Human Capital Analysis

Shanmugan Kukatla

Webster University

CSDA 6010 – Analytics Practicum

Professor Dr. Jiangping Wang

Dec 11th, 2024

Table of Contents

Table of Figures

Executive Summary

This project uses human capital analytics to address employee turnover by identifying key factors and predictors. Machine learning algorithms, particularly classification models like logistic regression, Naive Bayes, and regression trees, were used to predict employee turnover. The process involved exploratory data analysis, modeling, and clustering to uncover patterns and drivers of employee attrition. Despite being a classification technique, logistic regression was optimized to distinguish between employees likely to leave and those not. Alternative models were explored for comparison, but logistic regression best-predicted turnover. Clustering, such as K-Means, complemented the analysis by identifying the critical attributes and factors contributing to turnover. These insights enable targeted strategies to retain employees at risk of leaving.

Introduction

Human Resources are crucial to an organization's growth and long-term sustainability. Engaged and productive employees are key to maintaining a competitive edge in the market. Consequently, companies are prioritizing strategies to retain their most skilled and experienced workforce. This project utilizes a medium-sized engineering company's simulated Human Capital Analytics dataset to explore the reasons behind employee turnover. The analysis begins with descriptive and exploratory data techniques to understand patterns in the data. It progresses to predictive modeling to classify employees likely to leave the organization and employs clustering to identify the key factors influencing turnover. Finally, actionable recommendations are developed to effectively address the business challenge of employee retention.

Business Problem

Employee turnover poses significant challenges to organizations, including high costs (recruitment, training, productivity losses), loss of institutional knowledge, and negative impacts on innovation and competitiveness. It can severely hinder organizational growth or expose the company to financial risks, particularly if employees join competitors.

Business Goal

The goal is to **minimize employee turnover** by identifying and addressing the factors influencing their decisions to leave. Predicting potential attrition empowers the Human Resources (HR) department to implement proactive retention strategies, reducing associated risks and costs.

Analytical Goal

The analytical goal is to align predictive insights with the business objective of minimizing employee turnover and its associated challenges. By leveraging data-driven techniques, the aim is to **identify employees at risk** of leaving and **uncover** the **underlying factors influencing attrition.** These insights enable the HR team to implement proactive and targeted retention strategies, fostering a stable and productive workforce.

Analytical Approach

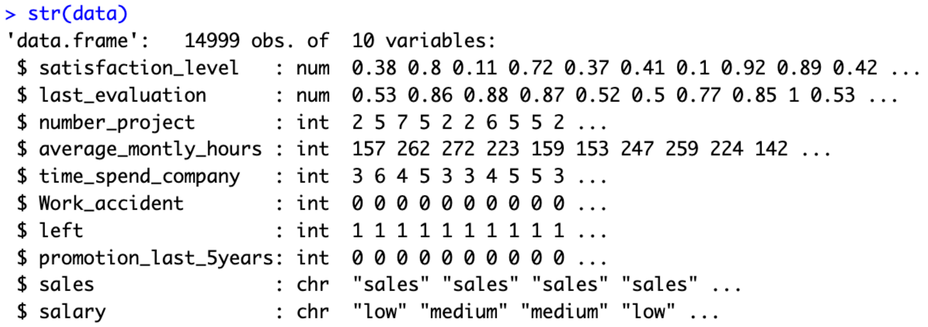
The analytical approach begins with data exploration and hypothesis testing to uncover initial patterns and validate assumptions about employee attrition. This phase examines data trends, distributions, and relationships to identify potential factors influencing turnover. Hypothesis testing is used to statistically confirm or reject these initial insights, providing a solid foundation for the subsequent analysis.

Building on these findings, classification models such as logistic regression, classification trees, and Naive Bayes are used to **predict which employees are likely to leave**. Clustering techniques like K-Means are then applied to group employees based on shared traits and turnover drivers. Together, these methods enable a deeper **understanding of attrition patterns** and support the design of targeted retention strategies aligned with business objectives.

Data Preprocessing

Data Understanding:

The Human Capital dataset is a simulated dataset from a medium-sized engineering company. It has 14,999 observations and 10 fields, which are combinations of variable types of integers, float, and char.



```
> str(data)
'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                : chr  "sales" "sales" "sales" "sales" ...
 $ salary               : chr  "low" "medium" "medium" "low" ...
```

FIGURE 1 : STRUCTURE OF THE DATASET

Attribute Definition:

Column Name	Description
'satisfaction_level'	The employee's level of job satisfaction typically ranges from 0 to 1, where 1 indicates high satisfaction.
'last_evaluation'	The employee's most recent performance evaluation score usually ranges from 0 to 1.
'number_project'	The number of projects the employee was involved in during their tenure.
'average_montly_hours'	The average number of hours worked per month by the employee.
'time_spend_company'	The number of years the employee has been with the company.
'work_accident'	Indicates whether the employee has had a work-related accident (1 if yes, 0 if no).
'left'	Indicates whether the employee has left the company (1 if yes, 0 if no)
'promotion_last_5years'	Indicates whether the employee was promoted in the last five years (1 if yes, 0 if no)
'department'	The department field is a department in which the employee works, such as "sales," "technical," "support," etc.
'salary'	The salary level of the employee is typically categorized as "low," "medium," or "high."

FIGURE 2 : ATTRIBUTE DEFINITION

Categorical data:

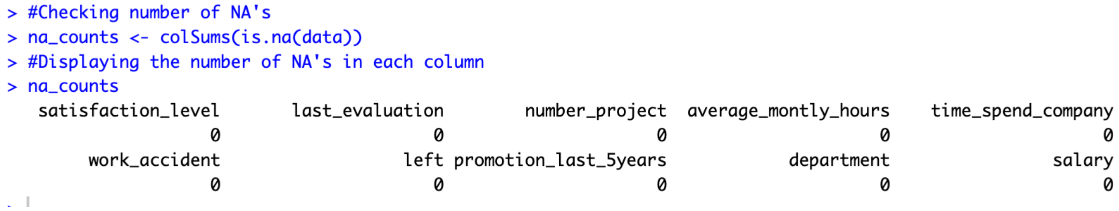
There are five categorial fields (work_accident, left, promotion_last_5years, department and salary). The sales(initially) field is a department in which the employee works, such as "sales," "technical," "support," etc. This field is later renamed as department, which feels more appropriate than "sales." The Salary field, which is the salary level of the employee, is typically categorized as "low," "medium," or "high.". The fields (work_accident, left, promotion_last_5years) hold 1 yes and 0 no, respectively.

Numeric Data:

There are two numeric (continuous) fields, satisfaction_level and last_evaluation, ranging from (0.09 - 1.0) and (0.36 – 1.0), respectively, and three integer fields (number_project, average_monthly_hours, and time_spend_company).

Check for missing values:

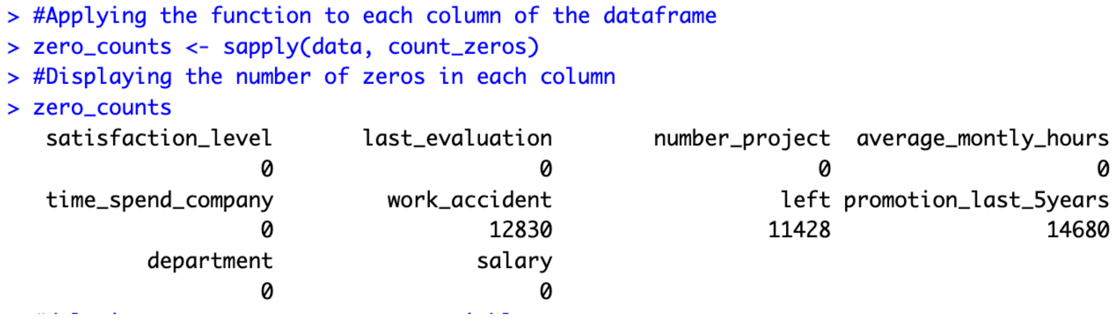
Check for Null values: Observed **no null values in** any of the columns.



```
> #Checking number of NA's
> na_counts <- colSums(is.na(data))
> #Displaying the number of NA's in each column
> na_counts
   satisfaction_level       last_evaluation        number_project  average_montly_hours    time_spend_company 
                    0                     0                     0                     0                     0 
        work_accident                  left promotion_last_5years            department                salary 
                    0                     0                     0                     0                     0 
```

FIGURE 3 : CHECKING FOR NA'S

Check for zeroes:



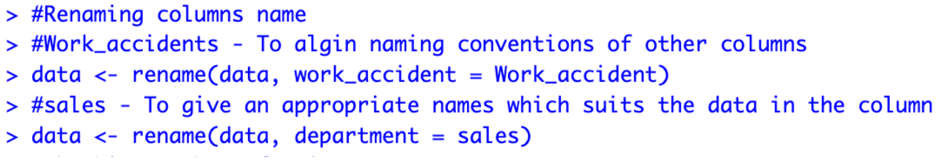
```
> #Applying the function to each column of the dataframe
> zero_counts <- sapply(data, count_zeros)
> #Displaying the number of zeros in each column
> zero_counts
   satisfaction_level       last_evaluation        number_project  average_montly_hours 
                    0                     0                     0                     0 
   time_spend_company         work_accident                  left promotion_last_5years 
                    0                 12830                 11428                 14680 
           department                salary 
                    0                     0 
```

FIGURE 4 : CHECKING FOR ZEROES

Although there are zeroes in work_accident, left, and promotion_5years, it is acceptable as they are binary; it does not mean the data is inconsistent.

Data Manipulation:

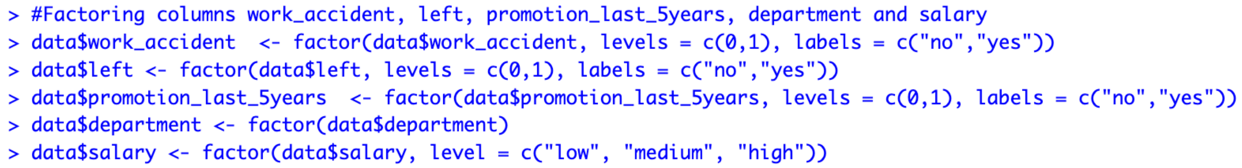
Renaming columns Work_accident and sales as "work_accident" and "department" will establish a similar naming convention among fields and give a meaningful name that reassembles the data in fields.



```
> #Renaming columns name
> #Work_accidents - To algin naming conventions of other columns
> data <- rename(data, work_accident = Work_accident)
> #sales - To give an appropriate names which suits the data in the column
> data <- rename(data, department = sales)
```

FIGURE 5 : RENAMING THE FIELDS OF DATASET

All five categorical fields were factored into the proper order specified for each column for modeling.

