ASSIGNMENT – 6

1. INTRODUCTION

Human Resources Analytics is of vital importance to any company. As the economic scenario keeps on changing due to globalization, HR department also needs to adapt to this new change. Hence, retaining top talents is the primary concern for the HR today. Better employee engagement and retaining strategies are the need of the hour. To do this, understanding employee behavior is critical as it gives an insight into their performance. In HR analytics, now-a-days the main focus is on employee retention vs. employee attrition which helps in forecasting workforce requirements. This will enable in the recognizing factors for employee satisfaction and productivity.

According to Momin & Taruna[1] ("HR analytics transforming human resource management, p.2, HR Analytics), "HR analytics or workforce analytics aids the organizations to make workforce decisions by reducing the costs, identifying the revenue streams, mitigate risks, and execute effective business strategies. HR analytics empowers the HR managers with accurate predictive analytics which determines the future, mainly for the organizations seeking more proactive role in driving business strategy."

Hypothesis for analysis in this report:

Null Hypothesis: No one left the company.
Alternative Hypothesis: At least one left the company.

2. DATA DESCRIPTION

After searching online for relevant datasets, I found the "Human Resources Analytics" hosted on Kaggle[2] to be an ideal choice. The dataset had the scope of analyzing why employees leave prematurely, and several models could be used which can enhance my understanding of the algorithms.

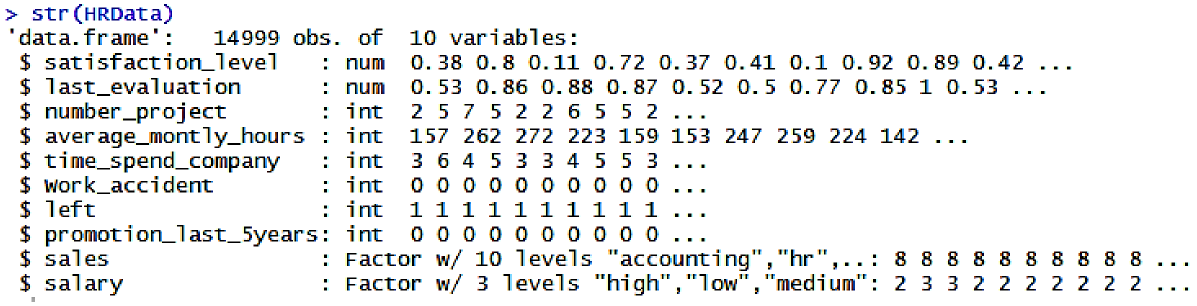
- The dataset has 10 variables and 14,999 observations.
- It consists of 8 numeric variables and 2 factor variables.

Fields in the dataset include:

- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years

- Department
- Salary
- Whether the employee has left

The variables ‘Sales’& ‘Salary’ are factor variables, while the rest were numeric variables.

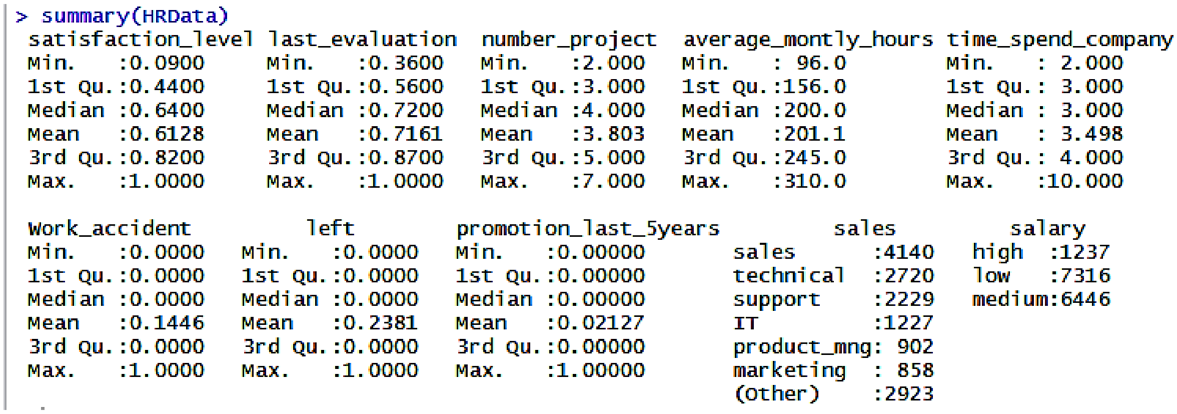


```
> str(HRData)
'data.frame':	14999 obs. of  10 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
 $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
 $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
 $ average_montly_hours : int  157 262 272 223 159 153 247 259 224 142 ...
 $ time_spend_company   : int  3 6 4 5 3 3 4 5 5 3 ...
 $ Work_accident        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ left                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
 $ sales                : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ salary               : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

3. ANALYSIS

The dataset does not have any missing values, and thus not require any imputation.

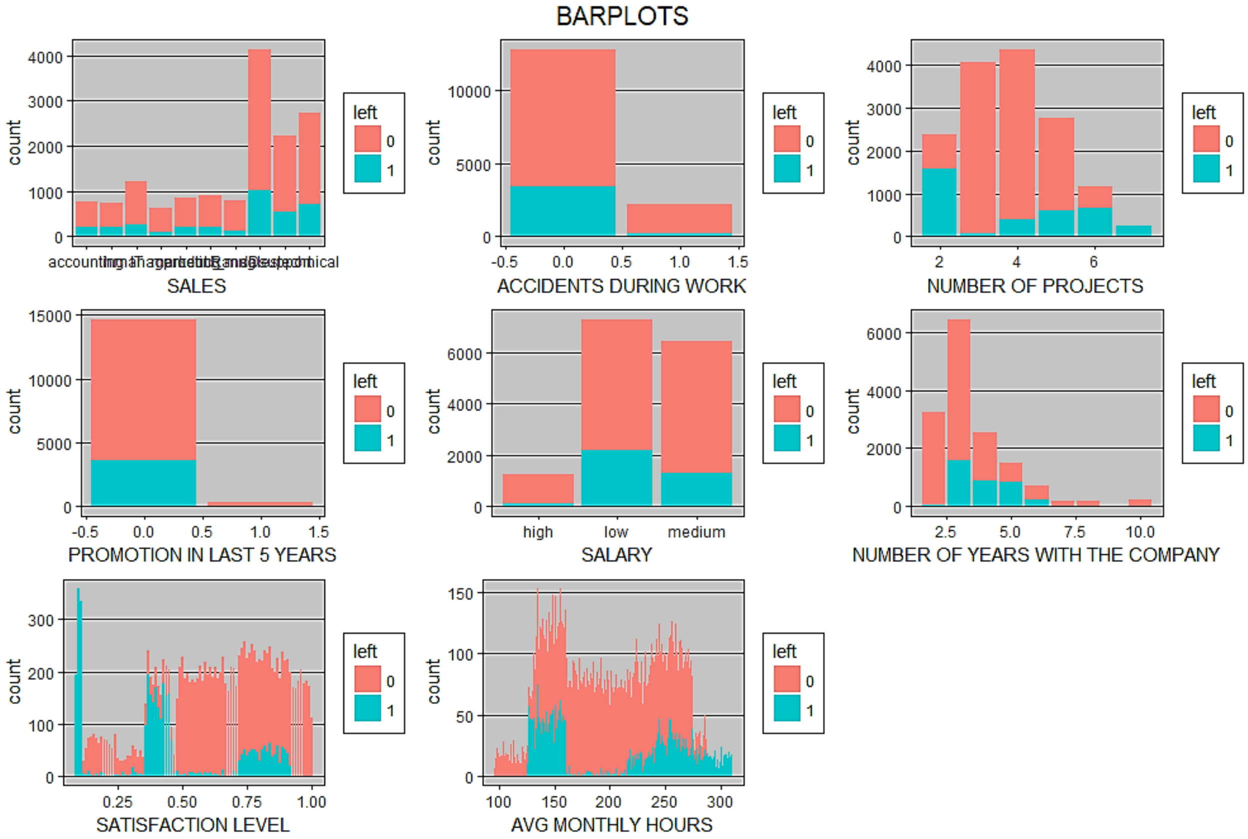
SUMMARY OF THE DATASET:



```
> summary(HRData)
 satisfaction_level last_evaluation  number_project  average_montly_hours time_spend_company
 Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0        Min.   : 2.000    
 1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0        1st Qu.: 3.000    
 Median :0.6400     Median :0.7200   Median :4.000   Median :200.0        Median : 3.000    
 Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1        Mean   : 3.498    
 3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0        3rd Qu.: 4.000    
 Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0        Max.   :10.000    
                                                                                            
