Human Capital Analysis

Karan Rajeshbhai Trivedi

M.Sc. Data Analytics

Webster University

CSDA 6010: Data Analytics Practicum

Prof. Dr. Ali Ovlia

Abstract

This project investigates the factors influencing employee turnover using a comprehensive dataset encompassing variables such as job satisfaction, promotion history, average monthly hours, and departmental assignments. Our analysis involved initial data exploration to identify patterns and relationships, followed by data transformation to prepare for modeling. Key visualizations, including job satisfaction distribution by turnover and promotion history by turnover, provided preliminary insights into the variables affecting employee retention.

To predict employee turnover, we used a variety of machine learning approaches, such as logistic regression, decision trees, k-nearest neighbors, and support vector machines (SVM). These models' performance was assessed using the following metrics: F1 score, recall, accuracy, and precision. The best model for predicting turnover turned out to be logistic regression, which identified work satisfaction and promotion history as important variables. The results provide useful information for human resource managers, including focused tactics to increase staff retention. This analysis lays the groundwork for future workforce management research and emphasizes the significance of data-driven strategies in tackling employee turnover.

Table of Contents

List of Figures

Executive Summary

In this project, we used a dataset including a variety of variables, including average monthly hours, job satisfaction, promotion history, and more, to perform a thorough analysis of employee turnover. We started our study by looking at the dimensions, summary statistics, and missing values of the dataset through data exploration and preprocessing. To comprehend the distribution and correlations of the important variables, we next converted and displayed them. The distribution of job satisfaction levels by turnover, the history of promotions by turnover, and the average monthly hours by turnover were among the visually striking displays. These plots provide preliminary insights into how these variables can affect workers' decisions to remain with or quit the company.

To determine the connections between the numerical variables, we also carried out a thorough correlation study, which led to the discovery of important trends that improve our comprehension of turnover. Logistic regression was used to assess the significance of each feature, highlighting important turnover predictors. Additionally, we standardized and one-hot encoded the data to get it ready for machine learning models. A number of predictive models,

such as logistic regression, decision trees, k-nearest neighbors, and support vector machines (SVM), were put into practice and compared. The models' performance was evaluated using F1 score, accuracy, precision, and recall. Based on the model's performance indicators, we ultimately chose a model to try and find the best predictor of employee turnover. Human resource professionals can use the practical findings from this project to create strategies that will lower employee turnover and increase retention.

1.0 Introduction

In today's competitive business landscape, having the most advanced and sophisticated machinery is no longer enough to rule a sector. Instead, companies that thrive to retain experienced and productive employees are proven to stay in competition. Thus, the significance of human capital management has increased because of this paradigm change, especially in terms of keeping top personnel. This project centers on a firm that is facing the difficult task of losing its most skilled and seasoned workers. This problem is not specific to this business; rather, it is a widespread worry in many different sectors. Long-term strategies will depend on keeping skilled workers around, since their exit can have a big effect on a company's growth trajectory and even put money at risk if they go to work for rival companies.

To address this, we will examine the Human Capital Analytics dataset (Employee.csv), which includes data on 14,999 employees across ten different attributes, to address this challenge. These characteristics include things like degree of satisfaction, results from the most recent evaluation, quantity of projects, average number of hours worked each month, duration of employment, work-related accidents, promotions, department, and pay. With this project, we hope to use data analytics techniques to better understand the factors that lead to employee turnover and create retention strategies.

2.0 Business Problems and Goals

2.1 Business Problems:

The organization has a hard time keeping knowledgeable and experienced workers. The productivity and financial stability of the company are at risk due to high employee turnover, particularly among top performers. Developing successful retention strategies and avoiding the loss of valuable talent to competitors require an understanding of the factors that lead to employee departures. Our goal will be to significantly reduce employee turnover, particularly among high-performing and experienced staff, thereby enhancing the company's overall performance, innovation capacity, and competitive position in the market. Also, to reduce annual employee turnover rate by 20% within the next fiscal year, with a focus on retaining high-performing employees who have been with the company for 3+ years.