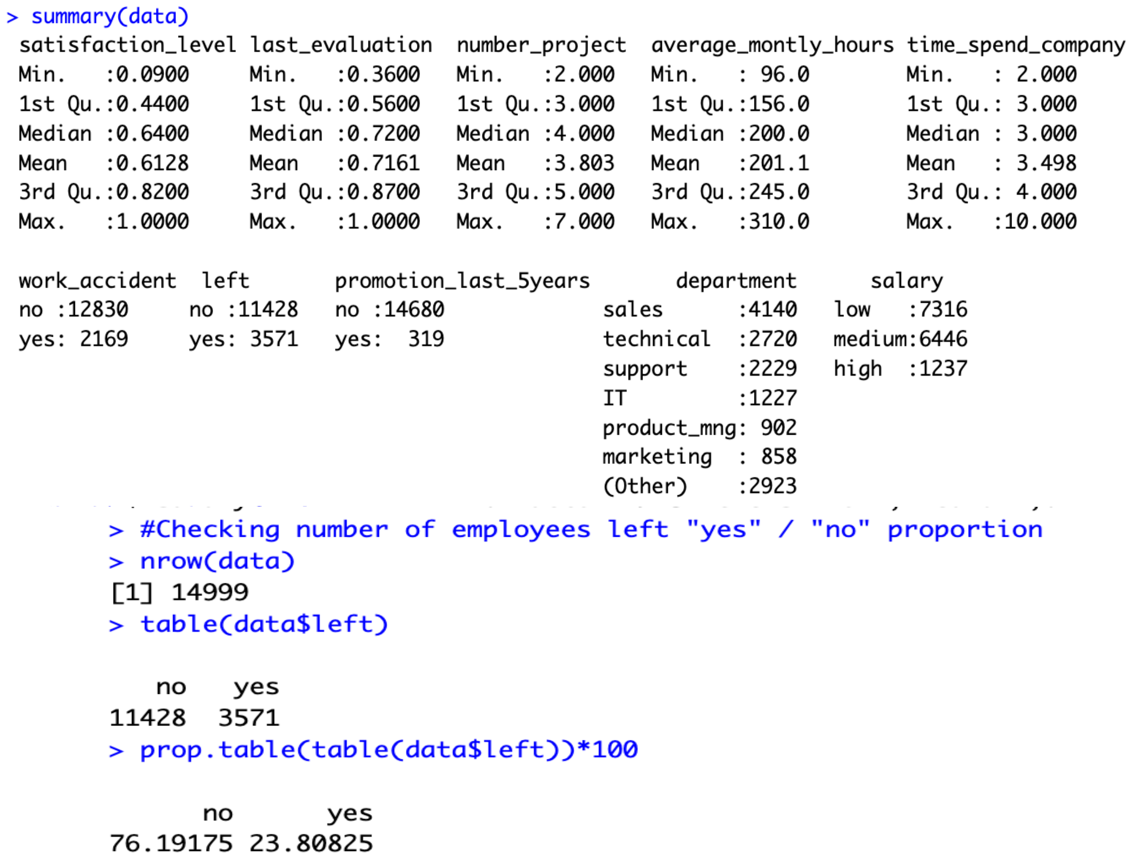


```
> #Factoring columns work_accident, left, promotion_last_5years, department and salary
> data$work_accident  <- factor(data$work_accident, levels = c(0,1), labels = c("no","yes"))
> data$left <- factor(data$left, levels = c(0,1), labels = c("no","yes"))
> data$promotion_last_5years  <- factor(data$promotion_last_5years, levels = c(0,1), labels = c("no","yes"))
> data$department <- factor(data$department)
> data$salary <- factor(data$salary, level = c("low", "medium", "high"))
```

FIGURE 6 : FACTORING CATEGORICAL DATA FIELDS

Descriptive Analysis

Descriptive statistics has provided an overview of key metrics, i.e. on average, employees rate their **satisfaction** as **0.61** on a scale from 0-1, **average time spent** by an employee in the company is about **3.5 years**. Approximately **23.8%** of the employees have **left** the company**.** **Only 2.1%** of the employees were promoted in the last 5 years. Some interesting facts from the data explored are that none of the employees has left the organization less than 2 years, **least evaluation score** given to an employee was **0.36,** and there exist even employees with almost 0 (0.09) satisfaction_level, average_montly_hours by the employees mean is 201.05 hours which is way beyond the stand hours if the organization is of USA.



```
> summary(data)
 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      department      salary    
 no :12830     no :11428   no :14680             sales      :4140   low   :7316  
 yes: 2169     yes: 3571   yes:  319             technical  :2720   medium:6446  
                                                 support    :2229   high  :1237  
                                                 IT         :1227                
                                                 product_mng: 902                
                                                 marketing  : 858                
                                                 (Other)    :2923                
```

```
> #Checking number of employees left "yes" / "no" proportion
> nrow(data)
[1] 14999
> table(data$left)

   no   yes 
11428  3571 
> prop.table(table(data$left))*100

      no      yes 
76.19175 23.80825 
```

FIGURE 7 : DESCRIPTIVE STATISTICS – SUMMARY OF DATA & PROPORTION DISTRIBUTION OF EMPLOYEE LEFT COLUMN

Boxplot of the numeric columns shows **outliers** in the **time_spend_company** field. Since it's real-world data and the scenario "the scale of senior employees is always less than another group of employees" makes scenes, outliers are **not handled**. All other four fields do not have any outliers, in general, these box plots summarize the data distribution. The majority group of the employees have less than five projects.

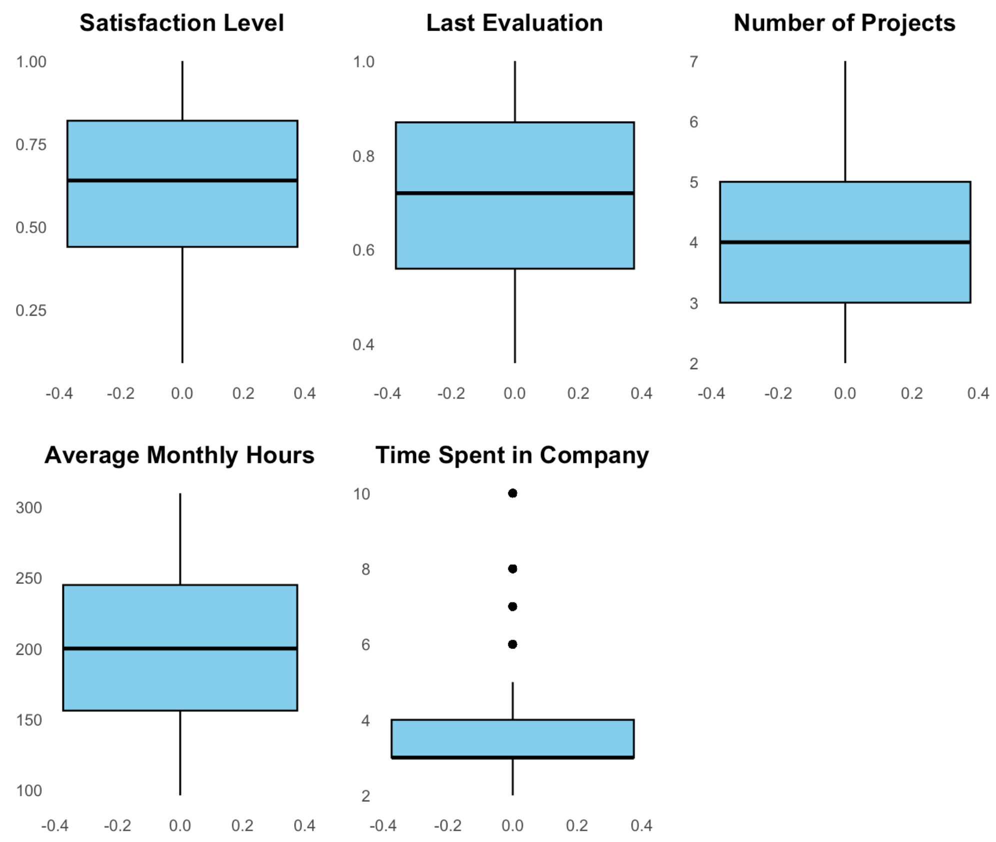


FIGURE 8 : BOX PLOT TO CHECK FOR ANY OUTLIERS

A bar plot of department-wise employee count and percentile. The **sales** department has the **most** employees, and **management** has the **fewest, 4.2%** of the overall employees.

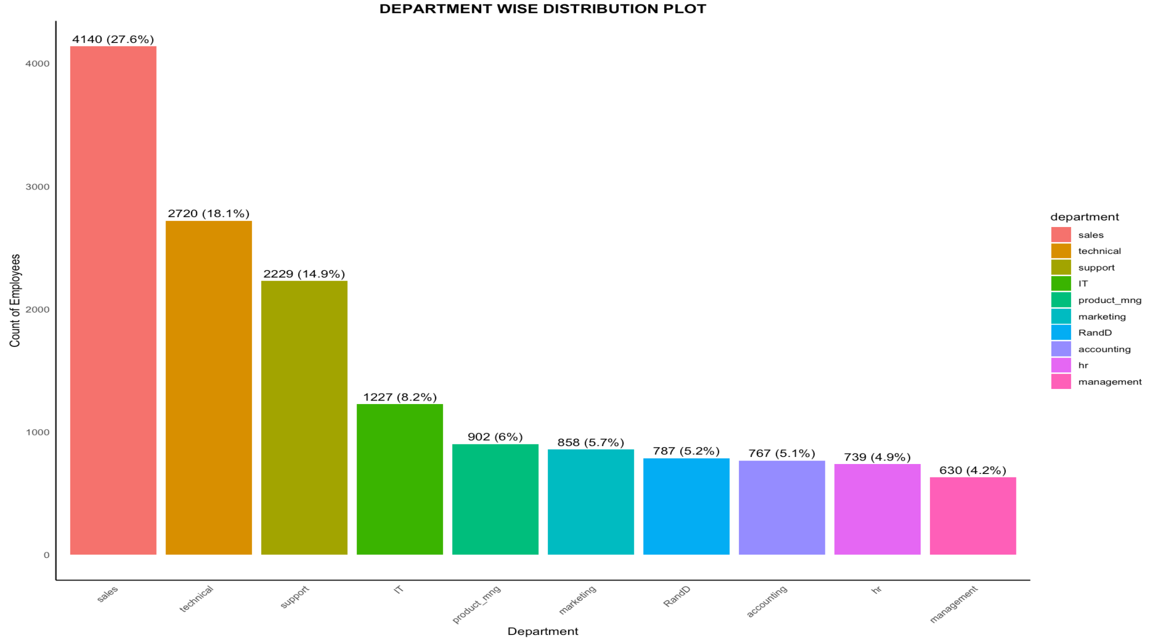


FIGURE 9 : DEPARTMENT WISE DISTRIBUTION PLOT

Approximately 49**%** of the employees belong to the low-paid category, whereas only **8.2% of the high-paid** employees in the company.

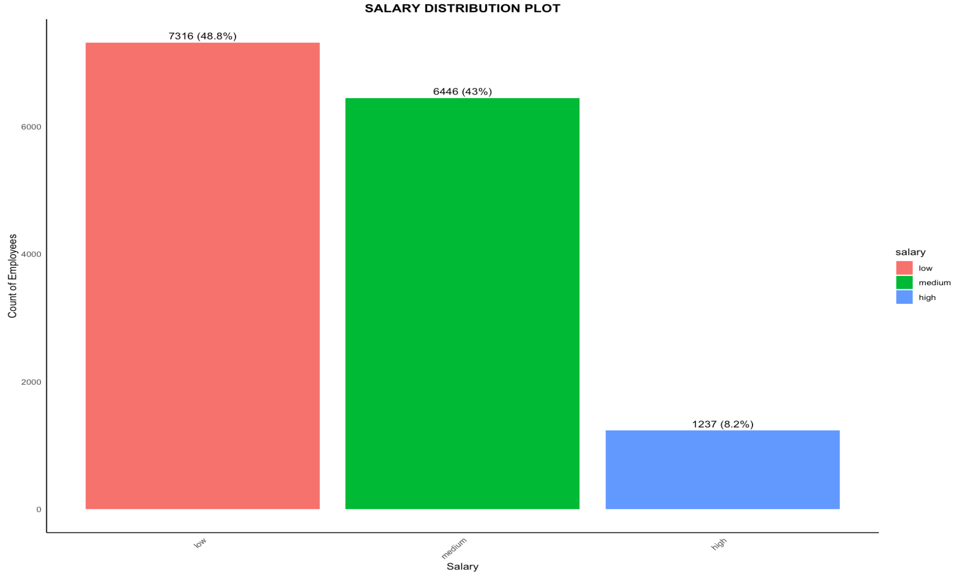


FIGURE 10 : SALARY DISTRIBUTION PLOT

Hypothesis Analysis

Hypothesis 1 - Salary is the reason why the employees left the company. Let's see if is this correct.

From the stats of the data, salary scales and employee turnover are in directly proportion. So, **yes**, salary is why employees leave the company based on the data.