 Work_accident         left        promotion_last_5years         sales         salary    
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000       sales      :4140   high  :1237  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000       technical  :2720   low   :7316  
 Median :0.0000   Median :0.0000   Median :0.00000       support    :2229   medium:6446  
 Mean   :0.1446   Mean   :0.2381   Mean   :0.02127       IT         :1227                
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000       product_mng: 902                
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000       marketing  : 858                
                                                         (Other)    :2923                
```

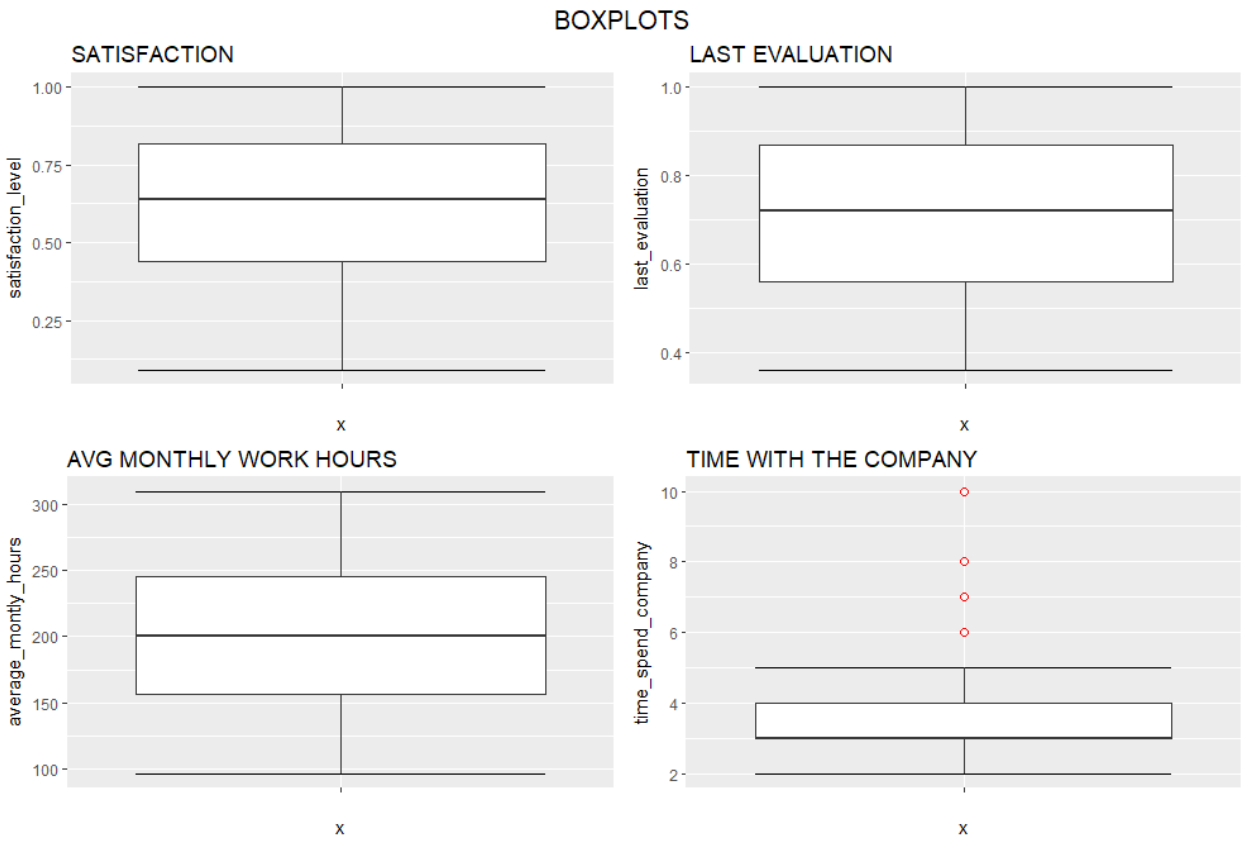
BAR PLOTS:

- The bar plots below show how each variable contributes to whether someone left the company or not.
- It can be seen that employees with high salary left the company, and so did people with higher satisfaction level.
- Employees with more average working hours stayed with the company, which is unusual.
- Employees on multiple projects seemed to stay back with the company.

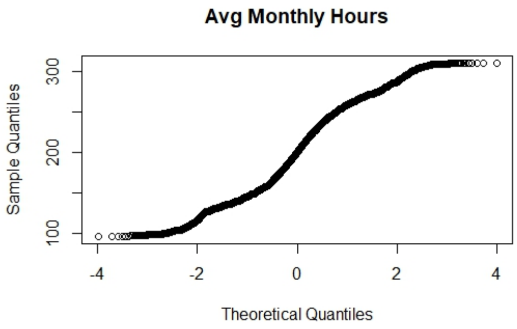


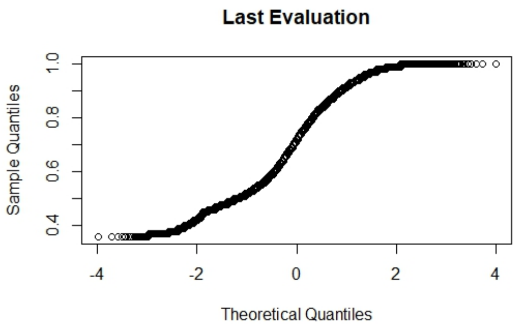
BOX PLOTS:

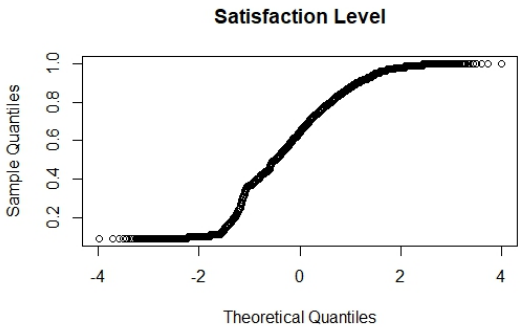
- The boxplots suggest that the dataset does not have outliers for most of the variables.
- The variable ‘Time with the company’ has 4 outliers which might not be wrong as some tend to stay with the company for a longer time.


BOXPLOTS
SATISFACTION
satisfaction_level
1.00
0.75
0.50
0.25
x
LAST EVALUATION
last_evaluation
1.0
0.8
0.6
0.4
x
AVG MONTHLY WORK HOURS
average_montly_hours
300
250
200
150
100
x
TIME WITH THE COMPANY