2.2 Analytics Goal:

To test whether key factors can reliably predict employee turnover and validate our hypothesis that certain attributes significantly impact the likelihood of an employee leaving the company. The goal is to develop a predictive model for turnover, assess its accuracy, and use the findings to enhance employee retention strategies.

Objectives:

1. **Hypothesis Testing through EDA:** Conduct exploratory data analysis (EDA) to uncover patterns and validate our hypothesis regarding the relationship between employee attributes and turnover.
2. **Identify Key Predictors:** Determine which factors are most significant in predicting employee turnover based on the data.

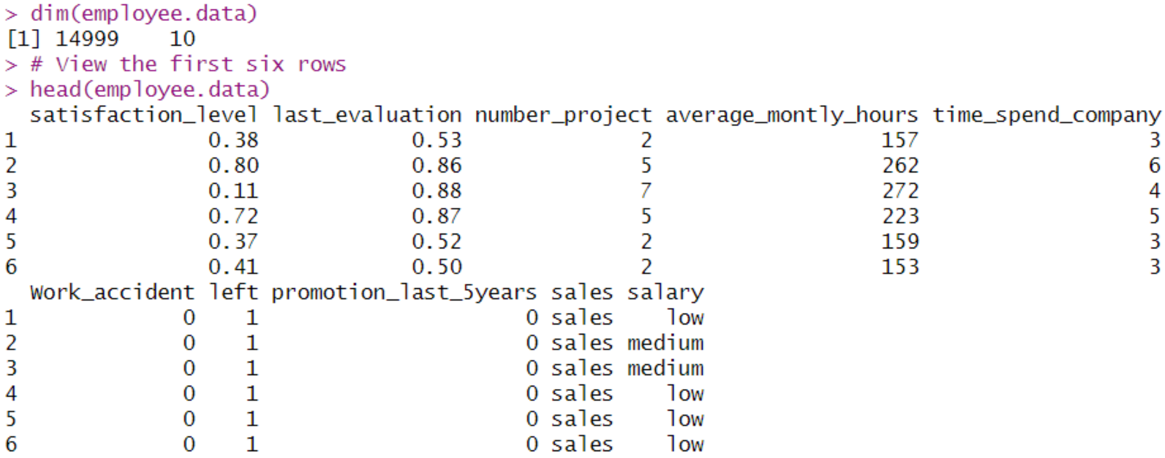
3.0 Data Exploration and Preprocessing

3.1 Attributes Definition:

Attribute	Description
satisfaction_level	A scale from 0 to 1 that indicates how satisfied an employee is with their work. Higher satisfaction is indicated by a higher value.
last_evaluation	The employee's score, which ranges from 0 to 1, represents their most recent performance evaluation. A higher rating denotes a better assessment.
number_project	The number of projects that the employee has worked on, with a range of 2 to 7.
average_monthly_hours	The average number of hours an employee works each month, ranging from 96 to 310 hours.
time_spend_company	The number of years an employee has worked for the company, ranging from 2 to 10 years.
work_accident	A binary variable indicating whether the employee has experienced a work accident (1 = Yes, 0 = No).
left	The target variable, indicating if the employee has left the organization (1 = Left, 0 = Stayed).
promotion_last_5years	A binary variable indicating whether the employee has received a promotion within the last 5 years (1 = Yes, 0 = No).
Sales	The department in which the employee is employed, represented by a categorical variable.
salary	A categorical variable classifying an employee's pay into three categories: low, medium, and high.