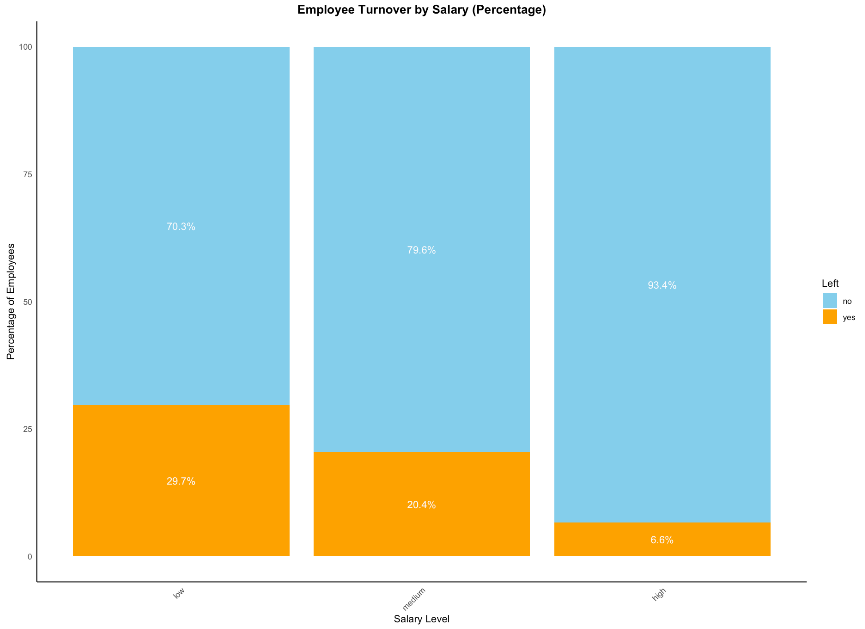


FIGURE 11 : EMPLOYEE TURNOVER BY SALARY

Hypothesis 2 - employees leave the company because work is not safe.

This hypothesis can be proved **wrong** based on the plot below, as the percentile of the employees leaving the company who did not have work accidents is higher, so work accidents might not be the reason for leaving.

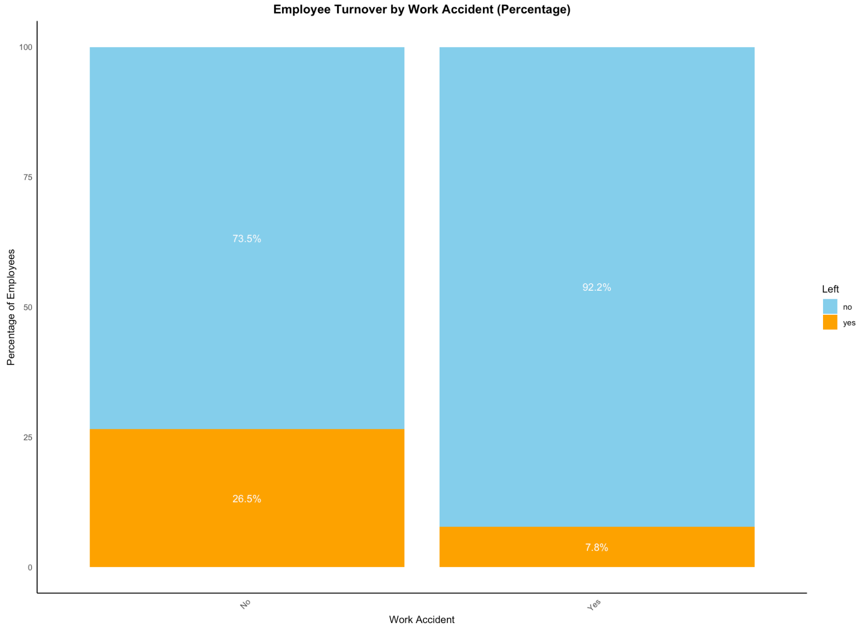


FIGURE 12 : EMPLOYEE TURNOVER BY WORK ACCIDENT

Hypothesis 3 - this company is a good place to grow professionally.

In the plot, employees promoted in the last five years are likelier to stay in the company. So, it can be inferred that employees might have believed this company was a good place to grow professionally.

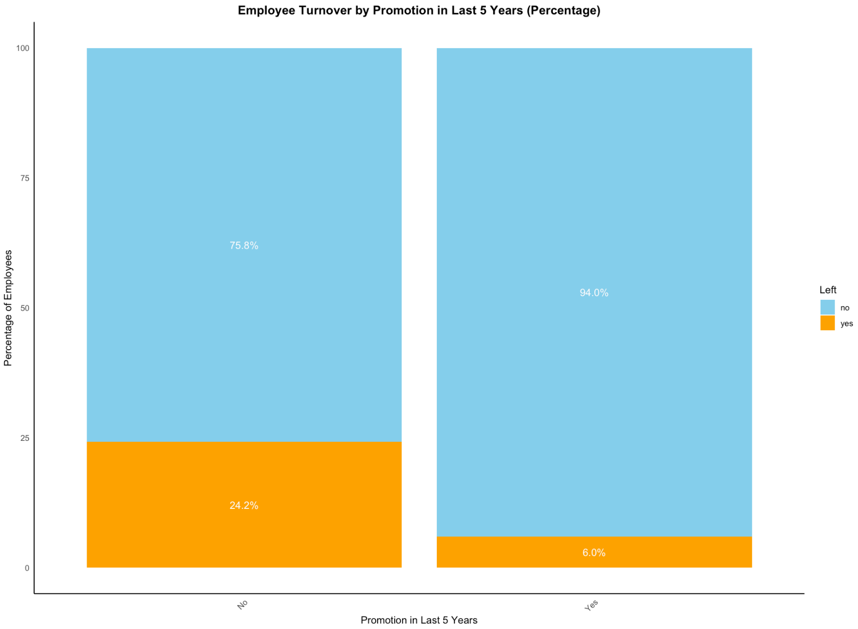


FIGURE 13 : EMPLOYEE TURNOVER BY PROMOTION

Predictor Analysis and Relevancy

In the correlation plot (among predictors), there are not highly (<-0.5 or >0.5) correlated fields. There is no high multicollinearity among predictors, so there is no need for feature elimination. However, a few slightly correlated fields tell some interesting facts.

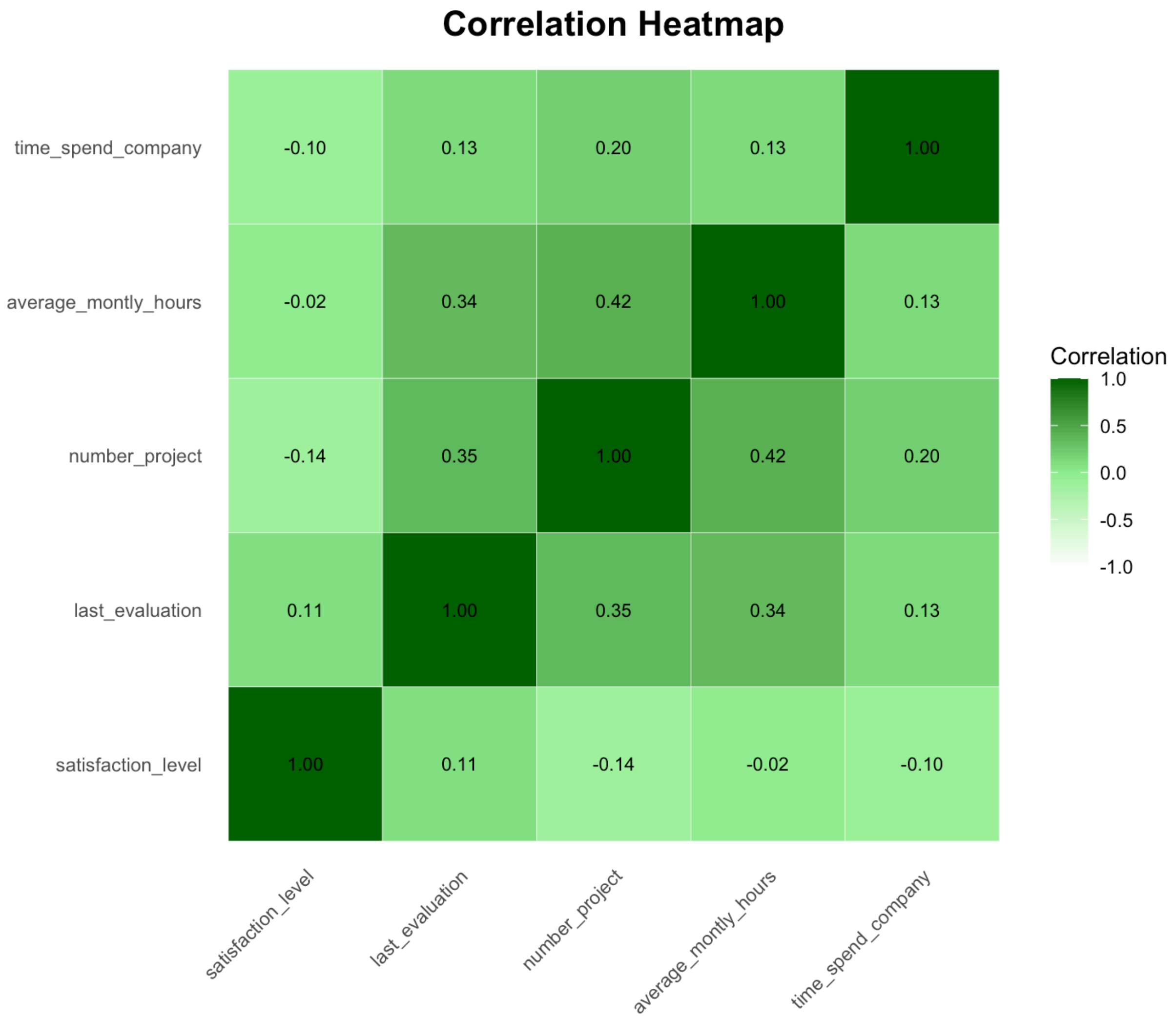


FIGURE 14 : CHECK FOR CORRELATION AMONG PREDICTORS

Firstly, employees with a greater number_of_project have higher average_monthly_hours (0.42), making more scenes for employees working on multiple projects to spend more time at work. Secondly, last_evaluation has decent correlation with number_of_project and average_monthly_hours (0.35 and 0.34 respectively), which tells employees with a more significant number of projects and average_monthly_hours trend to have good evaluation scores.

Finally, on the other hand, employees who have a more significant number of projects, average monthly hours, and time spent in the company have a negative correlation with satisfaction (-0.14, -0.02, -0.10, respectively), which suggests that such employees are not satisfied as they are indirectly proportional.

Correlate between predictors and outcome variable (left)

The plot shows that employees who left the company were less satisfied than those who stayed, as the mean satisfaction level among those who stayed and left was 0.67 and 0.44, respectively.

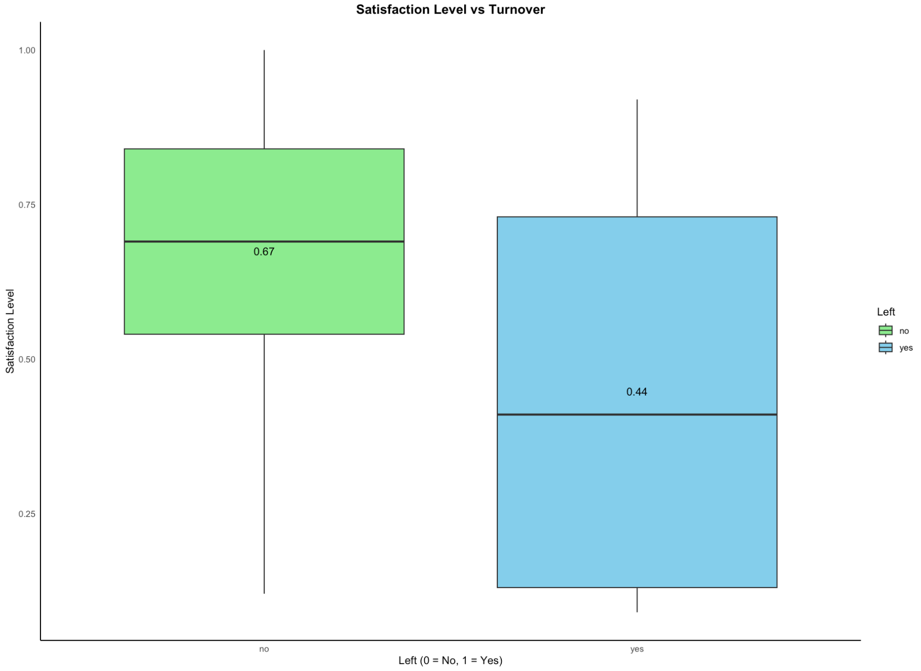


FIGURE 15 : SATISFACTION LEVEL VS TURNOVER

Turnover is higher for employees who have stayed in the company for a short time (2-5 years) compared to long-staying employees.

