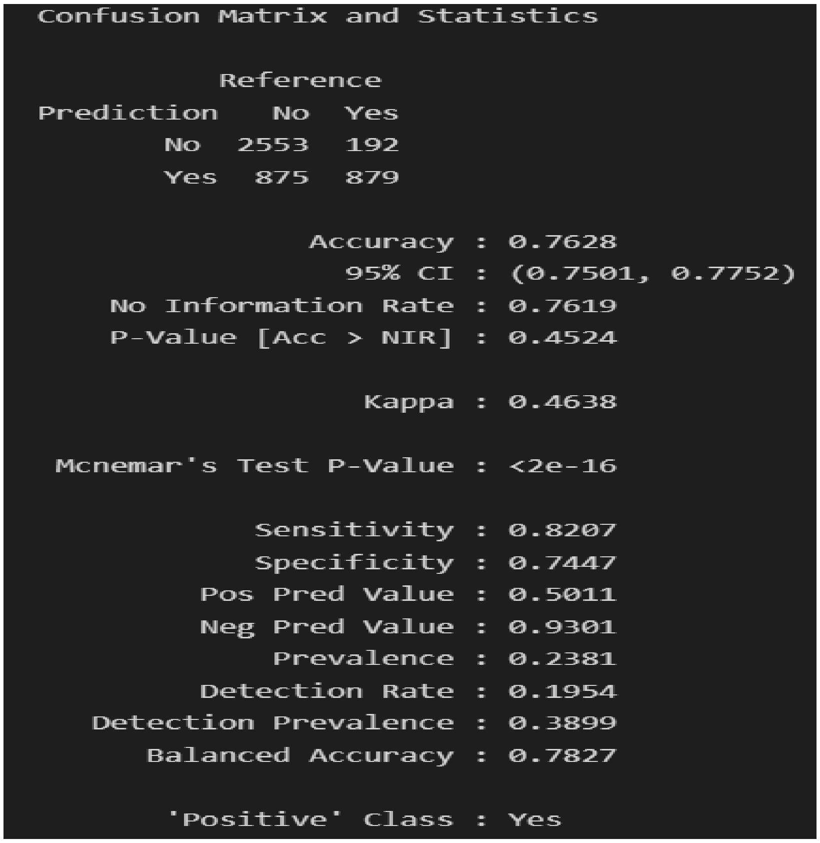
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1579  -0.8098   0.1137   0.8587   2.8539

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.0707810  0.1510772  -7.088 1.36e-12 ***
satisfaction_level       -4.3712884  0.0886838 -49.291  < 2e-16 ***
last_evaluation           1.2482484  0.1432372   8.715  < 2e-16 ***
number_project           -0.4178242  0.0201425 -20.743  < 2e-16 ***
average_monthly_hours     0.0050684  0.0005027  10.082  < 2e-16 ***
time_spend_company        0.4766529  0.0168188  28.340  < 2e-16 ***
work_accidentYes         -1.4772238  0.0696945 -21.196  < 2e-16 ***
promotion_last_5yearsYes -1.9352140  0.2266897  -8.537  < 2e-16 ***
depthr                    0.2281531  0.1146500   1.990 0.046591 *
deptIT                   -0.1284001  0.1070141  -1.200 0.230200
deptmanagement           -0.6668589  0.1357815  -4.911 9.05e-07 ***
deptmarketing            -0.1088763  0.1153302  -0.944 0.345149
deptproduct_mng          -0.0416111  0.1121834  -0.371 0.710697
deptRandD                -0.4665850  0.1224882  -3.809 0.000139 ***
deptsales                 0.0593739  0.0895981   0.663 0.507543
...
Residual deviance: 16617  on 15981  degrees of freedom
AIC: 16655

Number of Fisher Scoring iterations: 5
```

Figure 11

Key points learned from the model’s summary:

- The deviance residuals look good since they are close to being centered on 0 and symmetric as well.
- The coefficient that seem to have the biggest negative impact on the ‘left job’ variable is ‘satisfaction level’, which is understandable since if the employee is dissatisfied then they would leave their job at some point.
- On the other hand, the coefficient that seem to have the highest positive impact on the dependent variable is ‘last evaluation’. Meaning the higher the evaluation the more likely that this person won't end up leaving his/her job. Which is again very reasonable.

A confusion matrix was used on the test set to determine the model’s accuracy on a different given dataset and to observe whether the model suffers from an over/under-fitting problem.



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2553  192
       Yes  875  879

               Accuracy : 0.7628
                 95% CI : (0.7501, 0.7752)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : 0.4524

                  Kappa : 0.4638

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8207
            Specificity : 0.7447
         Pos Pred Value : 0.5011
         Neg Pred Value : 0.9301
             Prevalence : 0.2381
         Detection Rate : 0.1954
   Detection Prevalence : 0.3899
      Balanced Accuracy : 0.7827

       'Positive' Class : Yes
```