3.2 Data Exploration:

Exploring the data is essential to understand the dataset and getting it ready for additional examination. Loading the dataset and examining its basic properties is the first step. The read.csv function is used to load the dataset. We can gain an understanding of the dataset's structure and content by looking at its first few rows and dimensions. There are 10, columns (attributes) and 14,999 rows (records) in the dataset. If an employee has left the company, it is indicated by the target variable on the left (1 = Yes, 0 = No).



```
> dim(employee.data)
[1] 14999    10
> # View the first six rows
> head(employee.data)
  satisfaction_level last_evaluation number_project average_montly_hours time_spend_company
1               0.38            0.53              2                  157                  3
2               0.80            0.86              5                  262                  6
3               0.11            0.88              7                  272                  4
4               0.72            0.87              5                  223                  5
5               0.37            0.52              2                  159                  3
6               0.41            0.50              2                  153                  3
  Work_accident left promotion_last_5years sales salary
1             0    1                     0 sales    low
2             0    1                     0 sales medium
3             0    1                     0 sales medium
4             0    1                     0 sales    low
5             0    1                     0 sales    low
6             0    1                     0 sales    low
```

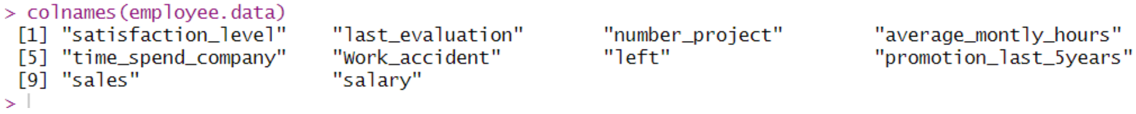
<Figure 3.2.1: First Six Rows and Dimension of the Data>

Column Names and Summary Statistics

The dataset includes the following attributes:

- satisfaction_level
- last_evaluation
- number_project
- average_montly_hours
- time_spend_company
- Work_accident
- left (target variable)

- promotion_last_5years
- sales (department)
- salary

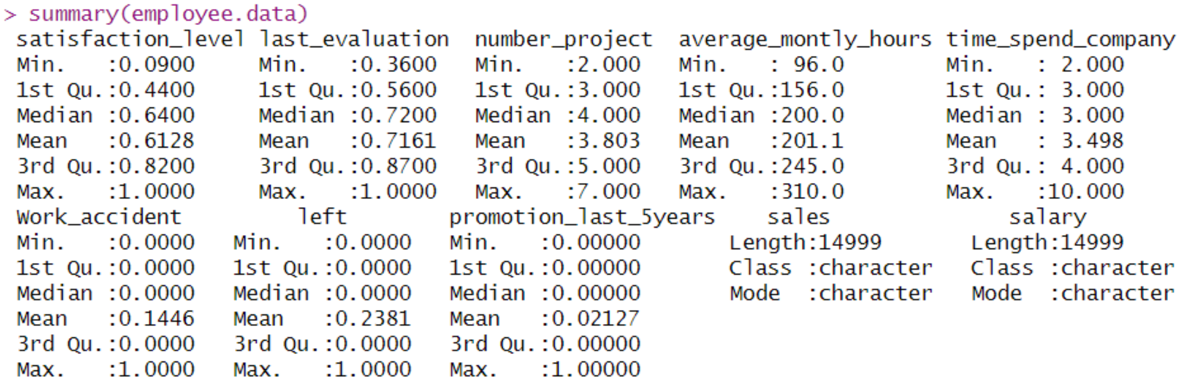


```
> colnames(employee.data)
 [1] "satisfaction_level"    "last_evaluation"       "number_project"        "average_montly_hours" 
 [5] "time_spend_company"    "Work_accident"         "left"                  "promotion_last_5years"
 [9] "sales"                 "salary"               
> 
```

<Figure 3.2.2: Column Names of the Dataset>

An overview of the features of the dataset is given by summary statistics:

- **Satisfaction Level**: Has a mean of 0.613 and a range of 0.09 to 1.00.
- **Last Evaluation**: Has a mean score of 0.716 and a range of scores from 0.36 to 1.00.
- **Number of Projects**: Has an average of 3.80 and ranges from 2 to 7.
- **Average Monthly Hours**: 201.1 on average, with a range of 96 to 310 hours.
- **Time Spent at Company**: Has a mean of 3.50 years and varies from 2 to 10 years.
- **Work Accident**: 14.46% of employees had a work accident.
- **Promotion in Last 5 Years**: Only 2.13% received a promotion.
- **Salary**: There are three categories: Low, Medium, and High.



```
> summary(employee.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    sales              salary         
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000       Length:14999       Length:14999      
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000       Class :character   Class :character  
 Median :0.0000   Median :0.0000   Median :0.00000       Mode  :character   Mode  :character  
 Mean   :0.1446   Mean   :0.2381   Mean   :0.02127                                            
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000                                            
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000                                            
```

<Figure 3.2.3: Summary Statistics of the Dataset>

Data Types and Missing Values

To ensure that the data types for the variables are appropriate, they are examined: Both character and numeric variables are present in this dataset. Sales and salary are examples of categorical variables. Let's talk about every kind of variable:

Numeric Data:

- **satisfaction_level**: Used to measure employee satisfaction levels, this continuous numeric variable ranges from 0.09 to 1.00.
- **last_evaluation**: this variable shows the outcomes of the most recent performance evaluation and ranges from 0.36 to 1.00.
- **number_project**: An employee's involvement in projects is counted using this integer variable, which ranges from 2 to 7.
- **average_monthly_hours**: This integer variable, which ranges from 96 to 310 hours, indicates the typical number of hours worked by staff members each month.
- **time_spend_company**: This integer variable counts the years that staff members have worked for the company, ranging from 2 to 10.

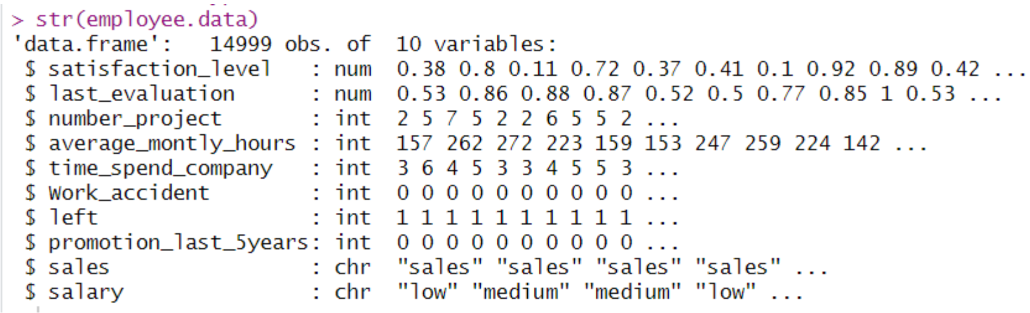
Binary Integer Data:

- **Work_accident**: This variable can be set to 1 (accident occurred) or 0 (no accident).
- **left**: Indicates if a worker has left the company (1 being out, 0 being employed).
- **promotion_last_5years**: Indicates if a worker has been promoted within the previous five years; a score of 0 means no promotion, while a score of 1 indicates a promotion.

Categorical Data:

- **sales**: The employee's department is indicated by a character string variable. Every entry in this dataset has the label "sales”,”IT”,”accounting” etc.

- **salary**: An additional character string variable that divides employee pay into three categories: "low," "medium," and "high."