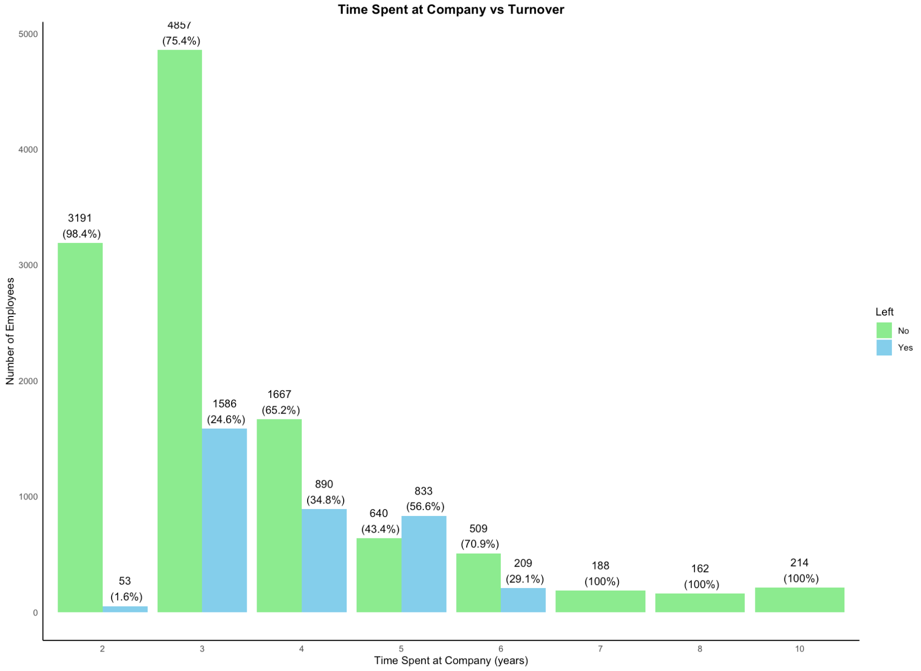


FIGURE 16 : TIME SPENT AT COMPANY VS TURNOVER

Employees who have experienced work accidents are less likely to leave compared to those who have no work accidents.

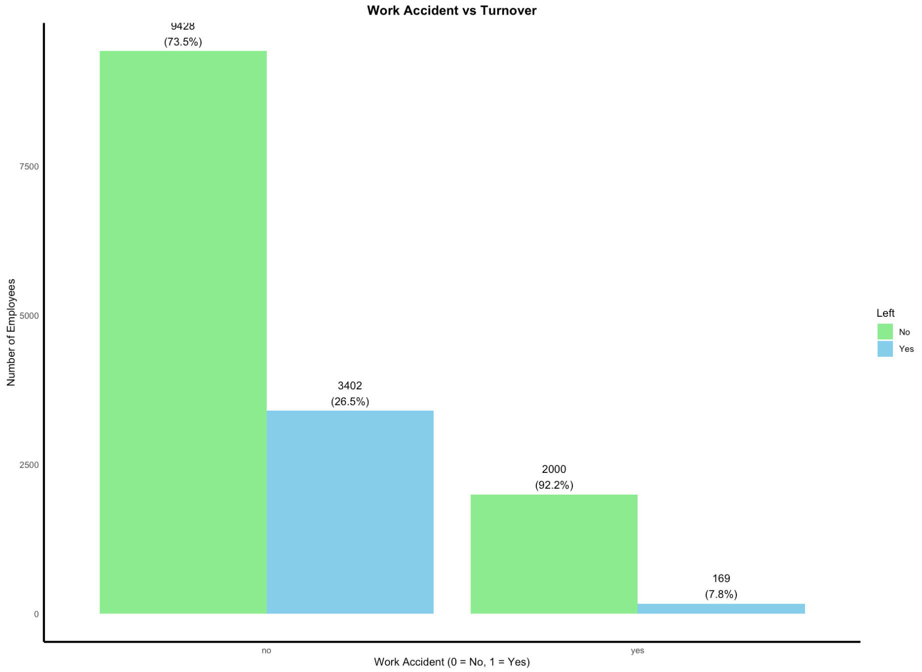


FIGURE 17 : WORK ACCIDENTS VS TURNOVER

Overall, satisfaction level is a good predictor of employee retention. Indirectly, the number of projects and average monthly hours correlate with satisfaction. However, they are negatively correlated, indicating that if an employee has a lower satisfaction level, they are likely to leave the organization which evident from below plot as well.

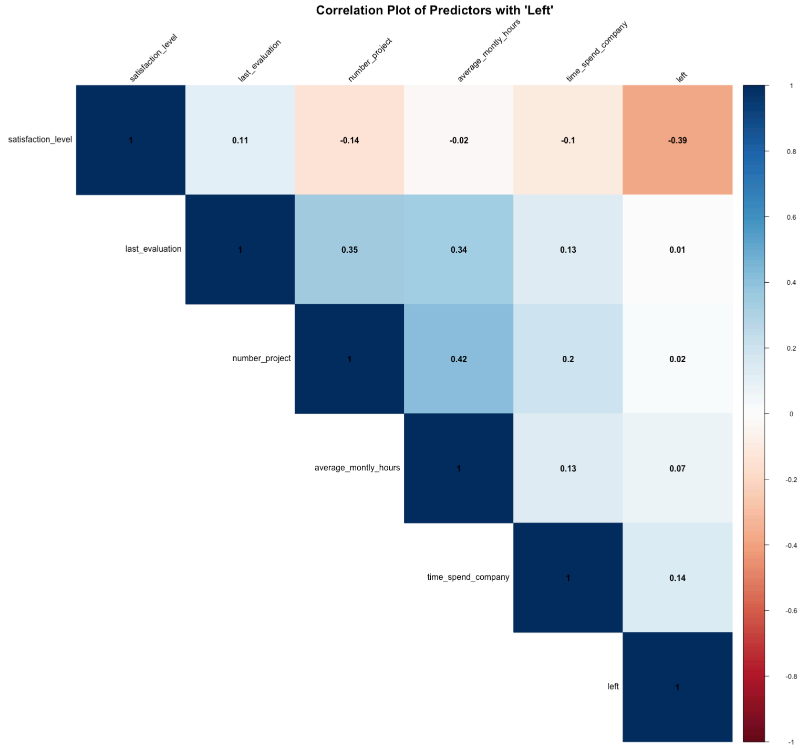
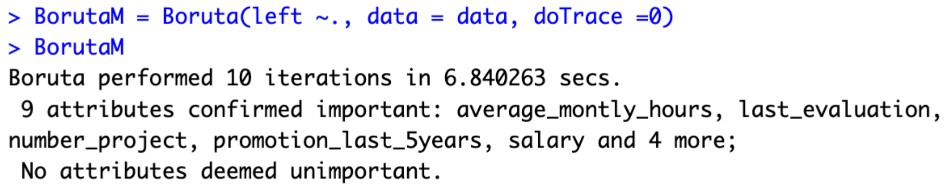


FIGURE 18 : CORRELATION PLOT OF PREDICTORS WITH OUTCOME(LEFT)

Dimension Reduction

Since there was no autocorrelation among the predictors, all of them are considered essential. Additionally, the Boruta model confirmed the importance of all predictors by identifying which variables are important, unimportant, or tentative. Therefore, dimensionality reduction is not necessary.

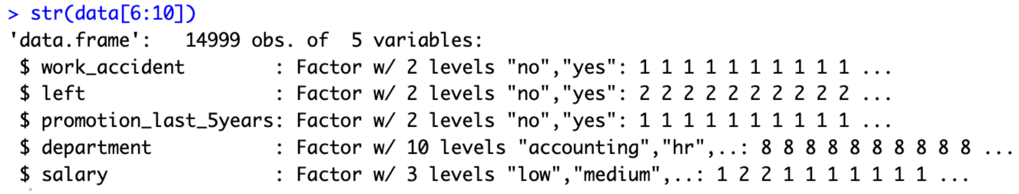


```
> BorutaM = Boruta(left ~., data = data, doTrace =0)
> BorutaM
Boruta performed 10 iterations in 6.840263 secs.
 9 attributes confirmed important: average_montly_hours, last_evaluation,
number_project, promotion_last_5years, salary and 4 more;
 No attributes deemed unimportant.
```

FIGURE 19 : BORUTA TO FIND ATTRIBUTE IMPORTANCE

Data Transformation

As part of the data transformation process for clustering, binary fields such as Work_accident, left, and promotion_last_5years were factored to appropriately handle their categorical nature, and categorical variables like department and salary were factored and converted into dummy variables. To ensure all variables were on a comparable scale, normalization was performed using the min_max() function. This technique scaled each feature to a range of 0 to 1, preserving the distribution and relationships of the data while making it suitable for K-means clustering. This comprehensive preprocessing enabled a robust and unbiased clustering analysis.



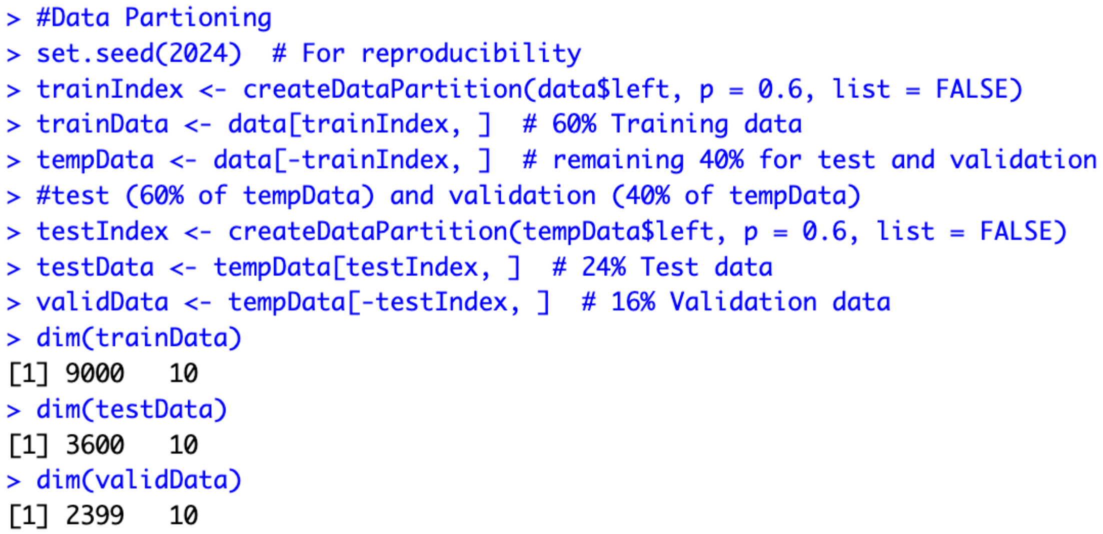
```
> str(data[6:10])
'data.frame':	14999 obs. of  5 variables:
 $ work_accident        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ left                 : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ promotion_last_5years: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ department           : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ salary               : Factor w/ 3 levels "low","medium",..: 1 2 2 1 1 1 1 1 1 1 ...
```

FIGURE 20 : DATA TRANSFORMATION

Data Partitioning methods

Data portioning is important to evaluate the model's performance on unseen data. During portioning, the data is split into two or three, i.e., train, test, and validate. If the model is built and tested on the same data, overfitting occurs, which results in poor performance on the unseen data. Sometimes, data is split into three: train, test, and validate/holdout, as train data is used to build various models, and the test is used to evaluate the performance of different models. Validate/holdout data will test how well the finally selected model performs.