time_spend_company
10
8
6
4
2
x

QQ-PLOTS:

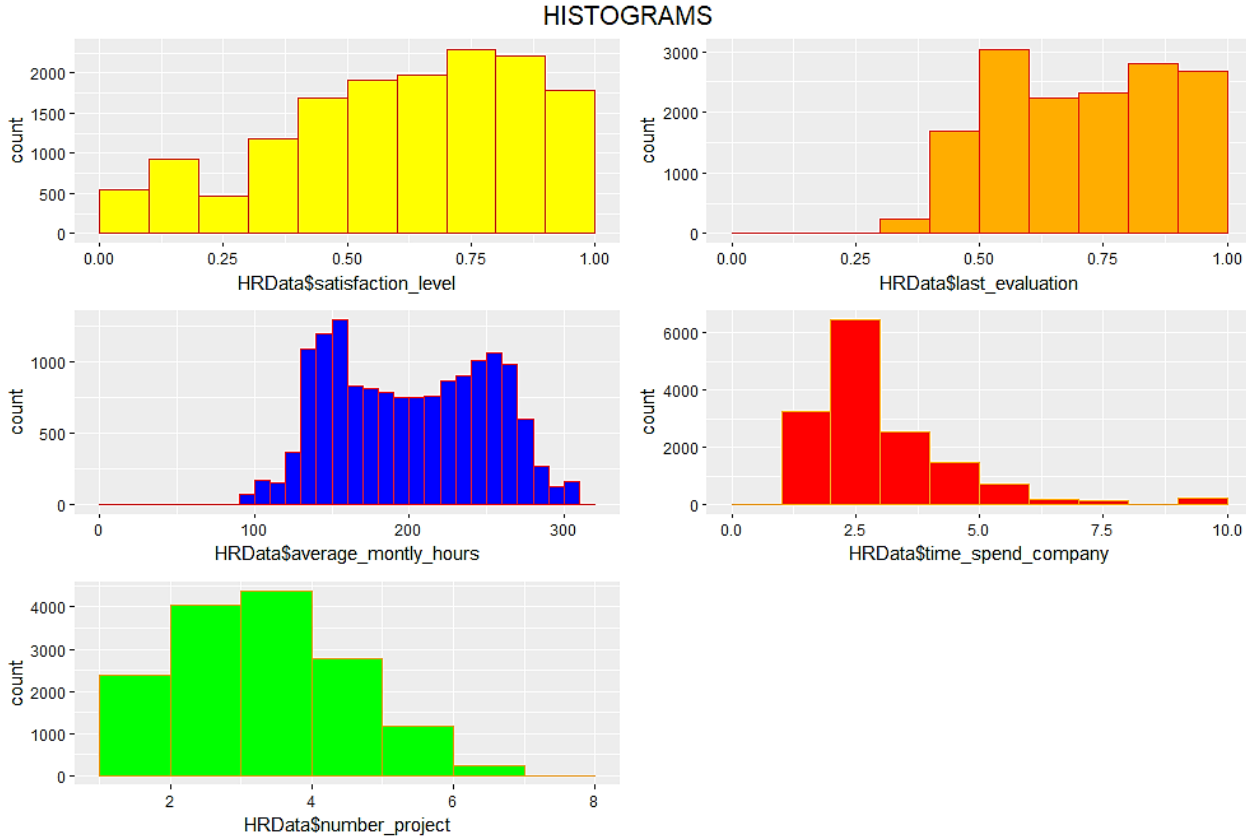

Avg Monthly Hours
Sample Quantiles
300
200
100
-4
-2
0
2
4
Theoretical Quantiles


Last Evaluation
Sample Quantiles
1.0
0.8
0.6
0.4
-4
-2
0
2
4
Theoretical Quantiles


Satisfaction Level
Sample Quantiles
1.0
0.8
0.6
0.4
0.2
-4
-2
0
2
4
Theoretical Quantiles

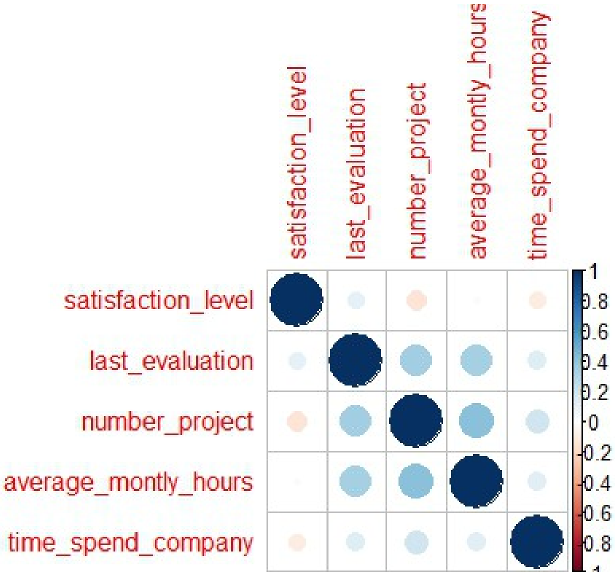
- It can be seen from the plots that the variables are not normally distributed.
- Satisfaction level is skewed to the left.
- Last evaluation are skewed to the right.
- Average monthly hours of work has no proper distribution.

HISTOGRAMS:



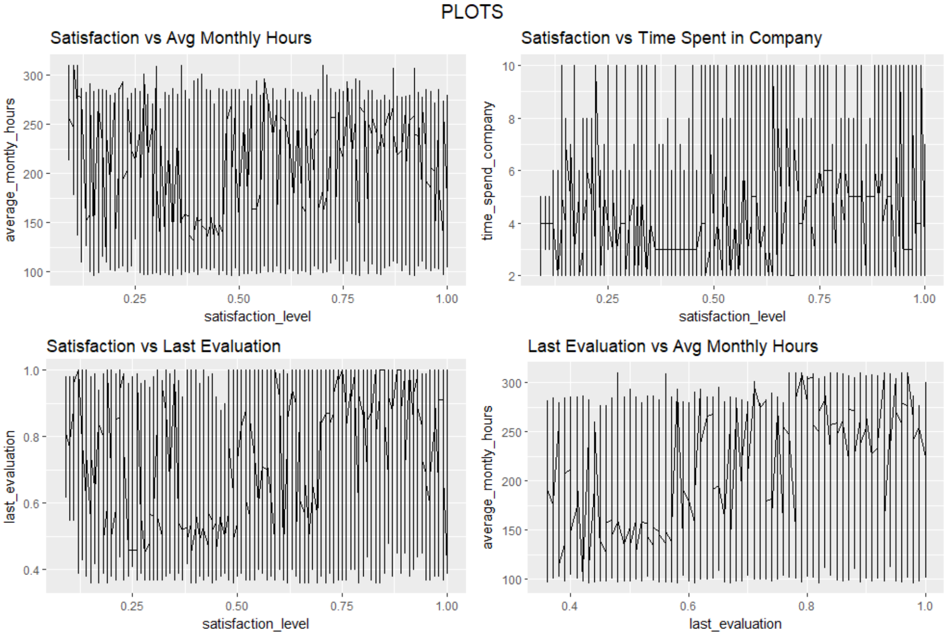
- The histograms reveal that no variable is distributed normally.
- The variable 'Number of projects' is right skewed.
- Satisfaction level is skewed to the left.
- Time spent with the company is also right skewed.

CORRELATION PLOT:



- It can be seen from the plot that there is not much correlation exists between the variables, which eliminates the possibility of multicollinearity.

PLOTS:



- The plots above also suggest that the variables do not have too much effect on each other, and that they are randomly distributed.

4. MODEL DEVELOPMENT AND APPLICATION OF MODEL(S)

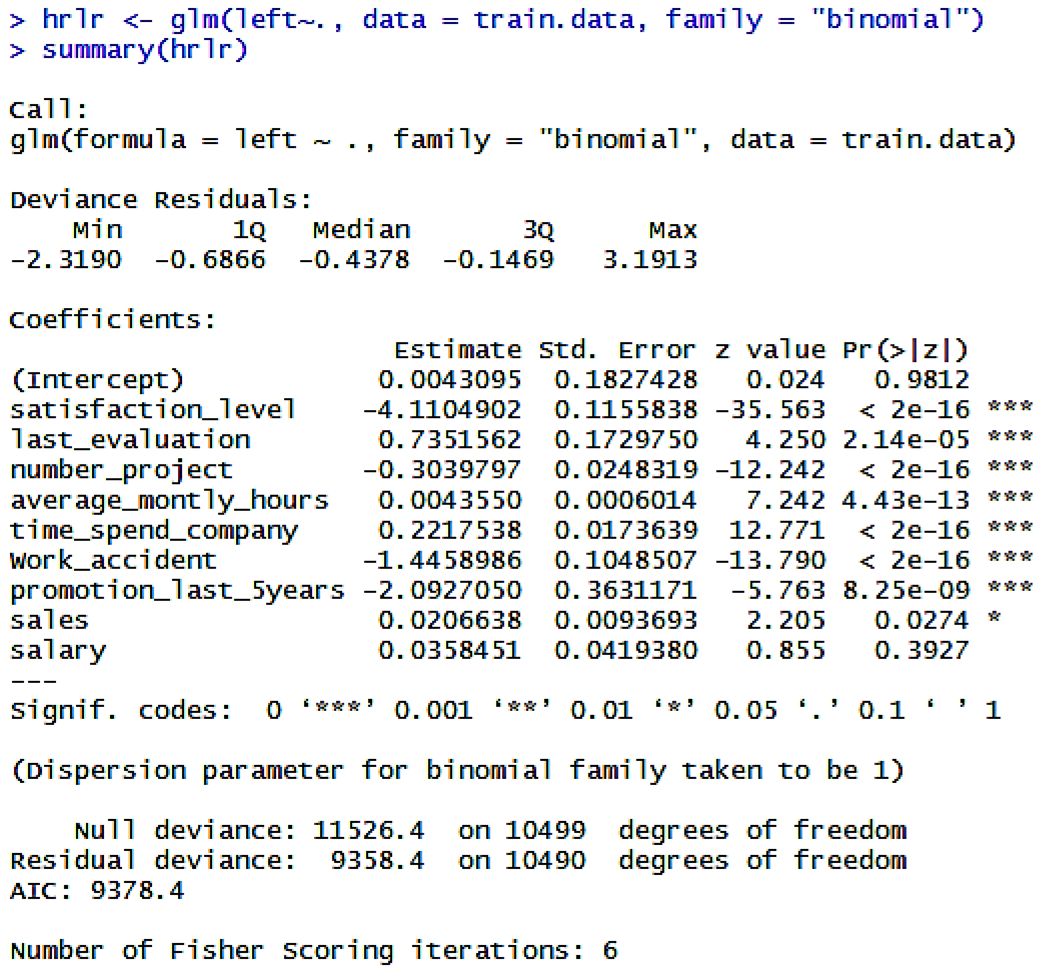
With ‘left’ as the dependent variable, it is a classification model as left = 0 or left =1.

With ‘satisfaction level’ as the dependent variable, it is a regression model as it is a continuous variable.

DEPENDENT VARIABLE – LEFT:

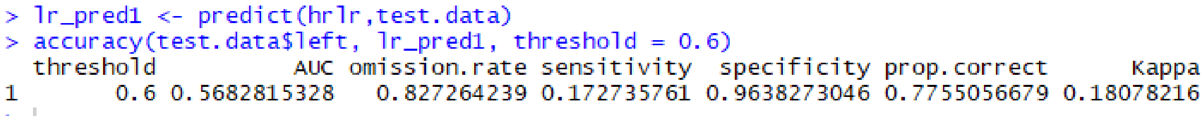
- The models logistic regression, step model, random forest, naïve bayes, support vector machine and k-nearest neighbors are used.

LOGISTIC REGRESSION:



```
> hrlr <- glm(left~., data = train.data, family = "binomial")
> summary(hrlr)

Call:
glm(formula = left ~ ., family = "binomial", data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3190  -0.6866  -0.4378  -0.1469   3.1913  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            0.0043095  0.1827428   0.024   0.9812    
satisfaction_level    -4.1104902  0.1155838 -35.563  < 2e-16 ***
last_evaluation        0.7351562  0.1729750   4.250 2.14e-05 ***
number_project        -0.3039797  0.0248319 -12.242  < 2e-16 ***
average_montly_hours   0.0043550  0.0006014   7.242 4.43e-13 ***
time_spend_company     0.2217538  0.0173639  12.771  < 2e-16 ***
Work_accident         -1.4458986  0.1048507 -13.790  < 2e-16 ***
promotion_last_5years -2.0927050  0.3631171  -5.763 8.25e-09 ***
sales                  0.0206638  0.0093693   2.205   0.0274 *  
salary                 0.0358451  0.0419380   0.855   0.3927    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11526.4  on 10499  degrees of freedom
Residual deviance:  9358.4  on 10490  degrees of freedom
AIC: 9378.4

Number of Fisher Scoring iterations: 6
```

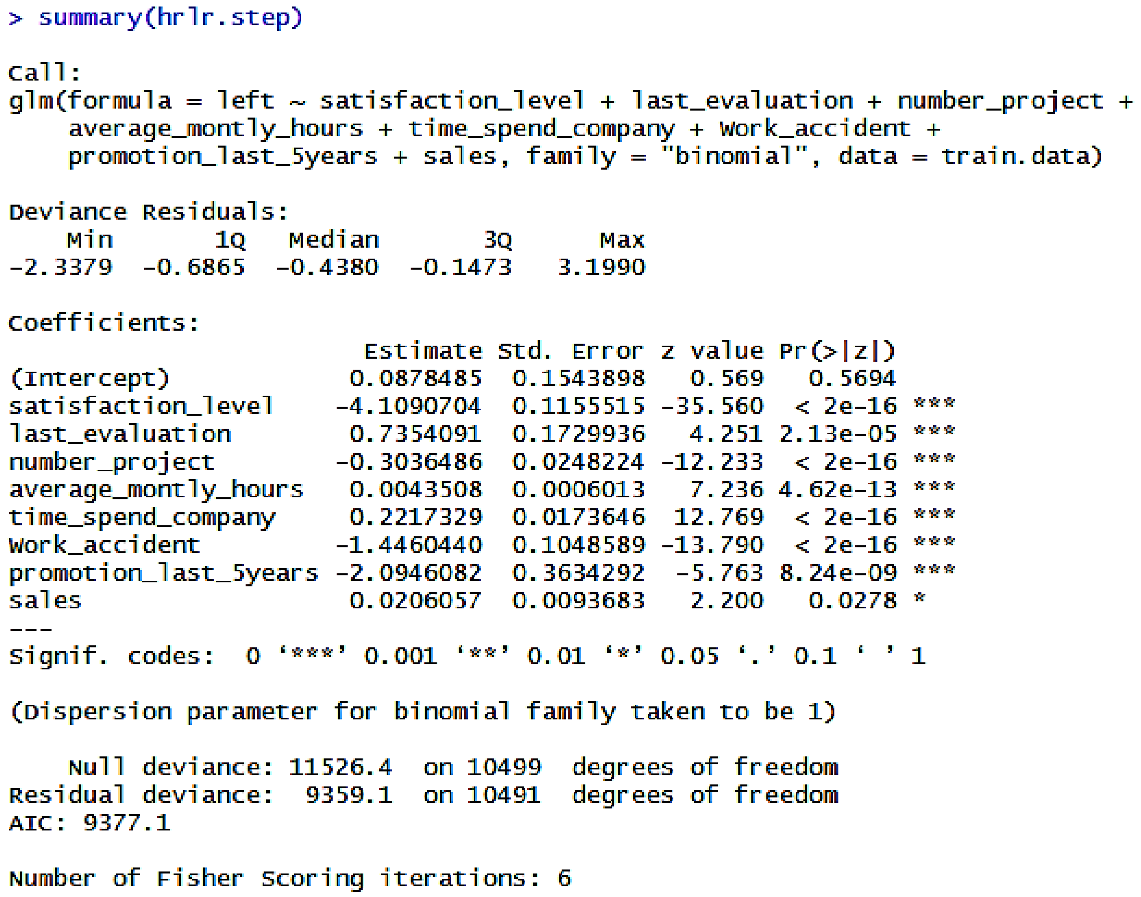
- The null and residual deviance are high indicating not a great fit.
- The variable ‘salary’ is not significant to the model.



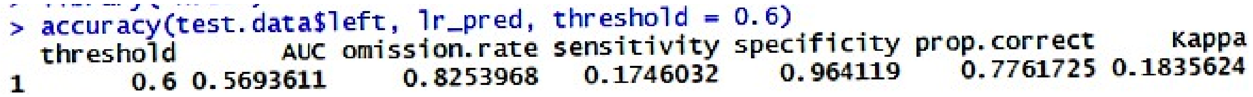
```
> lr_pred1 <- predict(hrlr,test.data)
> accuracy(test.data$left, lr_pred1, threshold = 0.6)
  threshold          AUC omission.rate sensitivity  specificity prop.correct      Kappa
1       0.6 0.5682815328   0.827264239 0.172735761 0.9638273046 0.7755056679 0.18078216
```

- The accuracy after running the model on the test dataset is 0.775.

STEP MODEL:



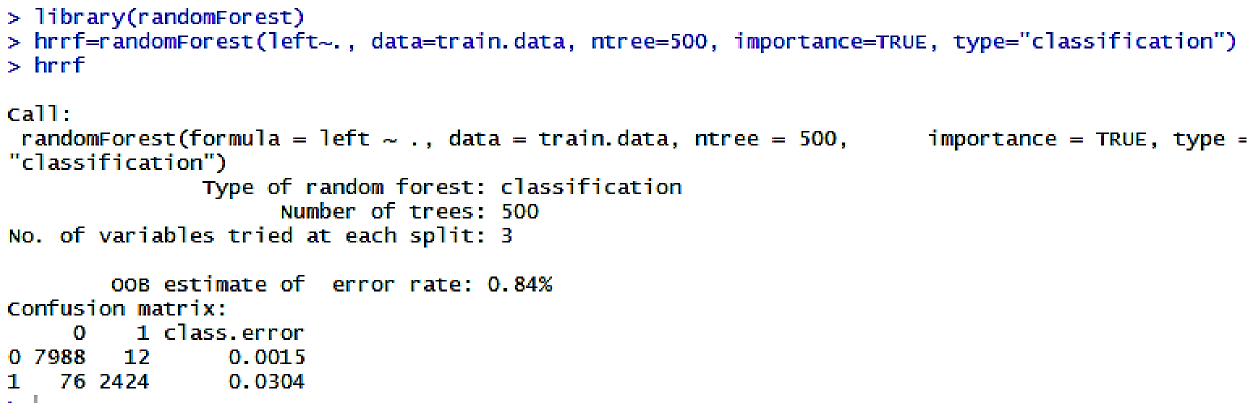
```
> summary(hrlr.step)

Call:
glm(formula = left ~ satisfaction_level + last_evaluation + number_project +
    average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years + sales, family = "binomial", data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3379  -0.6865  -0.4380  -0.1473   3.1990

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.0878485  0.1543898   0.569   0.5694
satisfaction_level    -4.1090704  0.1155515 -35.560  < 2e-16 ***
last_evaluation        0.7354091  0.1729936   4.251 2.13e-05 ***
number_project        -0.3036486  0.0248224 -12.233  < 2e-16 ***
average_montly_hours   0.0043508  0.0006013   7.236 4.62e-13 ***
time_spend_company     0.2217329  0.0173646  12.769  < 2e-16 ***
Work_accident         -1.4460440  0.1048589 -13.790  < 2e-16 ***
promotion_last_5years -2.0946082  0.3634292  -5.763 8.24e-09 ***
sales                  0.0206057  0.0093683   2.200   0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11526.4  on 10499  degrees of freedom
Residual deviance:  9359.1  on 10491  degrees of freedom
AIC: 9377.1