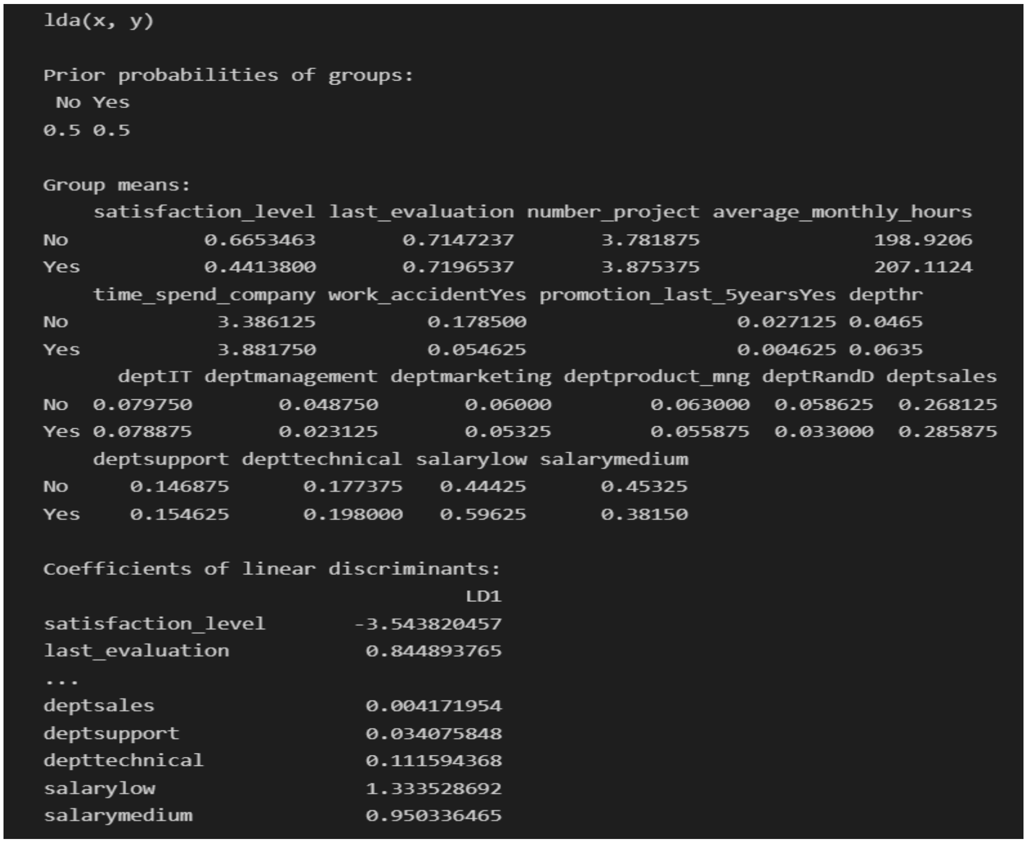
Figure 12

Inferences from the above confusion matrix:

- Model Accuracy is around: 76.3%.
- Sensitivity and specificity scores are good.
- By comparing the trained logistic regression model accuracy with the confusion matrix accuracy using the test set, we can notice that they're almost the same.
- Other models will be applied to determine if the logistic regression model is the most accurate or not.

B) <u>Linear Discriminant Analysis:</u>

The second model applied is LDA, the aim of this classification model is reducing the given dimensions by maximizing the separability among known categories/variables. The best separability can be achieved by maximizing the distance between means and minimizing the variation within each category. The model was first trained using k-fold cross validation and the accuracy was around 75%, very similar to the logistic regression trained model accuracy