A random sampling data partition method was used to split the data into 60% of training 24% testing, and 16% validation. The training dataset should always be larger than the test or validation dataset, giving the train 60% of the data. Train data is given more data as the model is built on the training dataset, allowing the model to learn more from the existing data and perform better.



```
> #Data Partioning
> set.seed(2024)  # For reproducibility
> trainIndex <- createDataPartition(data$left, p = 0.6, list = FALSE)
> trainData <- data[trainIndex, ]  # 60% Training data
> tempData <- data[-trainIndex, ]  # remaining 40% for test and validation
> #test (60% of tempData) and validation (40% of tempData)
> testIndex <- createDataPartition(tempData$left, p = 0.6, list = FALSE)
> testData <- tempData[testIndex, ]  # 24% Test data
> validData <- tempData[-testIndex, ]  # 16% Validation data
> dim(trainData)
[1] 9000   10
> dim(testData)
[1] 3600   10
> dim(validData)
[1] 2399   10
```

FIGURE 21 : DATA PARTITIONING

Model Selection

To achieve business goals, both classification and clustering models must be used. Classification modeling is a supervised learning that uses labeled data to categorize data into specific classes based on their characteristics. Clustering techniques are unsupervised learning that uses unlabeled data to group data points based on similarities, where the algorithm looks for the hidden patterns in the data.

To classify whether an employee might leave a company, **classification** models (**Naïve Bayes, Logistic regression, classification tree**) are used, and clustering techniques (K-Means clustering) are used to find factors behind an employee leaving the company, **clustering** techniques (**K-Means clustering**) are used.

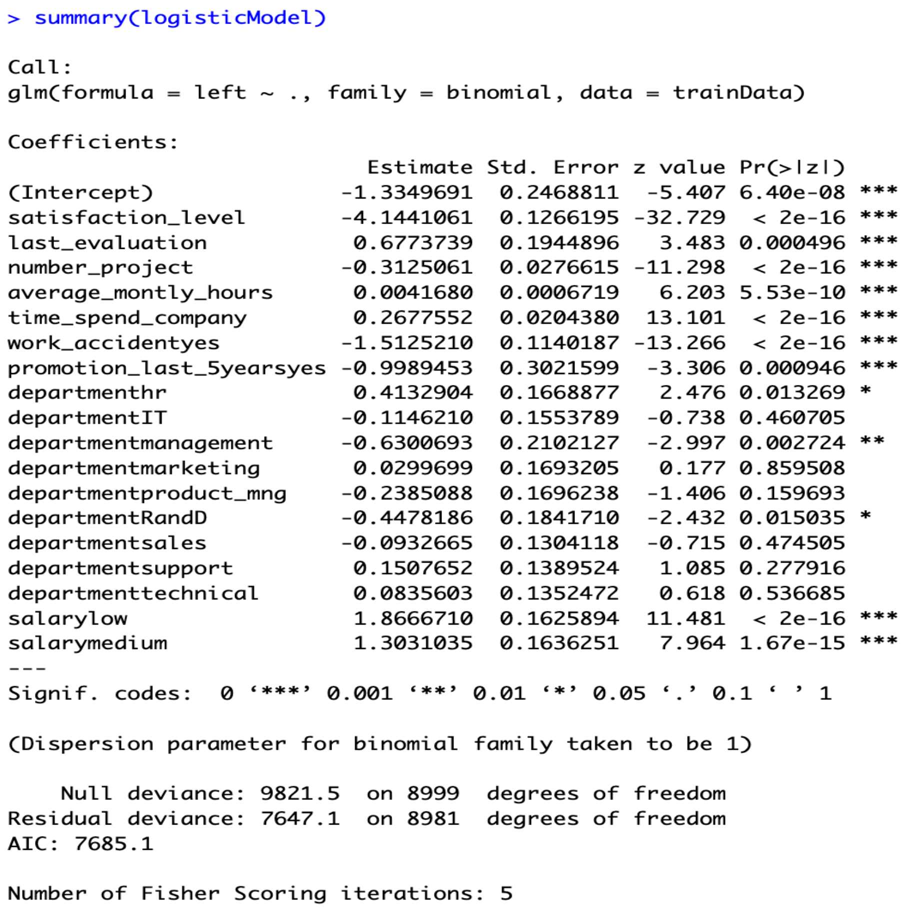
When selecting a model, the key points considered are data size, complexity, and interpretability. Interpret reasoning behind predictions is crucial for choosing logistic regression and decision tree models. These models give transparency, and easy to implement, and are good for understanding feature importance.

Modeling

Logistic Regression model

Logistic Regression model is used to predict the class of an outcome variable. In this case, the goal is to **classify employee turnover**, so left is the outcome variable, and the rest of the fields were considered as predictors. There were almost all significant predictors apart from departments. The model was built on 60% of the actual data (trainData) and was tested on 24% of actual data (testData).

The logistic regression model identifies key factors influencing employee turnover. Low satisfaction levels strongly increase the likelihood of leaving, with dissatisfaction being the most significant driver. Employees with higher last evaluation scores are also more likely to leave, suggesting that high performers may feel undervalued or seek better opportunities. A higher number of projects decreases turnover risk while working more hours slightly increases it. Longer tenure at the company raises the odds of leaving, potentially reflecting stagnation or burnout over time. Employees who have experienced work accidents or received promotions in the last five years are less likely to leave, highlighting the importance of recognition and safety. Departmentally, HR employees are more prone to leaving, while those in management and R&D are less likely. Salary plays a critical role, with employees earning medium or high salaries significantly less likely to leave, especially those with high salaries. Overall, the model fits the data well and provides actionable insights for improving retention by addressing satisfaction, workload balance, and career advancement opportunities.



```
> summary(logisticModel)

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

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.3349691  0.2468811  -5.407 6.40e-08 ***
satisfaction_level       -4.1441061  0.1266195 -32.729  < 2e-16 ***
last_evaluation           0.6773739  0.1944896   3.483 0.000496 ***
number_project           -0.3125061  0.0276615 -11.298  < 2e-16 ***
average_montly_hours      0.0041680  0.0006719   6.203 5.53e-10 ***
time_spend_company        0.2677552  0.0204380  13.101  < 2e-16 ***
work_accidentyes         -1.5125210  0.1140187 -13.266  < 2e-16 ***
promotion_last_5yearsyes -0.9989453  0.3021599  -3.306 0.000946 ***
departmenthr              0.4132904  0.1668877   2.476 0.013269 *
departmentIT             -0.1146210  0.1553789  -0.738 0.460705
departmentmanagement     -0.6300693  0.2102127  -2.997 0.002724 **
departmentmarketing       0.0299699  0.1693205   0.177 0.859508
departmentproduct_mng    -0.2385088  0.1696238  -1.406 0.159693
departmentRandD          -0.4478186  0.1841710  -2.432 0.015035 *
departmentsales          -0.0932665  0.1304118  -0.715 0.474505
departmentsupport         0.1507652  0.1389524   1.085 0.277916
departmenttechnical       0.0835603  0.1352472   0.618 0.536685
salarylow                 1.8666710  0.1625894  11.481  < 2e-16 ***
salarymedium              1.3031035  0.1636251   7.964 1.67e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9821.5  on 8999  degrees of freedom
Residual deviance: 7647.1  on 8981  degrees of freedom
AIC: 7685.1

Number of Fisher Scoring iterations: 5
```

FIGURE 22 : LOGISTIC REGRESSION MODEL

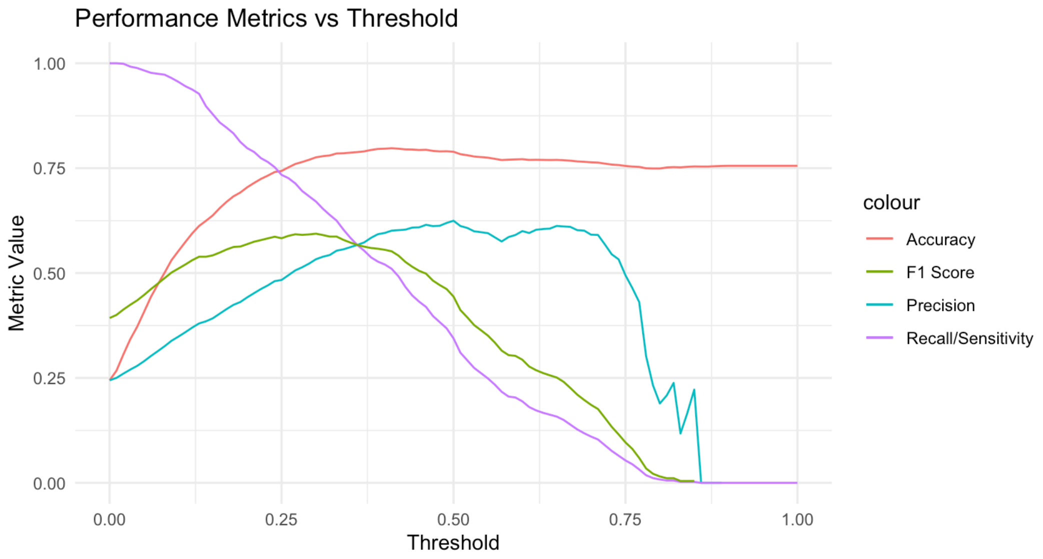
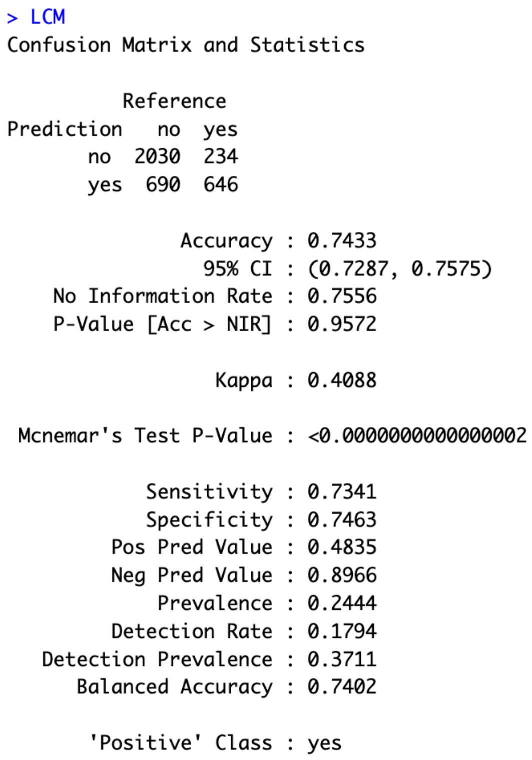


FIGURE 23 : PARAMETER TUNING

Based on the visual plot of parameter tuning, the threshold of 0.25 was chosen because it is the point where accuracy and sensitivity appear to be balanced. At this threshold, the sensitivity in the evaluation metrics represents the True Positive Rate, which is crucial for the model's performance in identifying employee turnover (class 1: yes). With an accuracy of 74.33%, the model demonstrates a solid overall performance. It performs particularly well in detecting actual turnover, with a sensitivity of 73.41%, correctly identifying 73.41% of employees who leave. This indicates that the model effectively recognizes true positive cases of employee turnover.