Number of Fisher Scoring iterations: 6
```



```
> accuracy(test.data$left, lr_pred, threshold = 0.6)
  threshold       AUC omission.rate sensitivity specificity prop.correct     Kappa
1       0.6 0.5693611     0.8253968   0.1746032    0.964119    0.7761725 0.1835624
```

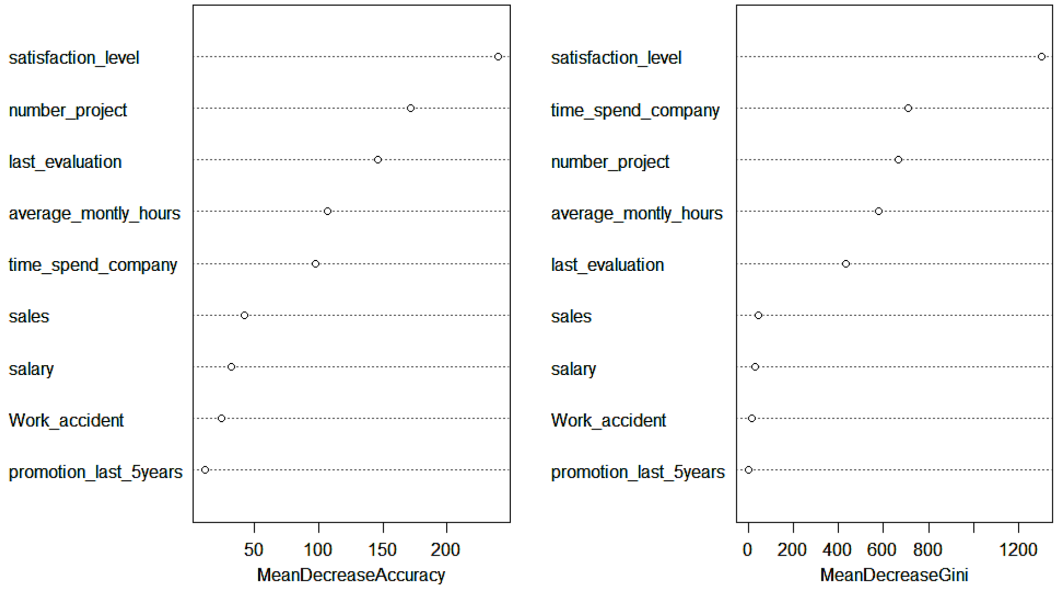
- The step model does not improve over the linear regression model.
- The accuracy on the test dataset is 0.776.
- The area under the curve also does not show any increase with the step model.

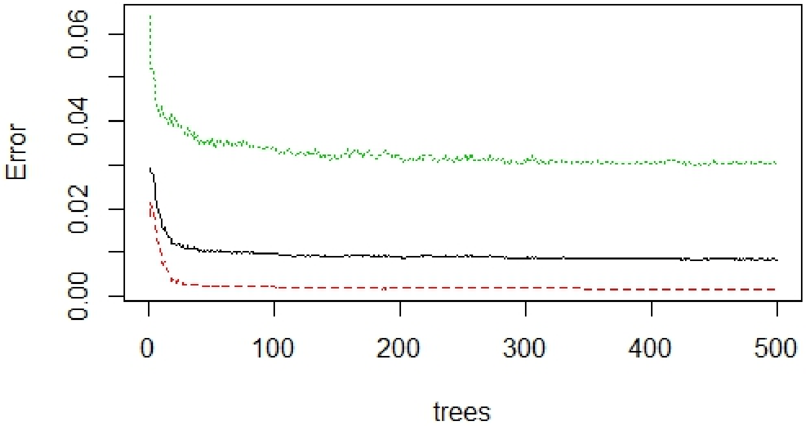
RANDOM FOREST:



```
> library(randomForest)
> hrrf=randomForest(left~., data=train.data, ntree=500, importance=TRUE, type="classification")
> hrrf

Call:
 randomForest(formula = left ~ ., data = train.data, ntree = 500,      importance = TRUE, type = 
"classification") 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.84%
Confusion matrix:
     0    1 class.error
0 7988   12      0.0015
1   76 2424      0.0304
```

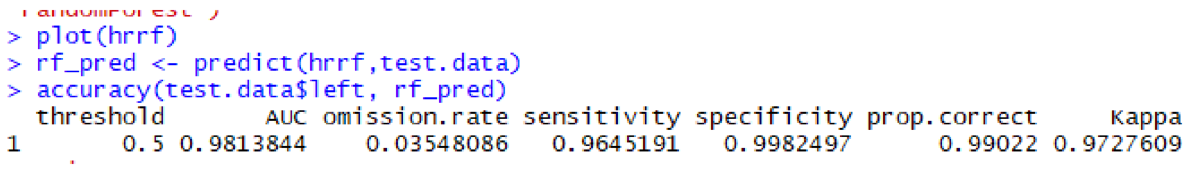
- The error rate is very low indicating a good fit on the train set.



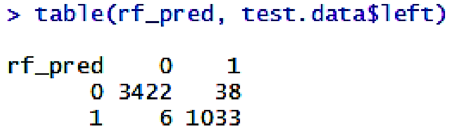


- The green and the red lines show how the error rate has decreased with the number of trees for the variable ‘left’ which has two levels, 0 & 1.

- The black line shows the overall decrease in error for the model.



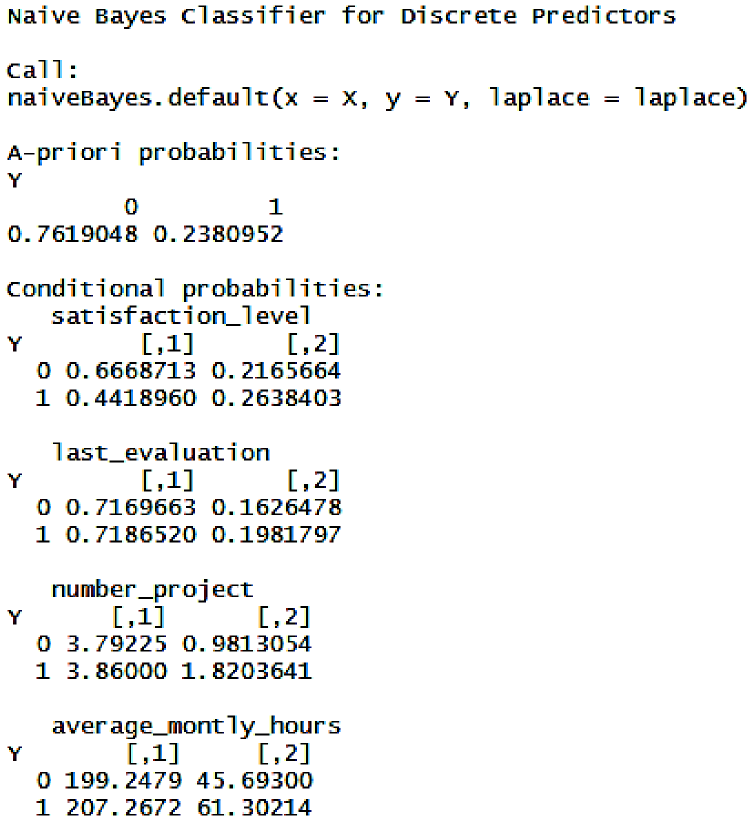
```
randomForest )
> plot(hrrf)
> rf_pred <- predict(hrrf,test.data)
> accuracy(test.data$left, rf_pred)
  threshold       AUC omission.rate sensitivity specificity prop.correct     Kappa
1       0.5 0.9813844    0.03548086   0.9645191   0.9982497      0.99022 0.9727609
```



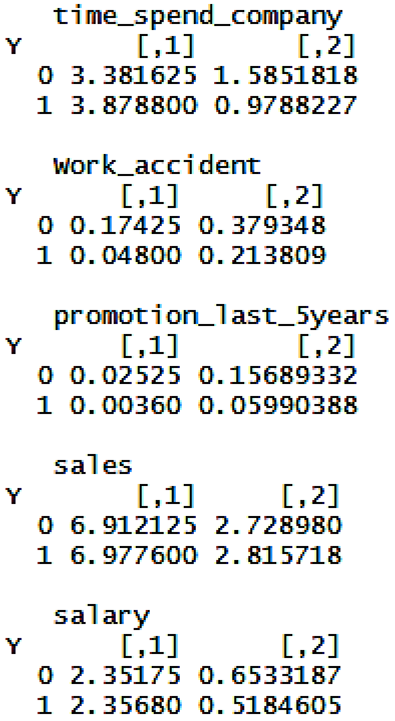
```
> table(rf_pred, test.data$left)

rf_pred    0    1
      0 3422   38
      1    6 1033
```

- The accuracy on the test set is 0.99.

NAÏVE BAYES:

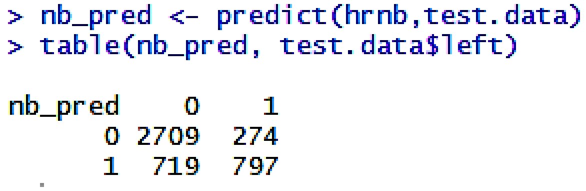