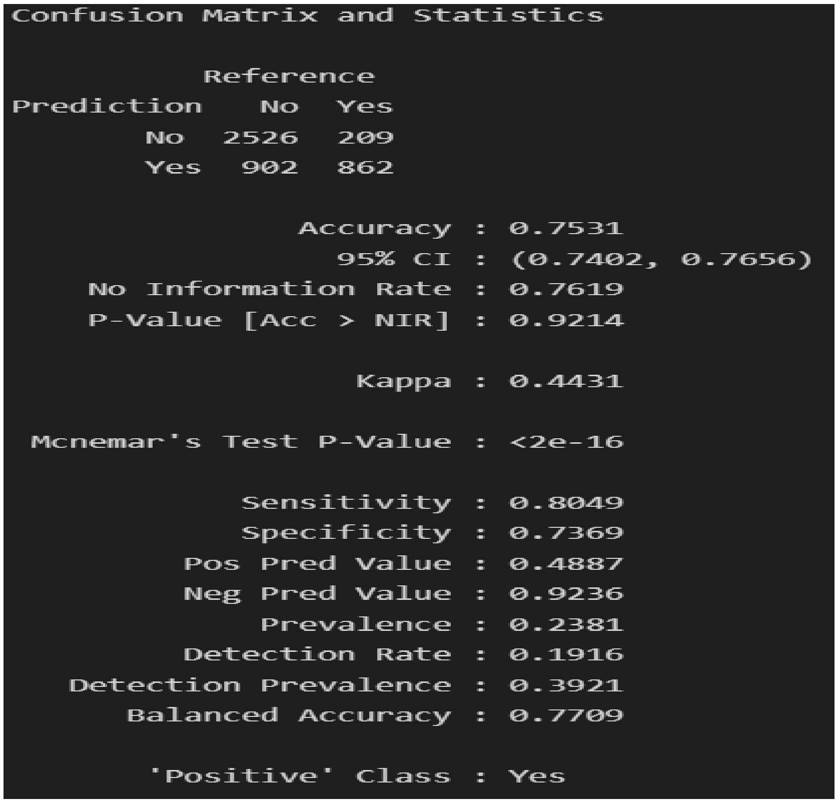


```
lda(x, y)

Prior probabilities of groups:
 No Yes
0.5 0.5

Group means:
    satisfaction_level last_evaluation number_project average_monthly_hours
No           0.6653463       0.7147237       3.781875              198.9206
Yes          0.4413800       0.7196537       3.875375              207.1124
    time_spend_company work_accidentYes promotion_last_5yearsYes depthr
No            3.386125         0.178500                 0.027125 0.0465
Yes           3.881750         0.054625                 0.004625 0.0635
      deptIT deptmanagement deptmarketing deptproduct_mng deptRandD deptsales
No  0.079750       0.048750       0.06000        0.063000  0.058625  0.268125
Yes 0.078875       0.023125       0.05325        0.055875  0.033000  0.285875
    deptsupport depttechnical salarylow salarymedium
No     0.146875      0.177375   0.44425      0.45325
Yes    0.154625      0.198000   0.59625      0.38150

Coefficients of linear discriminants:
                                 LD1
satisfaction_level       -3.543820457
last_evaluation           0.844893765
...
deptsales                 0.004171954
deptsupport               0.034075848
depttechnical             0.111594368
salarylow                 1.333528692
salarymedium              0.950336465
```

Figure 13

Figure 13 displays the summary of the LDA model, we can observe the group means and also an interesting point to visualize the number of employees who stayed vs the number of employees who left is the same, since we've applied up sampling to our training set.

A confusion matrix was used on the test set to determine the model’s accuracy on a different given dataset and to observe whether the model suffers from an over/under-fitting problem.



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2526  209
       Yes  902  862

               Accuracy : 0.7531
                 95% CI : (0.7402, 0.7656)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : 0.9214

                  Kappa : 0.4431

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8049
            Specificity : 0.7369
         Pos Pred Value : 0.4887
         Neg Pred Value : 0.9236
             Prevalence : 0.2381
         Detection Rate : 0.1916
   Detection Prevalence : 0.3921
      Balanced Accuracy : 0.7709