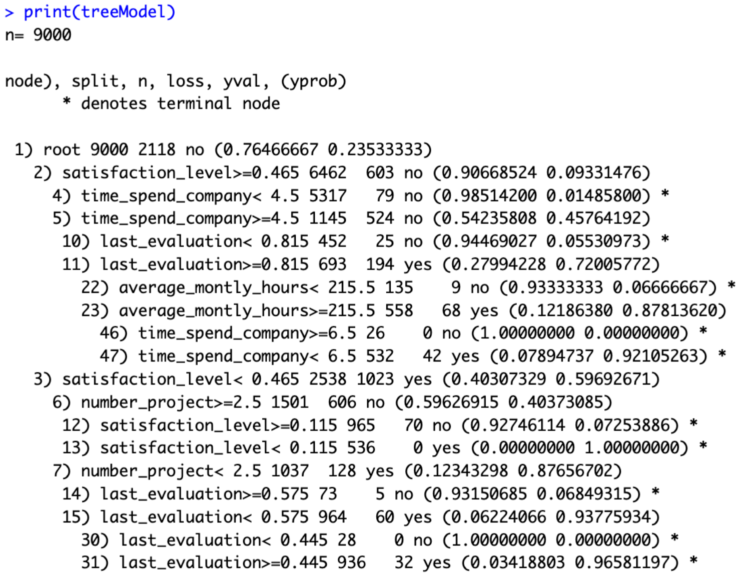
```
> LCM
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  2030  234
       yes  690  646

               Accuracy : 0.7433
                 95% CI : (0.7287, 0.7575)
    No Information Rate : 0.7556
    P-Value [Acc > NIR] : 0.9572

                  Kappa : 0.4088

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.7341
            Specificity : 0.7463
         Pos Pred Value : 0.4835
         Neg Pred Value : 0.8966
             Prevalence : 0.2444
         Detection Rate : 0.1794
   Detection Prevalence : 0.3711
      Balanced Accuracy : 0.7402

       'Positive' Class : yes
```

FIGURE 24 : LOGISTIC MODEL EVALUATION

Classification Tree

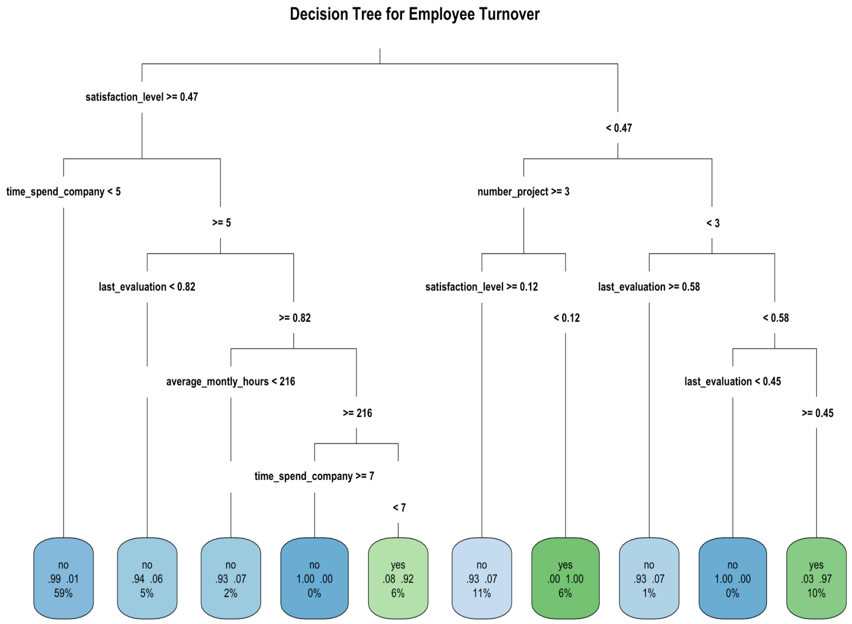
A classification tree recursively splits the dataset based on the most informative features, creating a tree structure where leaves represent distinct class labels. At each node, the algorithm selects the attribute that maximizes information gain or minimizes Gini impurity, resulting in a hierarchical model that efficiently classifies instances based on their feature values.

The classification tree has a root node with 9000 instances, split between two classes (76.46% no and 23.53% yes). It makes splits based on the satisfaction_level feature. For instances with 0.465 less than 0.5, it predicts class yes (59.69% accuracy). For instances with Freq greater than or equal to 0.465, it further splits based on time_spend_company, last_evaluation and average_monthly_hours producing terminal nodes with varying class predictions. The tree indicates a predictive pattern related to these features.

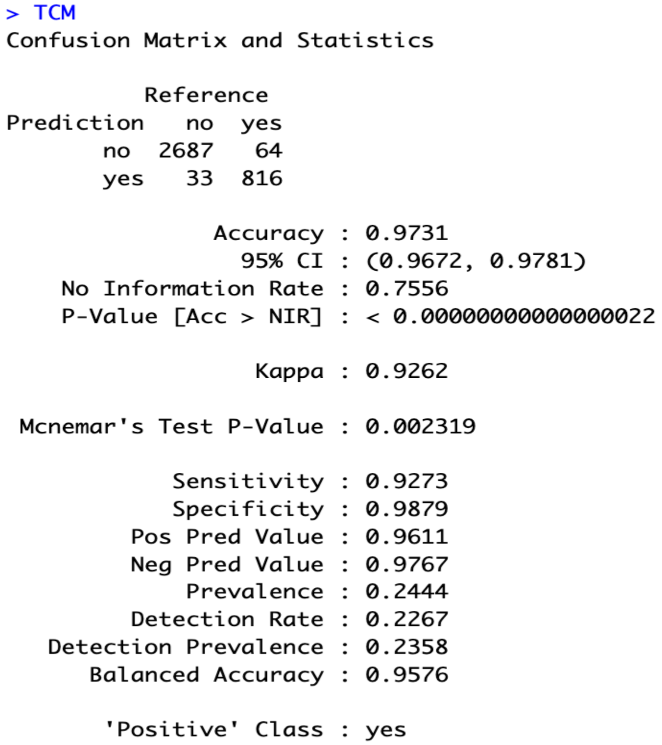


```
> print(treeModel)
n= 9000 

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 9000 2118 no (0.76466667 0.23533333)  
   2) satisfaction_level>=0.465 6462  603 no (0.90668524 0.09331476)  
     4) time_spend_company< 4.5 5317   79 no (0.98514200 0.01485800) *
     5) time_spend_company>=4.5 1145  524 no (0.54235808 0.45764192)  
      10) last_evaluation< 0.815 452   25 no (0.94469027 0.05530973) *
      11) last_evaluation>=0.815 693  194 yes (0.27994228 0.72005772)  
        22) average_montly_hours< 215.5 135    9 no (0.93333333 0.06666667) *
        23) average_montly_hours>=215.5 558   68 yes (0.12186380 0.87813620)  
          46) time_spend_company>=6.5 26    0 no (1.00000000 0.00000000) *
          47) time_spend_company< 6.5 532   42 yes (0.07894737 0.92105263) *
   3) satisfaction_level< 0.465 2538 1023 yes (0.40307329 0.59692671)  
     6) number_project>=2.5 1501  606 no (0.59626915 0.40373085)  
      12) satisfaction_level>=0.115 965   70 no (0.92746114 0.07253886) *
      13) satisfaction_level< 0.115 536    0 yes (0.00000000 1.00000000) *
     7) number_project< 2.5 1037  128 yes (0.12343298 0.87656702)  
      14) last_evaluation>=0.575 73    5 no (0.93150685 0.06849315) *
      15) last_evaluation< 0.575 964   60 yes (0.06224066 0.93775934)  
        30) last_evaluation< 0.445 28    0 no (1.00000000 0.00000000) *
        31) last_evaluation>=0.445 936   32 yes (0.03418803 0.96581197) *
```

FIGURE 25 : CLASSIFICATION TREE MODEL



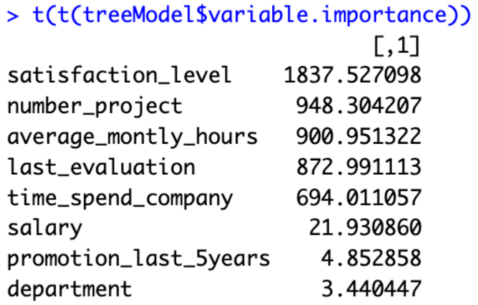
The classification tree model has achieved a sensitivity of 92.73%, which indicates that it correctly identifies employee turnover (i.e., the "yes" class) in most cases. Additionally, the model's overall accuracy is impressive at 97.31%, demonstrating that it performs very well in predicting employee retention and turnover. The model also shows strong specificity (98.79%), effectively identifying employees who do not leave the company.



```
> TCM
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  2687   64
       yes   33  816

               Accuracy : 0.9731
                 95% CI : (0.9672, 0.9781)
    No Information Rate : 0.7556
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9262

 Mcnemar's Test P-Value : 0.002319

            Sensitivity : 0.9273
            Specificity : 0.9879
         Pos Pred Value : 0.9611
         Neg Pred Value : 0.9767
             Prevalence : 0.2444
         Detection Rate : 0.2267
   Detection Prevalence : 0.2358
      Balanced Accuracy : 0.9576

       'Positive' Class : yes
```

FIGURE 26 : CLASSIFICATION TREE EVALUATION

The variable importance analysis for the classification tree model reveals the most significant predictors of employee turnover. Among the factors, satisfaction_level stands out as the most influential, with a variable importance score of 1837.53, followed by number_project (948.30) and average_monthly_hours (900.95), indicating that job satisfaction, the number of projects employees are involved in, and the hours worked monthly play a crucial role in predicting turnover. Other important variables include last_evaluation (872.99) and time_spend_company (694.01), both of which reflect the employee's engagement and tenure with the company. Salary (21.93) and promotion_last_5years (4.85) also have some predictive power, but to a lesser extent. Department (3.44) appears to be the least influential factor in determining turnover. These results provide valuable insights into which factors most strongly contribute to employee retention and departure.



```
> t(t(treeModel$variable.importance))
                             [,1]
satisfaction_level    1837.527098
number_project         948.304207
average_montly_hours   900.951322
last_evaluation        872.991113
time_spend_company     694.011057
salary                  21.930860
promotion_last_5years    4.852858
department               3.440447
```

FIGURE 27 : PREDICTOR IMPORTANCE FROM CLASSIFICATION TREE

Naïve Bayes model

Naïve Bayes is probabilistic learning built on the assumption all features are equally important and independent. Even when the assumptions are violated, naïve Bayes still performs accurately, particularly when features are large. This model is suitable for categorical and binary features. Built naïve Bayes model considering all predictors and left as the outcome variable to **classify an employee might turnover or not**.

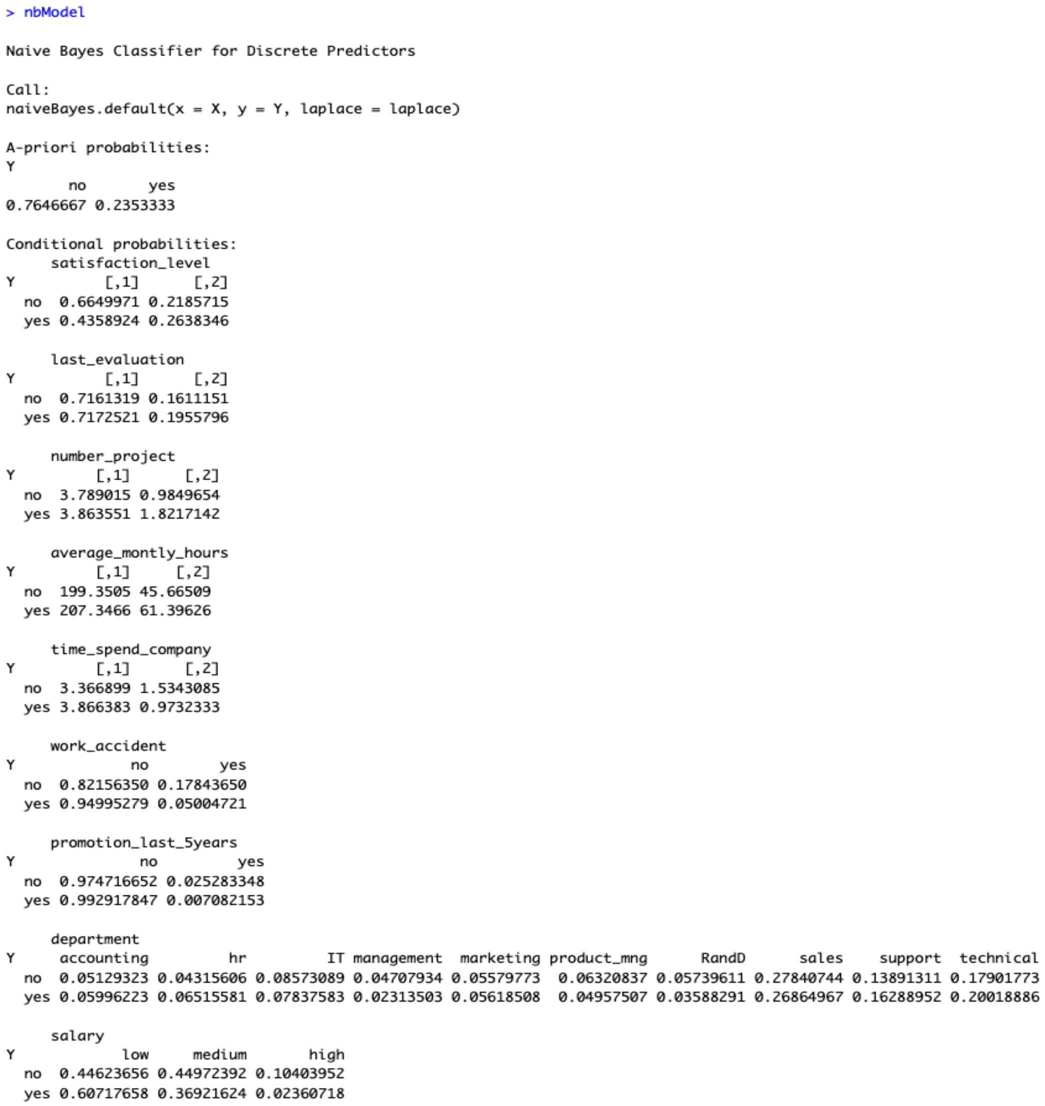
The Naive Bayes model provided insights into the relationships between various predictors and employee turnover (the dependent variable "Y"). The model's a-priori probabilities indicate that 76.47% of employees are classified as "no" (i.e., not leaving) and 23.53% as "yes" (i.e., leaving). This suggests that the company has a higher proportion of employees staying compared to those who leave.