```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.7619048 0.2380952 

Conditional probabilities:
   satisfaction_level
Y        [,1]      [,2]
  0 0.6668713 0.2165664
  1 0.4418960 0.2638403

   last_evaluation
Y        [,1]      [,2]
  0 0.7169663 0.1626478
  1 0.7186520 0.1981797

   number_project
Y      [,1]      [,2]
  0 3.79225 0.9813054
  1 3.86000 1.8203641

   average_montly_hours
Y       [,1]     [,2]
  0 199.2479 45.69300
  1 207.2672 61.30214
```



```
   time_spend_company
Y       [,1]      [,2]
  0 3.381625 1.5851818
  1 3.878800 0.9788227

   Work_accident
Y      [,1]     [,2]
  0 0.17425 0.379348
  1 0.04800 0.213809

   promotion_last_5years
Y      [,1]       [,2]
  0 0.02525 0.15689332
  1 0.00360 0.05990388

   sales
Y       [,1]     [,2]
  0 6.912125 2.728980
  1 6.977600 2.815718

   salary
Y      [,1]      [,2]
  0 2.35175 0.6533187
  1 2.35680 0.5184605
```

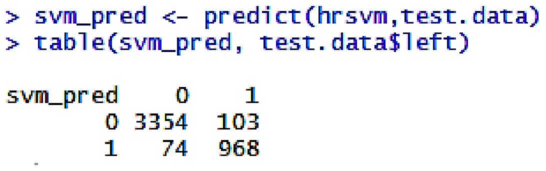
- The Naïve Bayes output gives the A-priori probabilities for left as 0.762 and 0.24.



```
> nb_pred <- predict(hrnb,test.data)
> table(nb_pred, test.data$left)

nb_pred    0    1
      0 2709  274
      1  719  797
```

- The predictions on the test set give an accuracy of 0.779.

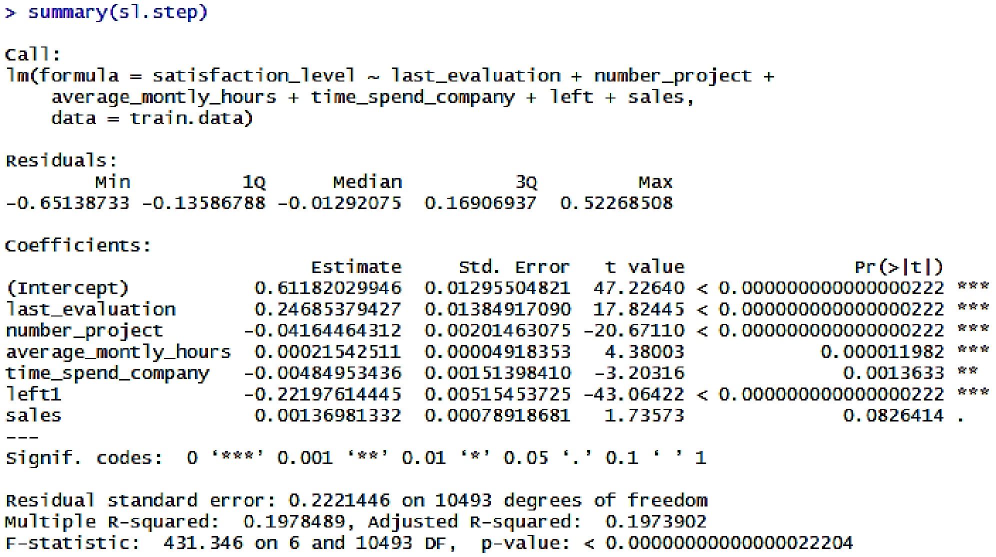
SUPPORT VECTOR MACHINE:



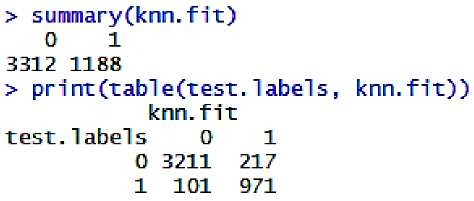
```
> svm_pred <- predict(hrsvm,test.data)
> table(svm_pred, test.data$left)

svm_pred    0    1
       0 3354  103
       1   74  968
```

- The SVM model on the test dataset gives an accuracy of 0.96.

K NEAREST NEIGHBOR:



```
> summary(sl.step)

Call:
lm(formula = satisfaction_level ~ last_evaluation + number_project + 
    average_montly_hours + time_spend_company + left + sales, 
    data = train.data)

Residuals:
        Min          1Q      Median          3Q         Max 
-0.65138733 -0.13586788 -0.01292075  0.16906937  0.52268508 