       'Positive' Class : Yes
```

Figure 14

Inferences from the above confusion matrix:

- Model Accuracy is around: 75.3%.
- Sensitivity and specificity scores are good.
- By comparing the trained LDA model accuracy with the confusion matrix accuracy using the test set, we can notice that they're almost the same.
- It seems that logistic regression is slightly performing better than the LDA model, one reason could be that the logistic regression is more flexible and more robust than LDA.

C) <u>Classification Tree:</u>

The third model applied is classification tree, the aim of this model is stratifying/segmenting the predictor space into a number of simple regions. This model was applied using the ‘rpart’ library which uses k-fold cross validation to validate the optimal cost complexity parameter cp. Misclassification for each class was as well adjusted using ‘prior probabilities’ in order to handle the imbalance our dataset is subject to.

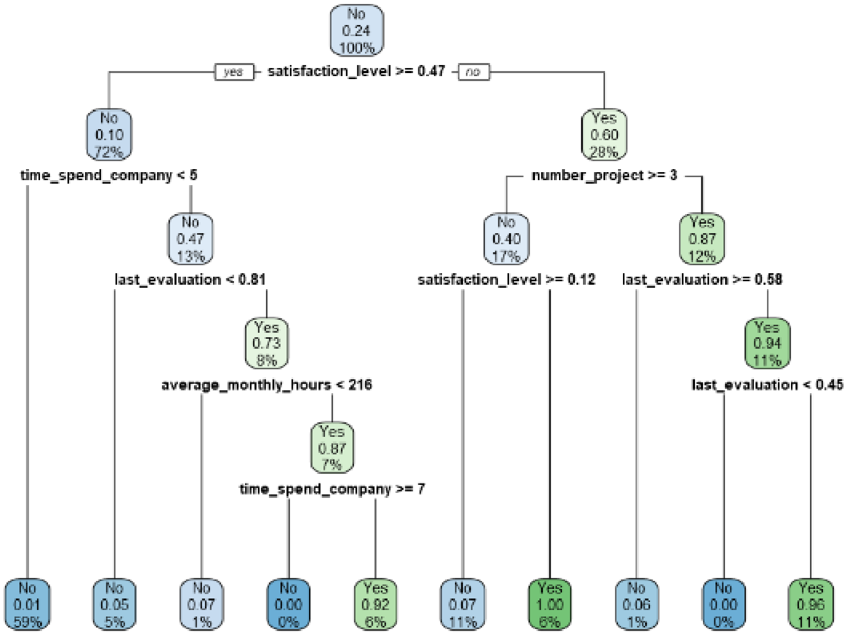
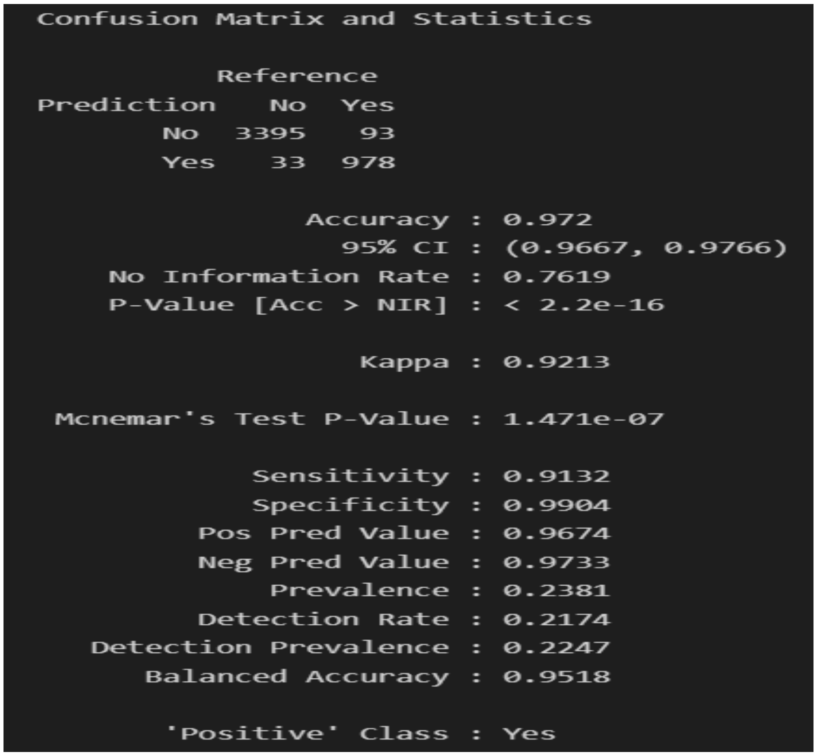


Figure 15

The classification tree is composed of 5 levels, where the root node is 'satisfaction level' and other two important nodes are 'time spend company' and 'number project'. A confusion matrix was used on the test set to determine the model's accuracy on a different given dataset.

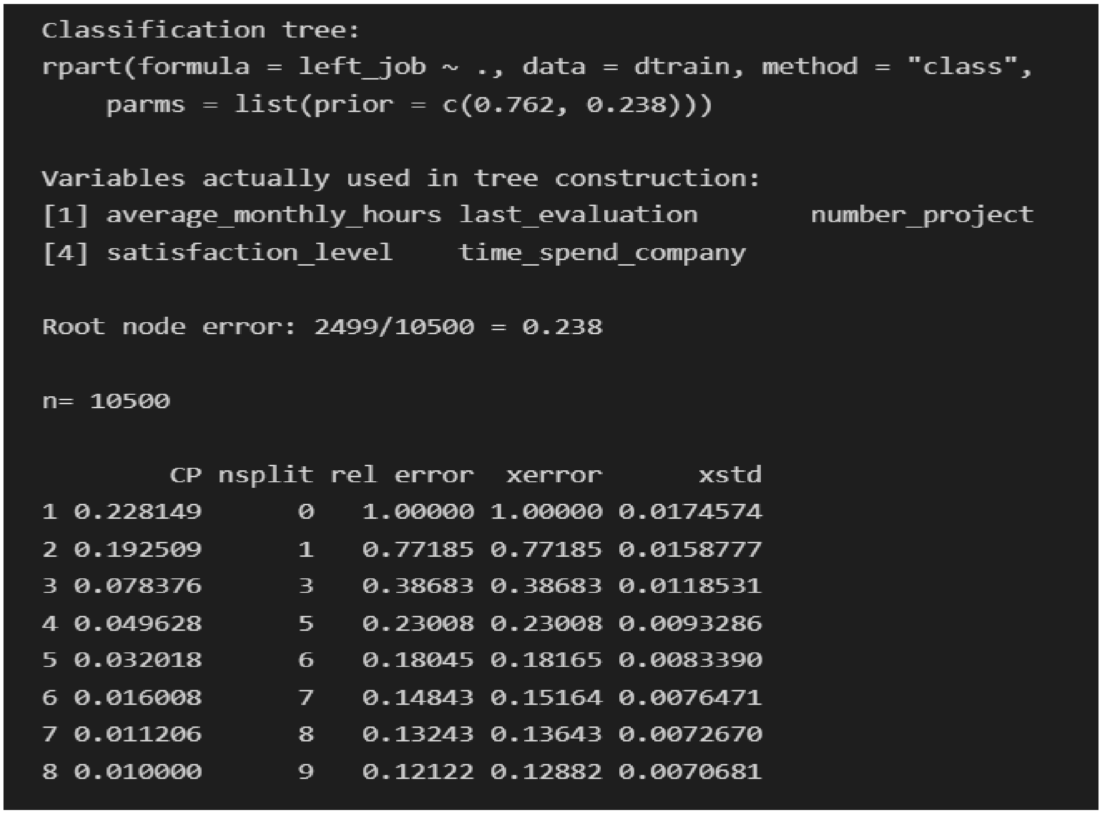