The conditional probabilities show how each predictor influences the likelihood of turnover. For instance, **satisfaction_level** is higher for employees who stay ("no") with a probability of 0.6649 compared to 0.4359 for those leaving ("yes"). Similarly, **number_project** and **average_monthly_hours** are slightly higher for employees who stay, suggesting these factors are associated with lower turnover risk. **Time_spend_company** is also higher for employees who stay, indicating that longer tenure reduces the likelihood of leaving.

Factors like **work_accident** and **promotion_last_5years** show a significant difference in conditional probabilities between employees who stay versus leave. For example, employees who have had a work accident are more likely to stay, as the probability for staying is 0.9499 versus 0.8216 for those who have not had an accident. Similarly, employees who received a promotion in the last 5 years are more likely to stay (0.9929 for "no" vs. 0.9747).

Salary also plays a role, with employees on lower salaries having a higher probability of leaving. The **department** variable indicates that employees in "support" and "technical" departments have a higher probability of staying compared to others, like those in "sales" or "marketing."

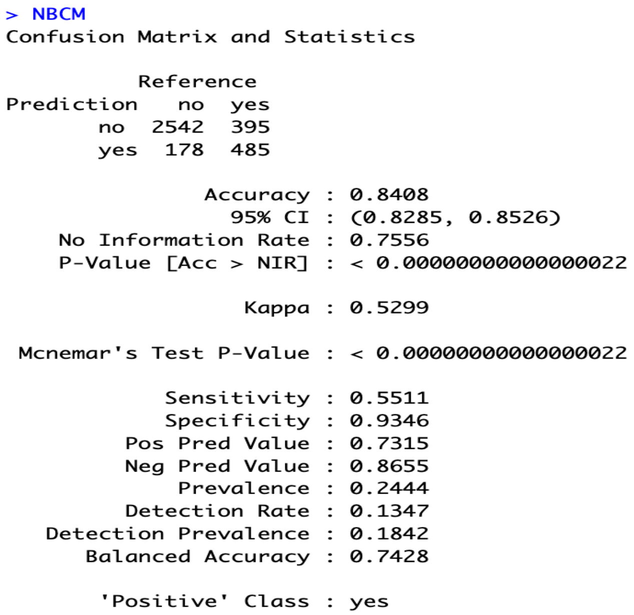
Overall, the Naive Bayes model highlights the influence of factors like job satisfaction, tenure, and work conditions (such as promotions and work accidents) on predicting employee turnover.



```
> nbModel

Naive Bayes Classifier for Discrete Predictors

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

A-priori probabilities:
Y
       no       yes 
0.7646667 0.2353333 

Conditional probabilities:
     satisfaction_level
Y          [,1]      [,2]
  no  0.6649971 0.2185715
  yes 0.4358924 0.2638346

     last_evaluation
Y          [,1]      [,2]
  no  0.7161319 0.1611151
  yes 0.7172521 0.1955796

     number_project
Y         [,1]      [,2]
  no  3.789015 0.9849654
  yes 3.863551 1.8217142

     average_montly_hours
Y         [,1]     [,2]
  no  199.3505 45.66509
  yes 207.3466 61.39626

     time_spend_company
Y         [,1]      [,2]
  no  3.366899 1.5343085
  yes 3.866383 0.9732333

     work_accident
Y             no        yes
  no  0.82156350 0.17843650
  yes 0.94995279 0.05004721

     promotion_last_5years
Y              no         yes
  no  0.974716652 0.025283348
  yes 0.992917847 0.007082153

     department
Y      accounting         hr         IT management  marketing product_mng      RandD      sales    support  technical
  no  0.05129323 0.04315606 0.08573089 0.04707934 0.05579773  0.06320837 0.05739611 0.27840744 0.13891311 0.17901773
  yes 0.05996223 0.06515581 0.07837583 0.02313503 0.05618508  0.04957507 0.03588291 0.26864967 0.16288952 0.20018886

     salary
Y            low     medium       high
  no  0.44623656 0.44972392 0.10403952
  yes 0.60717658 0.36921624 0.02360718
```

FIGURE 28 : NAÏVE BAYES MODEL

The Naive Bayes Classifier model has achieved a sensitivity of 55.11%, meaning it correctly identifies approximately 55% of employees likely to leave the company (i.e., the "yes" class). While this sensitivity is moderate, the model struggles somewhat with identifying turnover cases. The model's overall accuracy is 84.08%, which is quite strong and shows that the model performs well in distinguishing between employees who stay and those who leave. The specificity is 93.46%, demonstrating that the model effectively identifies employees not at risk of leaving. The positive predictive value (PPV) is 73.15%, indicating a good proportion of employees predicted to leave are actually leaving. Additionally, the model's Kappa value of 0.5299 suggests a moderate agreement between the predicted and actual values, while the balanced accuracy of 74.28% further confirms its overall decent performance in predicting employee turnover and retention.



```
> NBCM
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  2542  395
       yes  178  485

               Accuracy : 0.8408
                 95% CI : (0.8285, 0.8526)
    No Information Rate : 0.7556
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5299

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.5511
            Specificity : 0.9346
         Pos Pred Value : 0.7315
         Neg Pred Value : 0.8655
             Prevalence : 0.2444
         Detection Rate : 0.1347
   Detection Prevalence : 0.1842
      Balanced Accuracy : 0.7428

       'Positive' Class : yes
```

FIGURE 29 : NAÏVE BAYES MODEL EVALUATION

Performance Evaluation on Test Data

The **Classification Tree** model outperforms both the **Logistic Regression** and **Naïve Bayes** models in terms of both **accuracy** and **sensitivity**. With an impressive accuracy of **97.31%** and a sensitivity (True Positive Rate) of **92.73%**, the classification tree is highly effective at predicting employee turnover. In contrast, **Logistic Regression** achieves an accuracy of **74.33%** and sensitivity of **73.41%**, making it less reliable in identifying turnover. The **Naïve Bayes** model, while showing a decent accuracy of **84.08%**, has a much lower sensitivity of **55.11%**, indicating it misses a significant number of true positives. Thus, the classification tree is the most effective model for identifying employees likely to leave the company.

Model Name	Accuracy	Sensitivity (True Positive Rate)
Logistic Regression (threshold = 0.25)	74.33%	73.41%
Classification Tree	97.31%	92.73%
Naïve Bayes	84.08%	55.11%

FIGURE 30 : PERFORMANCE EVALUATION ON TEST DATA

Performance Evaluation of Model on Validation Data

The classification tree model was applied to the validation data, achieving an impressive **accuracy of 96.96%** and a **sensitivity of 90.23%**. The model demonstrated strong performance with high specificity (**99.07%**) and positive predictive value (**96.82%**), indicating its effectiveness in accurately identifying both employees likely to leave and those who would stay. The balanced accuracy of **94.65%** reflects the model's robust ability to handle both classes effectively, making it a reliable tool for predicting employee turnover.

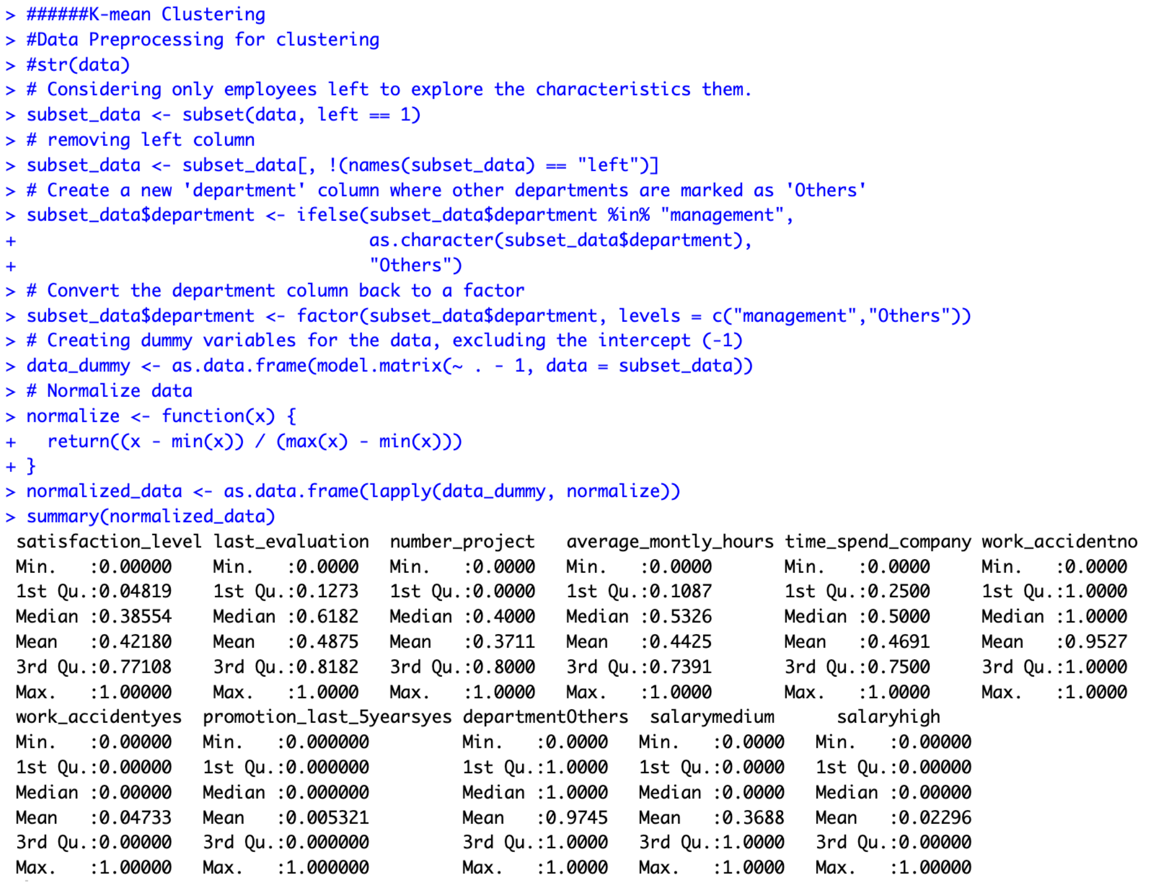
Model Name	Accuracy	Sensitivity (TPR)
Classification Tree	96.96%	90.23%

FIGURE 31 : PERFORMANCE EVALUATION ON VALIDATION DATA

K-Mean Clustering

Clustering approaches are for unsupervised learning used to cluster unlabeled cases into groups. K-Mean Clustering was appropriate to **achieve the business goal of finding important factors contributing to employee turnover**.