Coefficients:
                         Estimate     Std. Error   t value              Pr(>|t|)    
(Intercept)            0.61182029946  0.01295504821  47.22640 < 0.000000000000000222 ***
last_evaluation        0.24685379427  0.01384917090  17.82445 < 0.000000000000000222 ***
number_project        -0.04164464312  0.00201463075 -20.67110 < 0.000000000000000222 ***
average_montly_hours   0.00021542511  0.00004918353   4.38003            0.000011982 ***
time_spend_company    -0.00484953436  0.00151398410  -3.20316              0.0013633 ** 
left1                 -0.22197614445  0.00515453725 -43.06422 < 0.000000000000000222 ***
sales                  0.00136981332  0.00078918681   1.73573              0.0826414 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2221446 on 10493 degrees of freedom
Multiple R-squared:  0.1978489, Adjusted R-squared:  0.1973902 
F-statistic:  431.346 on 6 and 10493 DF,  p-value: < 0.0000000000000022204
```

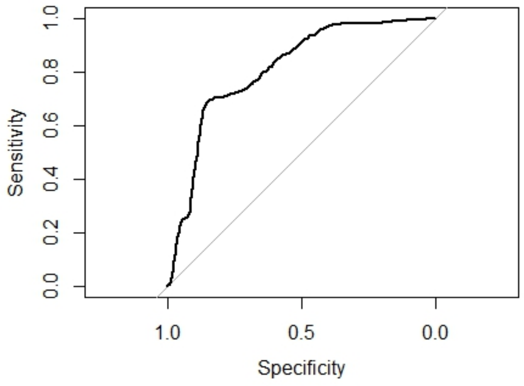


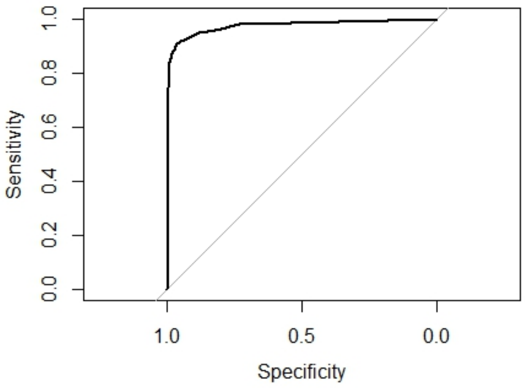
```
> summary(knn.fit)
   0    1 
3312 1188 
> print(table(test.labels, knn.fit))
           knn.fit
test.labels    0    1
          0 3211  217
          1  101  971
```

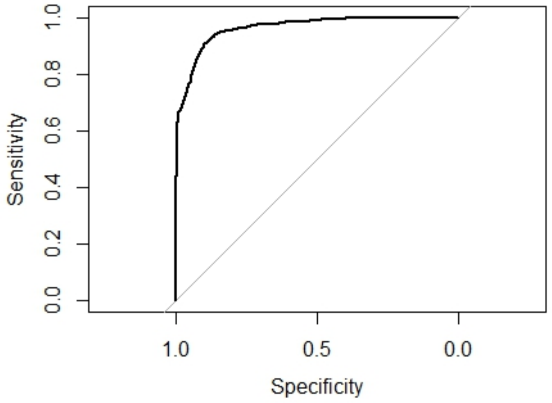
- The KNN model on the test dataset gives an accuracy of 0.93.
- The k-value was set to 10.

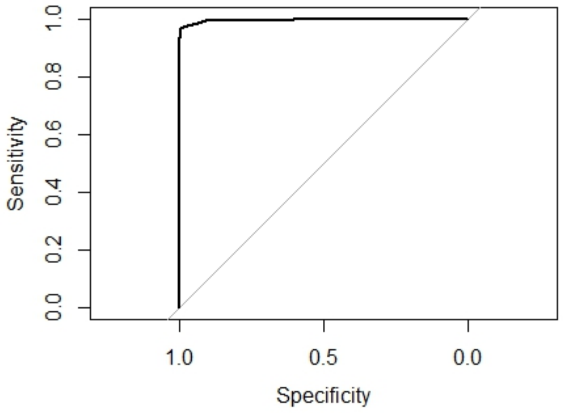
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE:

The ROC curves for logistic regression, k-nearest neighbors, naïve bayes and random forest in counter clockwise direction. Random forest has the highest ROC value.





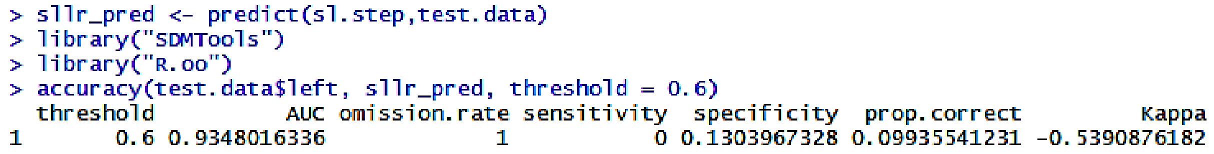




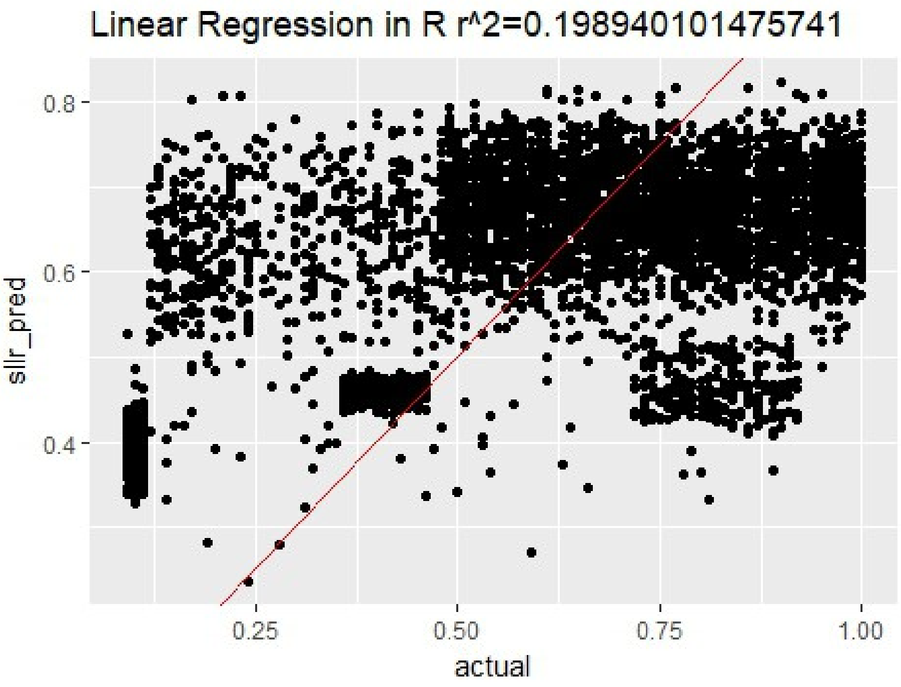
DEPENDENT VARIABLE – SATISFACTION LEVEL:

LINEAR REGRESSION – STEP MODEL:

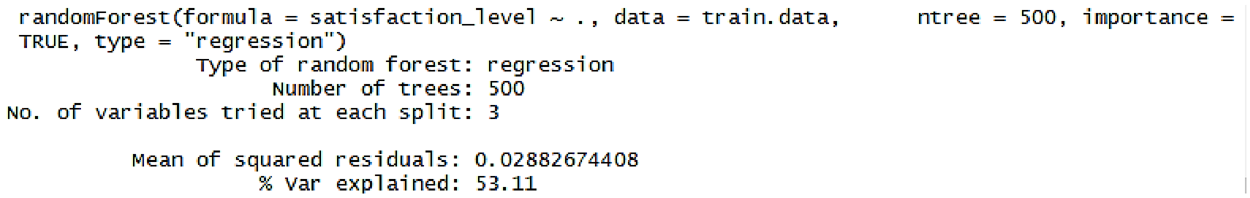
- The linear regression model with satisfaction level as the dependent variable gave an r-squared of 0.1978.



```
> sllr_pred <- predict(sl.step,test.data)
> library("SDMTools")
> library("R.oo")
> accuracy(test.data$left, sllr_pred, threshold = 0.6)
  threshold          AUC omission.rate sensitivity  specificity  prop.correct         Kappa
1       0.6 0.9348016336             1           0 0.1303967328 0.09935541231 -0.5390876182
```

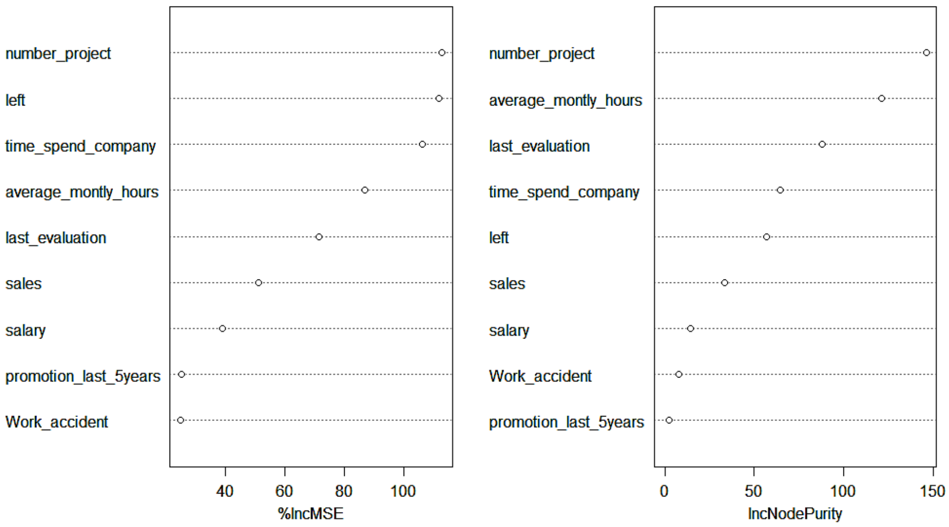


RANDOM FOREST:

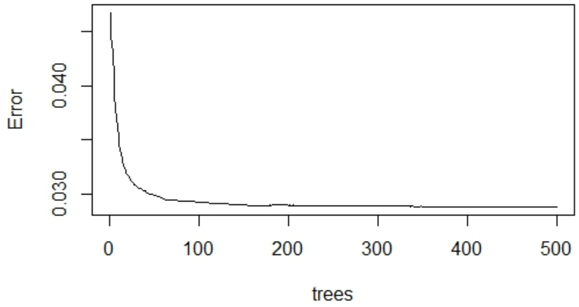