```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3395   93
       Yes   33  978

               Accuracy : 0.972
                 95% CI : (0.9667, 0.9766)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9213

 Mcnemar's Test P-Value : 1.471e-07

            Sensitivity : 0.9132
            Specificity : 0.9904
         Pos Pred Value : 0.9674
         Neg Pred Value : 0.9733
             Prevalence : 0.2381
         Detection Rate : 0.2174
   Detection Prevalence : 0.2247
      Balanced Accuracy : 0.9518

       'Positive' Class : Yes
```

Figure 16

Inferences from the above confusion matrix:

- Model Accuracy is around: 97.2%.
- Sensitivity and specificity scores are very good.
- It seems that classification tree model is performing much better than the LDA and the logistic regress models.
- Pruning the tree might improve the model’s accuracy and reduce the levels of the tree.



```
Classification tree:
rpart(formula = left_job ~ ., data = dtrain, method = "class",
    parms = list(prior = c(0.762, 0.238)))

Variables actually used in tree construction:
[1] average_monthly_hours last_evaluation       number_project
[4] satisfaction_level    time_spend_company

Root node error: 2499/10500 = 0.238

n= 10500

        CP nsplit rel error  xerror      xstd
1 0.228149      0   1.00000 1.00000 0.0174574
2 0.192509      1   0.77185 0.77185 0.0158777
3 0.078376      3   0.38683 0.38683 0.0118531
4 0.049628      5   0.23008 0.23008 0.0093286
5 0.032018      6   0.18045 0.18165 0.0083390
6 0.016008      7   0.14843 0.15164 0.0076471
7 0.011206      8   0.13243 0.13643 0.0072670
8 0.010000      9   0.12122 0.12882 0.0070681
```

Figure 17