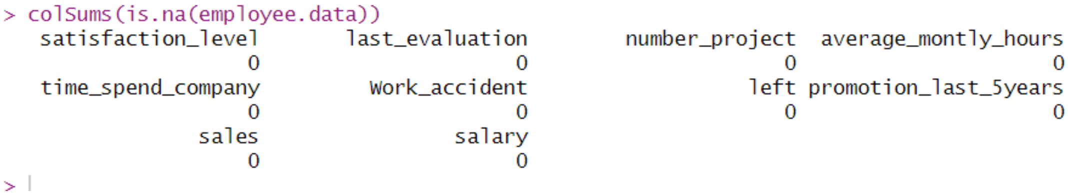


```
> str(employee.data)
'data.frame':	14999 obs. of  10 variables:
 $ satisfaction_level   : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
 $ last_evaluation      : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
 $ number_project       : int  2 5 7 5 2 2 6 5 5 2 ...
 $ average_montly_hours : 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 3.2.4: Data Types of the Variables>

Checking for missing values:

The dataset is complete and contains all the information that is needed because there are no missing values in any of the columns.



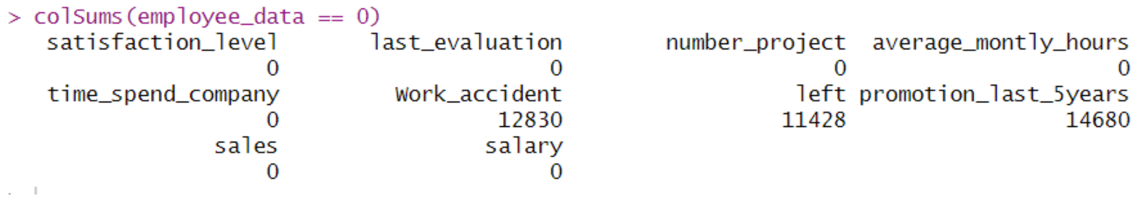
```
> colSums(is.na(employee.data))
   satisfaction_level       last_evaluation        number_project  average_montly_hours 
                    0                     0                     0                     0 
   time_spend_company         Work_accident                  left promotion_last_5years 
                    0                     0                     0                     0 
                sales                salary 
                    0                     0 
> 
```

<Figure 3.2.5: Checking for Missing Values>

Analysis of Zeros

Examining the zeros in particular columns shed light on particular data points:

- **Work Accident**: 85.54% of workers report no history of workplace accidents.
- **Promotion Last 5 Years**: Out of all the workforce, 97.87% did not receive a promotion in the previous five years.
- **left**: This variable contains 11,428 zeros, meaning that while the remaining employees have left, a sizable portion of the workforce—roughly 76.02 percent—remains with the company.



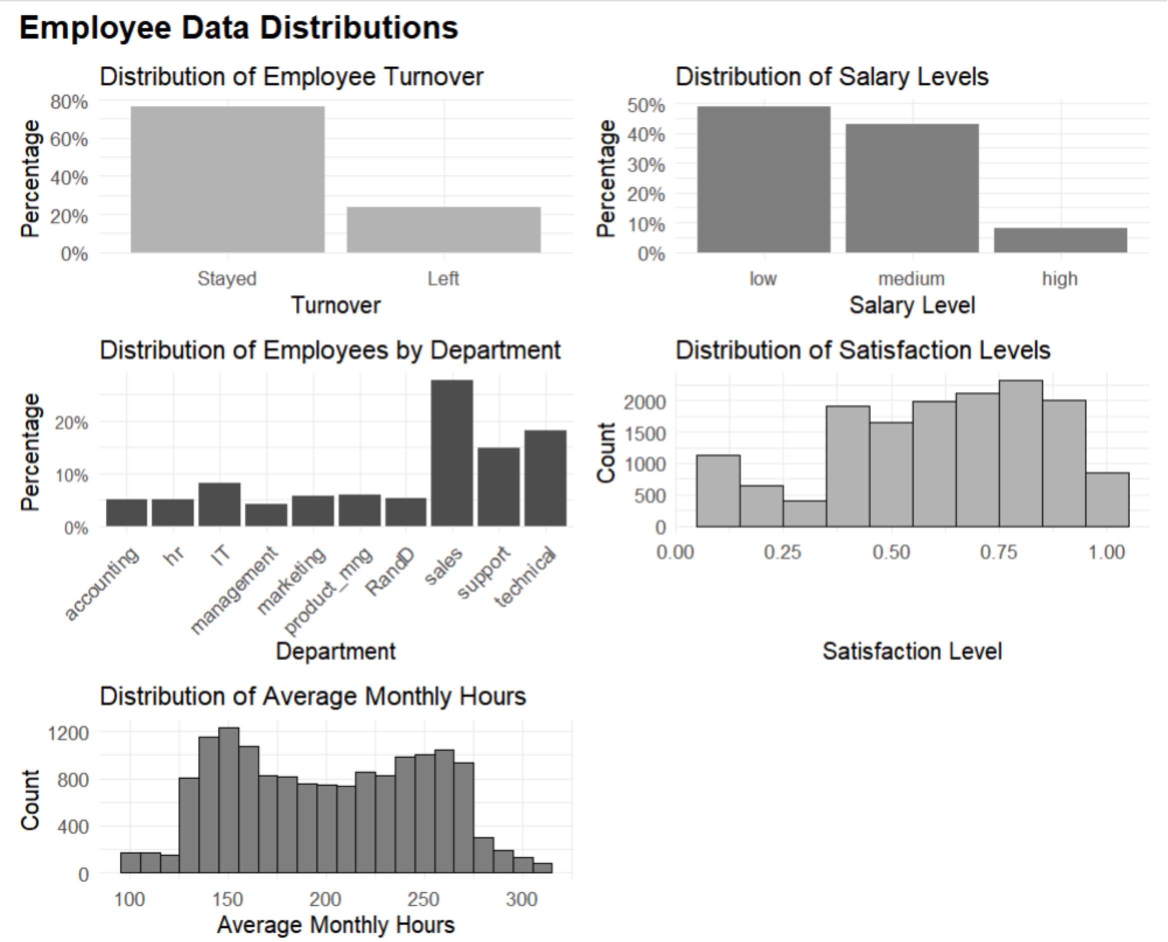
```
> colSums(employee_data == 0)
   satisfaction_level       last_evaluation        number_project  average_montly_hours 
                    0                     0                     0                     0 
   time_spend_company         Work_accident                  left promotion_last_5years 
                    0                 12830                 11428                 14680 
                sales                salary 
                    0                     0 
```

<Figure 3.2.6: Check for Zeros>

3.3 Distribution of Variables

We'll examine the distribution of important variables to gain insight into patterns and trends. To start with we will plot multiple visualizations to have an idea of what exactly is the root cause or does this factor matters in our project. The Figure 3.3.1 shows multiple plots that provide insights into employee data:

- **Employee Turnover**: More employees stayed than left the company.
- **Salary Levels**: Most employees have medium salaries, followed by low, then high salaries.



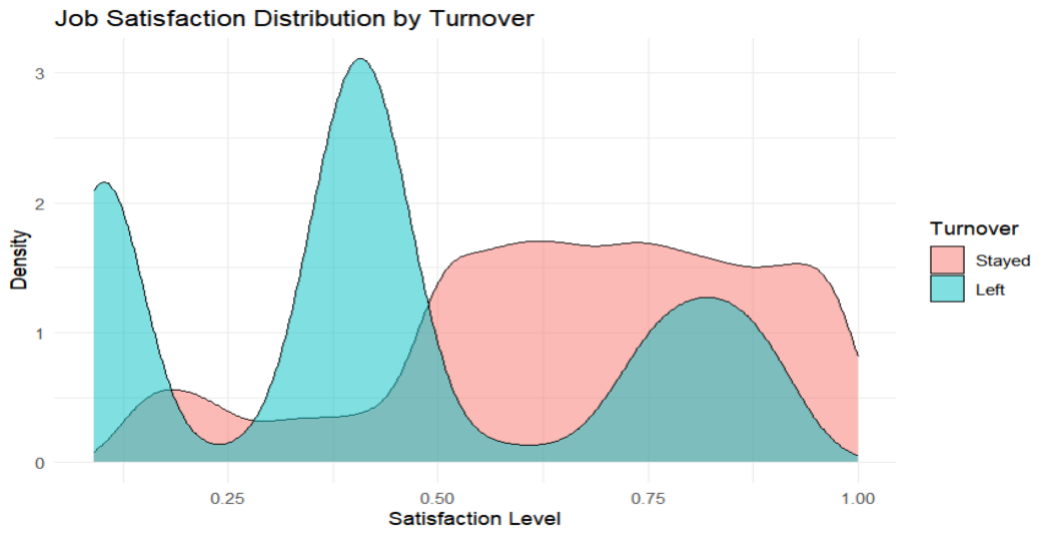
<Figure 3.3.1: Distribution of Employee Data>

- **Department Distribution**: The technical department has the highest percentage of employees, while accounting has the lowest.
- **Satisfaction Levels**: There's a wide range of satisfaction levels, with peaks at very low and high satisfaction.
- **Average Monthly Hours**: Most employees work between 150-250 hours per month, with the highest number working around 200 hours.

These plots together give a comprehensive overview of the workforce, showing patterns in turnover, salaries, departmental distribution, job satisfaction, and working hours

Job Satisfaction Distribution by Turnover:

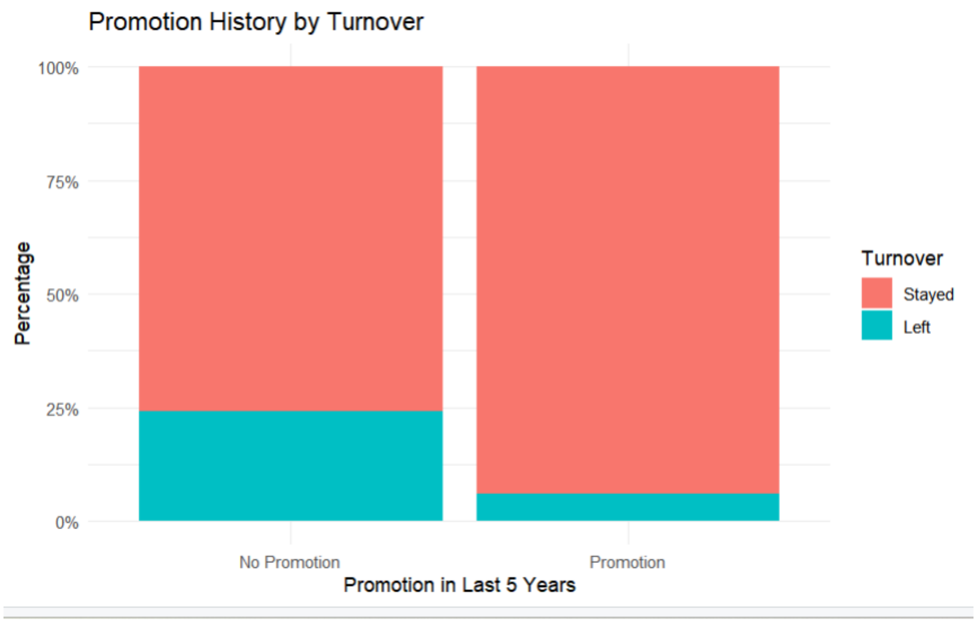
The graph shows how satisfied employees are with their jobs, comparing those who stayed versus those who left. Interestingly, employees who stayed were either very satisfied or very unsatisfied with their jobs. On the other hand, employees who left were more likely to have medium levels of satisfaction. This suggests that just making employees somewhat satisfied isn't enough to keep them - there are likely other important factors influencing whether people stay or leave their jobs.