```
 randomForest(formula = satisfaction_level ~ ., data = train.data,      ntree = 500, importance =
 TRUE, type = "regression")
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.02882674408
                    % Var explained: 53.11
```

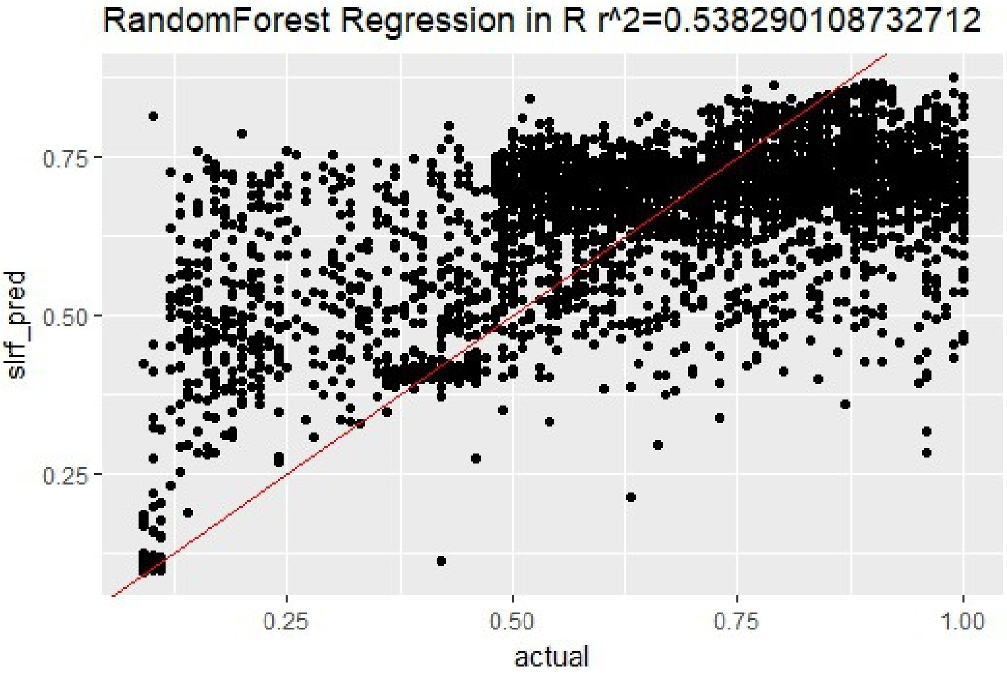
- The random forest model explains 53.11 percentage of variation in the data.



- IncNodePurity relates to the loss function which by best splits are chosen. It is gini-impurity for classification.
- %Inc MSE shows how a variable is assigned values by random permutation and by how much will the MSE increase.

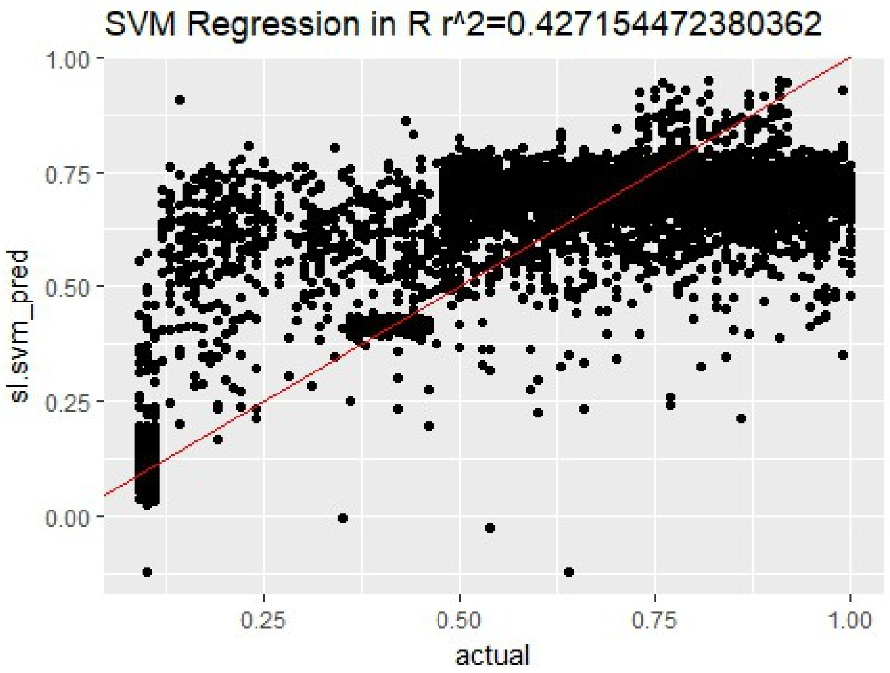


- The error decreases as we go towards 500 trees.



- The r-squared value for the predictions on the test dataset is 0.538.

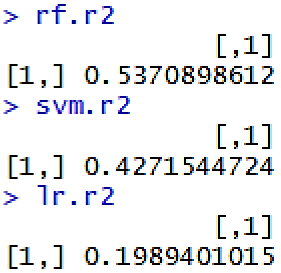
SUPPORT VECTOR MACHINE:



- The r-squared value for the predictions on the test dataset is 0.537.

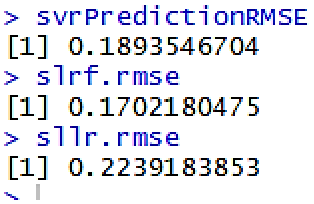
STATISTIC MEASURES:

- The r-squared for the predictions on test set using random forest, SVM and linear regression models are as below.



```
> rf.r2
             [,1]
[1,] 0.5370898612
> svm.r2
             [,1]
[1,] 0.4271544724
> lr.r2
             [,1]
[1,] 0.1989401015
```

- The RMSE value for the test set using SVM, random forest and linear regression are as below.



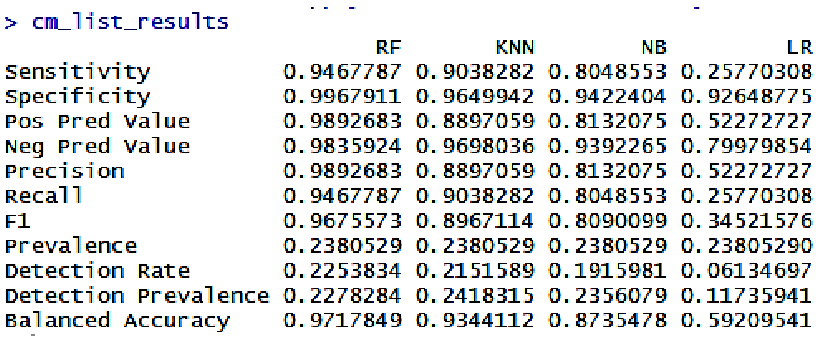
```
> svrPredictionRMSE
[1] 0.1893546704
> slrf.rmse
[1] 0.1702180475
> sllr.rmse
[1] 0.2239183853
>
```

5. CONCLUSIONS AND DISCUSSION

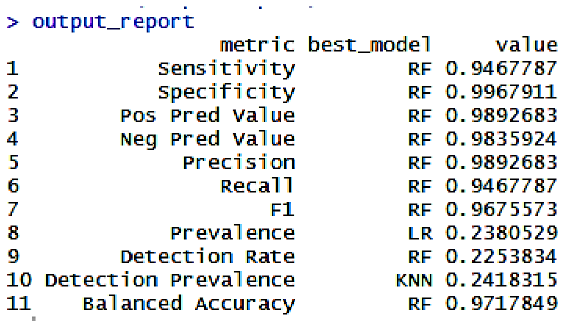
FOR ‘LEFT’ AS THE DEPENDENT VARIABLE:

MODEL	TEST SET ACCURACY
Logistic Regression	0.775

Step Model	0.776
Naïve Bayes	0.779
K Nearest Neighbor	0.93
Support Vector Machine	0.96
Random Forest	0.99



```
> cm_list_results
                             RF       KNN        NB          LR
Sensitivity          0.9467787 0.9038282 0.8048553 0.25770308
Specificity          0.9967911 0.9649942 0.9422404 0.92648775
Pos Pred Value       0.9892683 0.8897059 0.8132075 0.52272727
Neg Pred Value       0.9835924 0.9698036 0.9392265 0.79979854
Precision            0.9892683 0.8897059 0.8132075 0.52272727
Recall               0.9467787 0.9038282 0.8048553 0.25770308
F1                   0.9675573 0.8967114 0.8090099 0.34521576
Prevalence           0.2380529 0.2380529 0.2380529 0.23805290
Detection Rate       0.2253834 0.2151589 0.1915981 0.06134697
Detection Prevalence 0.2278284 0.2418315 0.2356079 0.11735941
Balanced Accuracy    0.9717849 0.9344112 0.8735478 0.59209541
```



```
> output_report
                 metric best_model     value
1           Sensitivity         RF 0.9467787
2           Specificity         RF 0.9967911
3        Pos Pred Value         RF 0.9892683
4        Neg Pred Value         RF 0.9835924
5             Precision         RF 0.9892683
6                Recall         RF 0.9467787
7                    F1         RF 0.9675573
8            Prevalence         LR 0.2380529
9        Detection Rate         RF 0.2253834
10 Detection Prevalence        KNN 0.2418315
11    Balanced Accuracy         RF 0.9717849
```

- It can be said that the random forest model predicts better than any other.
- SVM & KNN also have a high accuracy on test set.
- Satisfaction level, number of project, average monthly working hours and last evaluation are the most important factors for an employee leaving a company.
- This was asserted in all the models.
- Random forest has high specificity and sensitivity meaning that it predicts those who left the company and those who stayed back correctly.

FOR ‘SATISFACTION LEVEL’ AS THE DEPENDENT VARIABLE:

MODEL	RMSE	R-SQUARED
RANDOM FOREST	0.17	0.537
SVM	0.189	0.427
LINEAR REGRESSION	0.223	0.199

- It can be that the random forest model gave the best RMSE value.

- Number of projects, left the company or not, time spent in the company and average monthly working hours are the motivating factors for levels of satisfaction.
- From the R-squared value, it can be seen that random forest explains the data much better than the other models.

SUMMARY:

- Though the dataset was predicting the variable left/not in a well-defined manner, it was not able to do so for other variables.
- The models for satisfaction level revealed that they need additional factors not captured in this dataset to understand the underlying picture for satisfaction.
- In subsequent exploration, using additional variables to gauge the levels of satisfaction would be taken care of.

References:

1. http://www.allresearchjournal.com/archives/2015/vol1issue9/PartK/1-9-143.pdf
2. https://www.kaggle.com/ludobenistant/hr-analytics

Packages used:

Packages used were corrplot, caret, SDMTools, R.oo, randomForest, e1071, class, Metrics, miscTools, pROC, arm, nnet, ggthemes, ggplot2, grid, gridExtra