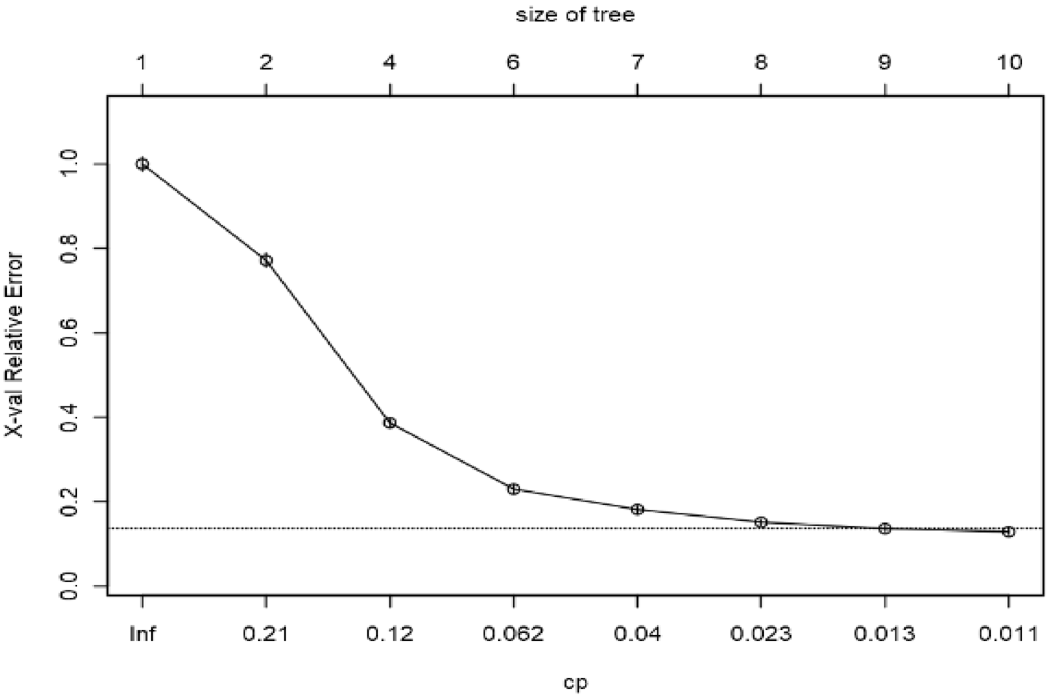


Figure 18

A list of complexity parameters have been observed (Figure 17 & 18), where the minimum cp is selected in order to have the optimal pruned tree. However, using cp = 0.01 didn't prune the original tree. Hence, according to the 1-SE rule (Selecting the smallest tree that has a misclassification rate below the horizontal reference line), 0.013 was selected as cp.

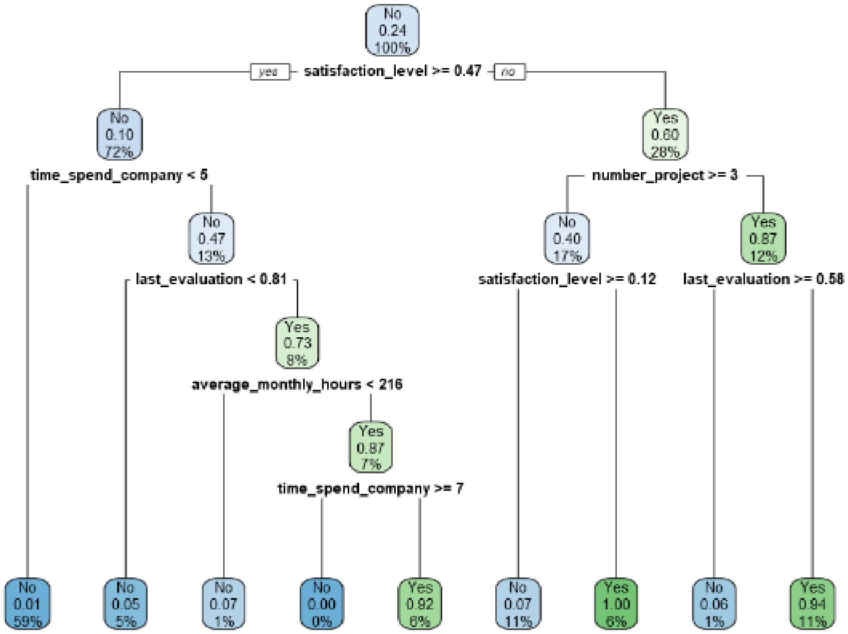
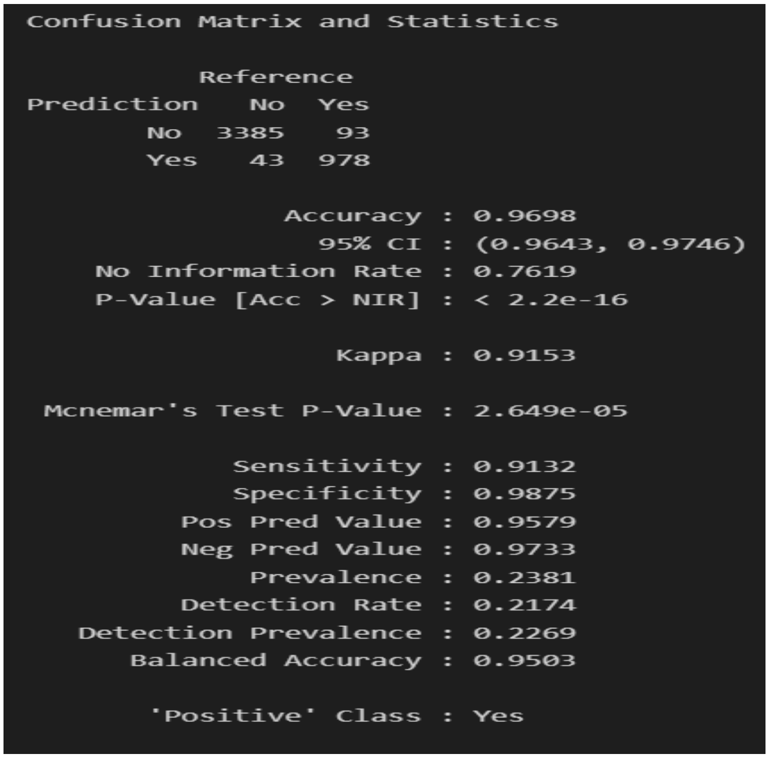


Figure 19

The tree has been pruned slightly (Figure 19), however not significantly. A confusion matrix is applied to determine whether the model's accuracy will improve.



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3385   93
       Yes   43  978

               Accuracy : 0.9698
                 95% CI : (0.9643, 0.9746)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9153

 Mcnemar's Test P-Value : 2.649e-05

            Sensitivity : 0.9132
            Specificity : 0.9875
         Pos Pred Value : 0.9579
         Neg Pred Value : 0.9733
             Prevalence : 0.2381
         Detection Rate : 0.2174
   Detection Prevalence : 0.2269
      Balanced Accuracy : 0.9503

       'Positive' Class : Yes
```

Figure 20

Comparing the accuracy of both confusion matrices (Figure 16 & Figure 20), the pruned tree seems slightly less accurate. Two methods have been applied to prune the original tree, pre-pruning and post-pruning, both result in the same output. Therefore, because pruning the tree doesn't improve the original tree, we can determine that the best accuracy for this model is using the original tree, where it's 97.2% accurate.