In the data preprocessing for clustering, first filtered the data to focus on employees who had left the company. The "left" column was removed from the dataset, and a new 'department' column was created, categorizing non-management departments as "Others." Dummy variables were generated for categorical variables, and the data was normalized using a custom function to scale the features between 0 and 1. The resulting normalized dataset was ready for clustering analysis, with each feature now on a comparable scale.



```
> ######K-mean Clustering
> #Data Preprocessing for clustering
> #str(data)
> # Considering only employees left to explore the characteristics them.
> subset_data <- subset(data, left == 1)
> # removing left column
> subset_data <- subset_data[, !(names(subset_data) == "left")]
> # Create a new 'department' column where other departments are marked as 'Others'
> subset_data$department <- ifelse(subset_data$department %in% "management",
+                                  as.character(subset_data$department),
+                                  "Others")
> # Convert the department column back to a factor
> subset_data$department <- factor(subset_data$department, levels = c("management","Others"))
> # Creating dummy variables for the data, excluding the intercept (-1)
> data_dummy <- as.data.frame(model.matrix(~ . - 1, data = subset_data))
> # Normalize data
> normalize <- function(x) {
+   return((x - min(x)) / (max(x) - min(x)))
+ }
> normalized_data <- as.data.frame(lapply(data_dummy, normalize))
> summary(normalized_data)
 satisfaction_level last_evaluation  number_project   average_montly_hours time_spend_company work_accidentno
 Min.   :0.00000    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000       Min.   :0.0000     Min.   :0.0000
 1st Qu.:0.04819    1st Qu.:0.1273   1st Qu.:0.0000   1st Qu.:0.1087       1st Qu.:0.2500     1st Qu.:1.0000
 Median :0.38554    Median :0.6182   Median :0.4000   Median :0.5326       Median :0.5000     Median :1.0000
 Mean   :0.42180    Mean   :0.4875   Mean   :0.3711   Mean   :0.4425       Mean   :0.4691     Mean   :0.9527
 3rd Qu.:0.77108    3rd Qu.:0.8182   3rd Qu.:0.8000   3rd Qu.:0.7391       3rd Qu.:0.7500     3rd Qu.:1.0000
 Max.   :1.00000    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000       Max.   :1.0000     Max.   :1.0000
 work_accidentyes  promotion_last_5yearsyes departmentOthers  salarymedium      salaryhigh
 Min.   :0.00000   Min.   :0.000000         Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.00000   1st Qu.:0.000000         1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :0.00000   Median :0.000000         Median :1.0000   Median :0.0000   Median :0.00000
 Mean   :0.04733   Mean   :0.005321         Mean   :0.9745   Mean   :0.3688   Mean   :0.02296
 3rd Qu.:0.00000   3rd Qu.:0.000000         3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.00000   Max.   :1.000000         Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
```

FIGURE 32 : K-MEAN CLUSTERING DATA PREPROCESSING

Silhouette measures the similarity of a data point to its cluster and relative to other clusters (cohesion vs separation).

The silhouette width increases with each increment of k up to k = 10, suggesting that k = 10 is the optimal clustering quality based on silhouette width. While this may offer the best separation, practical considerations, such as the clusters' interpretability and actionability, might lead you to choose a lower k. Considering k = 4 might balance the need for clustering performance with simplicity and actionable insights, making the results more useful for business strategy.

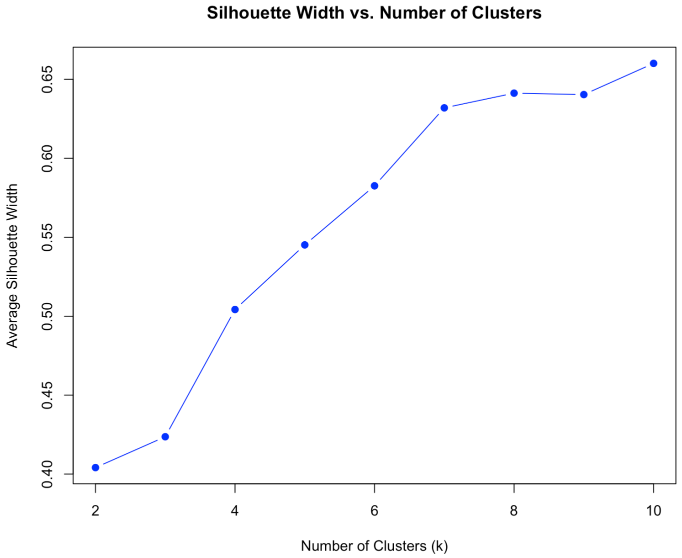


FIGURE 33 : KMEAN CLUSTERING TUNING K-MEAN PARAMETER

After preprocessing the data and finding the optimal k value, clustering was performed using the K-means algorithm with **k = 4**, resulting in 4 distinct clusters. These clusters group employees who left the company based on similar characteristics, such as satisfaction level, average monthly hours, time spent in the company, and other relevant features. The clustering helps identify patterns or segments within the employee base, providing insights into different groups of employees who might have experienced similar reasons for leaving.

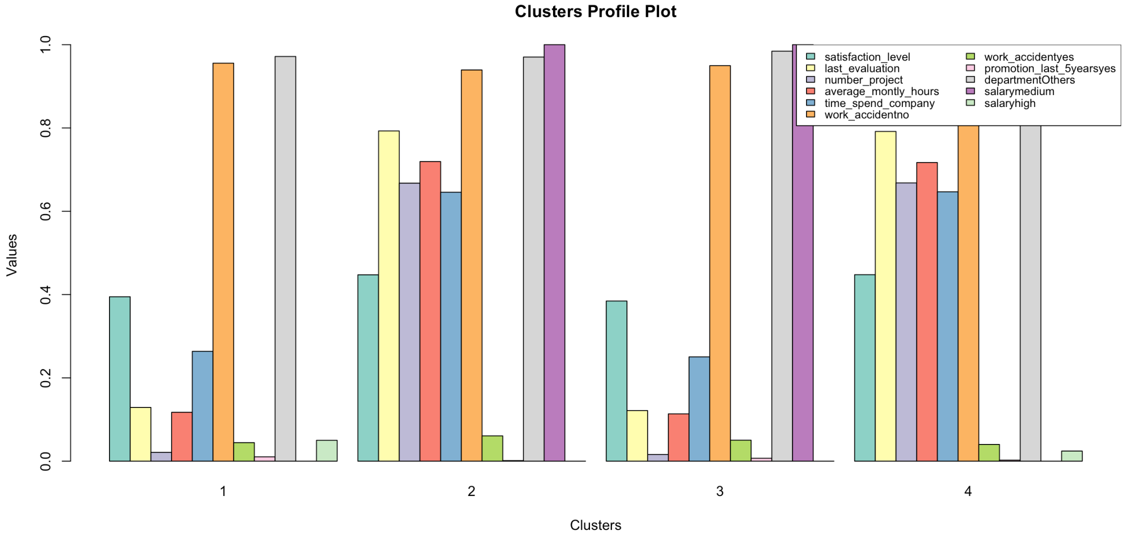


FIGURE 33 : PROFILE PLOT OF ALL 4 CLUSTERS

Based on the characteristics of each cluster, they were named that reflect the employee behavior and attributes in each group:

1. **Engaged but Unrewarded**: These employees are moderately satisfied, have decent performance evaluations, and work long hours, but have limited opportunities for promotions. They have medium salaries and tend to stay longer with the company.
2. **Disengaged and Underperforming**: This group is characterized by low satisfaction, poor performance evaluations, minimal project involvement, and very few work hours.

They also have a short tenure and are less likely to receive promotions or experience work accidents.

3. **High Potential but Unrecognized**: These employees exhibit similar satisfaction and evaluation levels to the first group, but they spend longer with the company and work more hours. However, they are rarely promoted, indicating untapped potential. They also have low work accident rates and are primarily in medium salary brackets.
4. **Frustrated and Stagnant**: Employees in this group have low satisfaction, poor evaluations, minimal involvement, and very low hours. Despite being more likely to experience promotions, they are likely frustrated and stagnant in their roles. They tend to have higher salaries but shorter tenure.

The silhouette plot summary provided insight into the quality of the clustering solution. It indicates that the clustering model has divided the 3,571 units (employees who left) into 4 clusters of varying sizes: 741, 1,058, 1,196, and 576 units. The average silhouette widths for the clusters range from 0.37 to 0.71, suggesting that some clusters are better defined than others. The individual silhouette widths, which reflect how well each unit fits its assigned cluster, have a median of 0.44 and a mean of 0.50, indicating a moderate level of cluster cohesion. The minimum silhouette value is negative (-0.14), suggesting that a few units may be poorly assigned to their respective clusters, but the majority of data points have positive silhouette widths, with values ranging up to 0.80, indicating that most clusters are reasonably well-separated and defined.

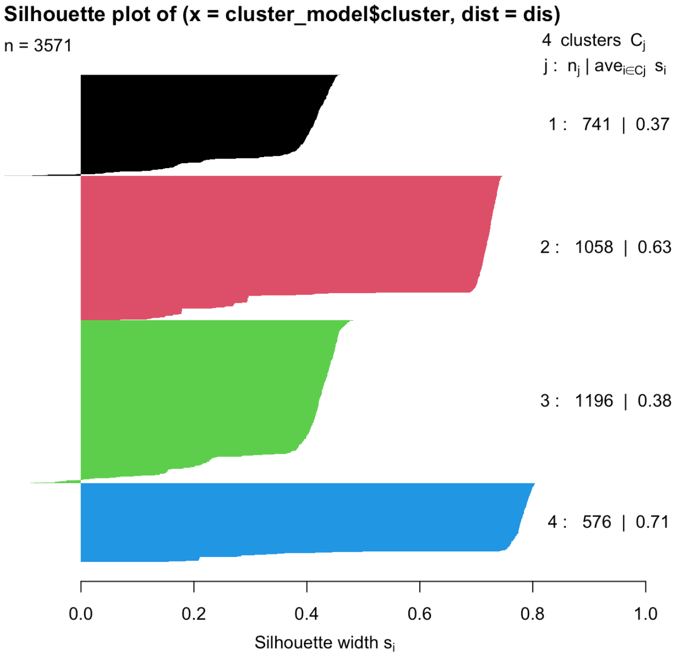


FIGURE 34 : SILHOUETTE PLOT FOR K =4

The clustering identified distinct employee groups with differing engagement levels, performance, and career progression. These clusters provided valuable insights into the underlying reasons for turnover, helping organizations better understand employee behavior and tailor retention strategies accordingly. By analyzing these segments, companies can focus on improving specific employee experiences, such as addressing dissatisfaction in high-risk groups, offering more recognition to high-potential employees, or improving support for disengaged workers. This will ultimately reduce turnover and enhance overall retention.

Conclusion

The classification tree has been performing the best model among other models to predict employee turnover, achieving an accuracy of 96.96% and sensitivity of 90.23%, effectively solving the first analytical goal. On the other hand, K-means clustering has been used to identify distinct employee segments based on characteristics like satisfaction level, time spent in the company, last evaluation, average monthly hours, and the number of projects. The clustering process has highlighted key factors contributing to employee turnover, helping to uncover patterns within different groups of employees. By analyzing these clusters, organizations can gain a deeper understanding of the underlying causes of turnover, allowing for more targeted and effective retention strategies to address the specific needs of each segment, ultimately improving employee satisfaction and reducing turnover.

Recommendations for the Business

The organization can potentially enhance its business strategies by adopting predictive analytics. The company should implement targeted actions on the employees classified from predictive models to reduce the employee turnover rate. Incorporate continuous model monitoring and adjusted predicted probabilities based on the regular audit data, ensuring long-term relevance and effectiveness.

Targeted action based on the important factors contributing to employee turnover is utilizing the classification tree model to classify existing employees as prone to turnover. Then, get the list of employees prone to turnover and conduct drives to increase employee satisfaction level, introduce schemes for employee’s milestones in the longer run at the organization, which might boost employees’ interest in staying with the company. Convey the employee’s target and goals clearly at the beginning of the year or quarter so employees can improve evaluation scores. And make sure employees are not overburdened with high working hours and the number of projects.