D) <u>Random Forest:</u>

The fourth and final model applied is Random Forest. The aim of this model is to combine the simplicity of the classification tree with flexibility in order to result in a vast improvement in accuracy. Where an ensemble of trees will be built and instead of searching for the most important feature while splitting a node, the algorithm looks for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

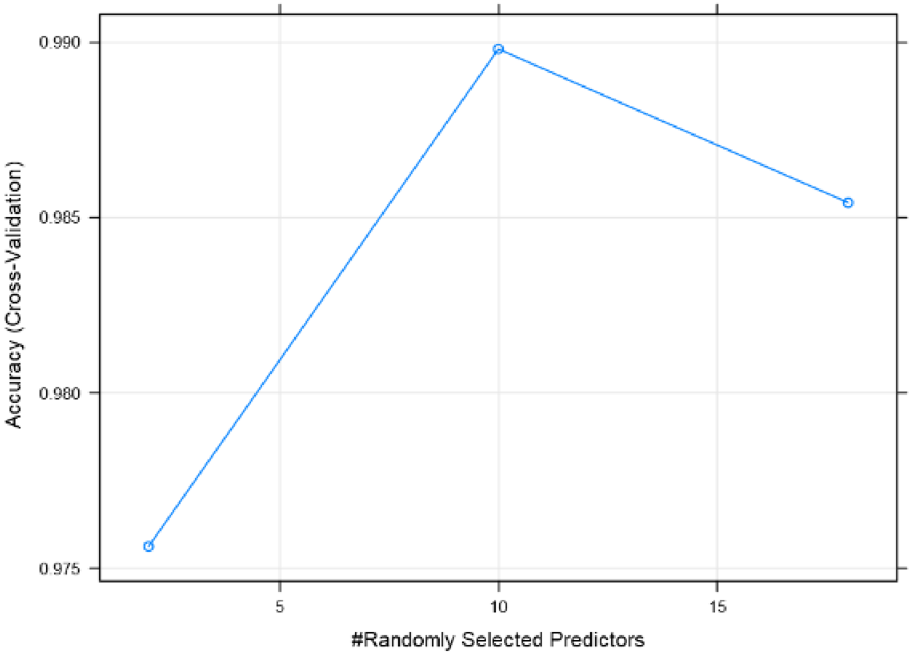


Figure 21

The model was first trained using k-fold cross validation and the accuracy was around 99%, best accuracy achieved. The number of random subset of features is 10. (Figure 21)

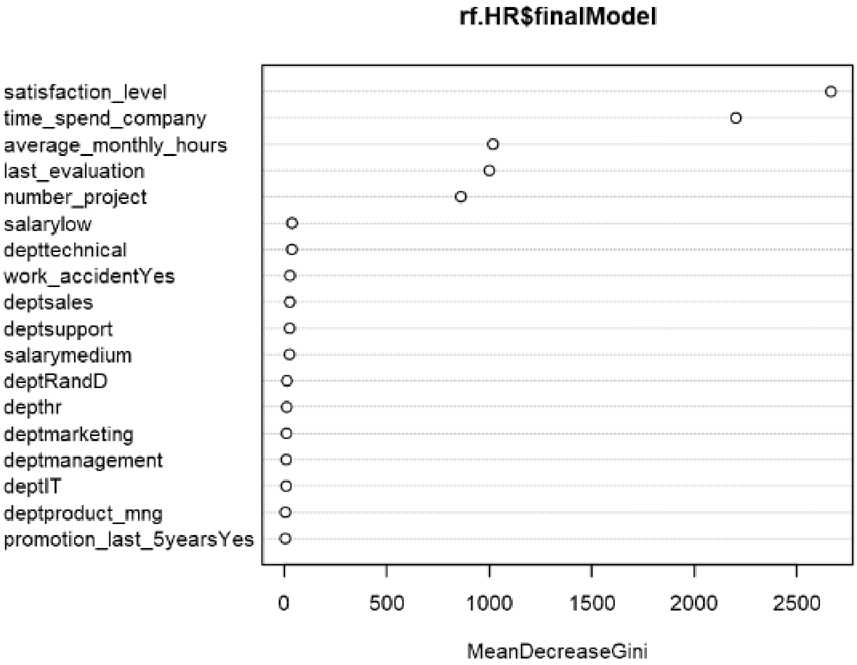
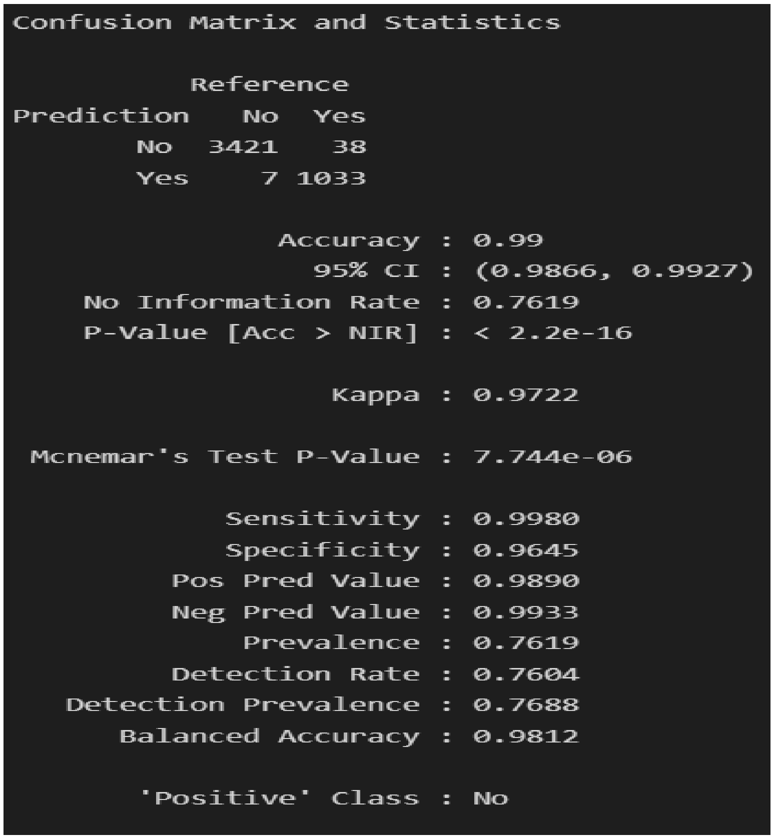


Figure 22

From the Mean Decrease Gini plot (Figure 22), it seems that the main important features are time spent in the company, average monthly hours and last evaluation, number of projects and the most important of them all is satisfaction level.



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3421   38
       Yes    7 1033

               Accuracy : 0.99
                 95% CI : (0.9866, 0.9927)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9722

 Mcnemar's Test P-Value : 7.744e-06

            Sensitivity : 0.9980
            Specificity : 0.9645
         Pos Pred Value : 0.9890
         Neg Pred Value : 0.9933
             Prevalence : 0.7619
         Detection Rate : 0.7604
   Detection Prevalence : 0.7688
      Balanced Accuracy : 0.9812

       'Positive' Class : No
```

Figure 23

Inferences from the above confusion matrix:

- Model Accuracy is around: 99%.
- Sensitivity and specificity scores are very good.

5. Conclusion & Model Evaluation:

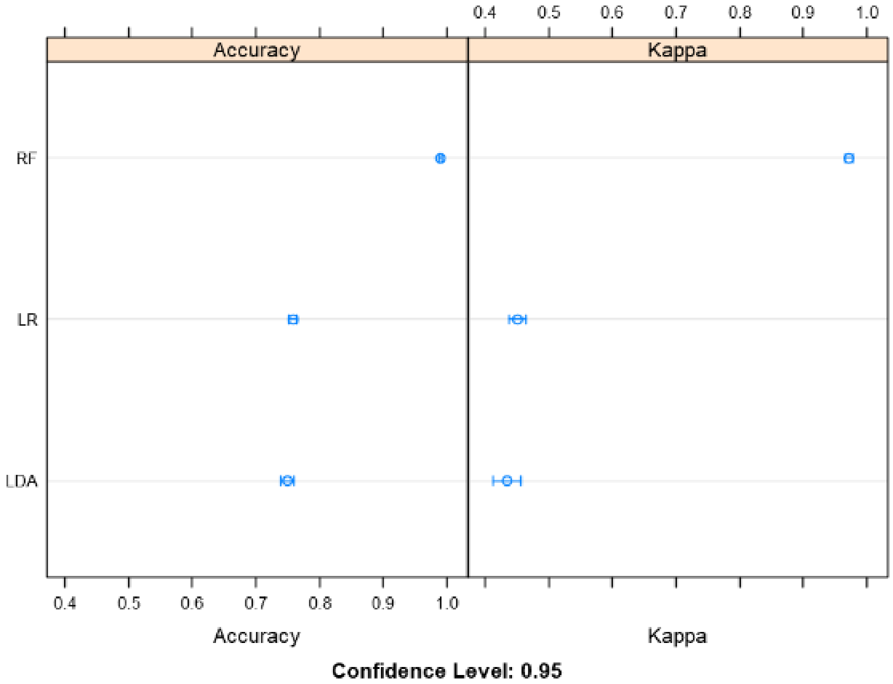


Figure 24

Figure 24 displays the accuracy of each model, only the models where the K-fold cross validation was used (using the caret library) are displayed. The reason is to have a standard comparison, where the same library has been used across the models. Random Forest seems to be the most train/test accurate, 99% accuracy. The variables with the highest importance scores (the ones that give the best prediction and contribute most to the model) are satisfaction level and time spent in the company. This leads us to our initial intuition where we observed that there exists a significant satisfaction level difference between employees who stayed and employees who left (Figure 10).

In conclusion, this data set gave an insight on what could be the main determinants (satisfaction level & time spent in the company) on why employees resign and also gave a time span of when it's most likely this will occur.