<Figure 3.3.2: Job Satisfaction distribution by Turnover>

Promotion History by Turnover:

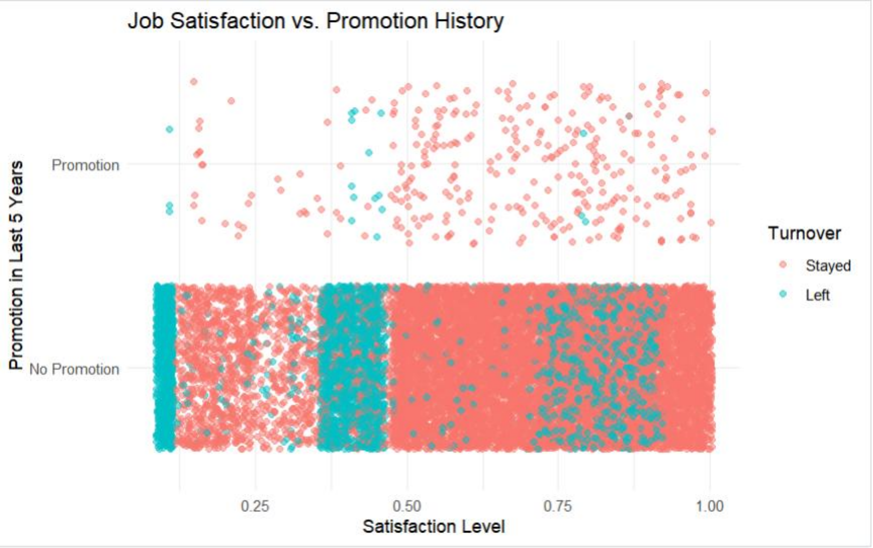
This graph helps us understand how promotions relate to employee turnover. Most employees, whether they stayed or left, didn't get promoted in the last 5 years. Interestingly, among those who did get promoted, a higher percentage stayed with the company. This suggests that promotions might help retain employees, but they're not the only factor since many employees stay even without promotions.



<Figure 3.3.3: Promotion History by Turnover>

Job Satisfaction vs Promotion History:

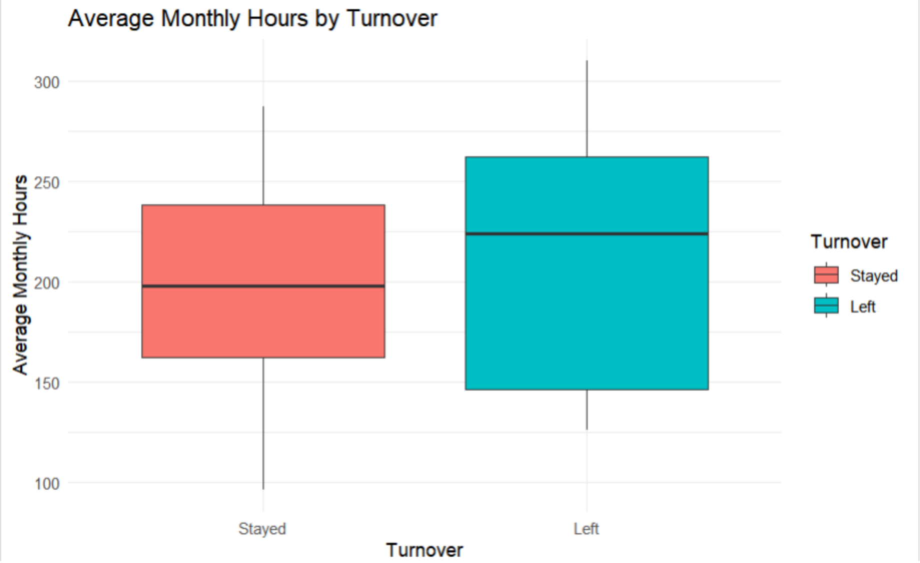
Let's relate job satisfaction and promotions, here most employees, whether they stayed or left, didn't get promoted in the last 5 years (shown by the dense cluster at the bottom). Promotions are rare across all satisfaction levels. Interestingly, there's no clear pattern linking job satisfaction to promotions, suggesting that other factors might influence who gets promoted.



<Figure 3.3.4: Jobs Satisfaction vs Promotion History>

Average Monthly Hours by Turnover:

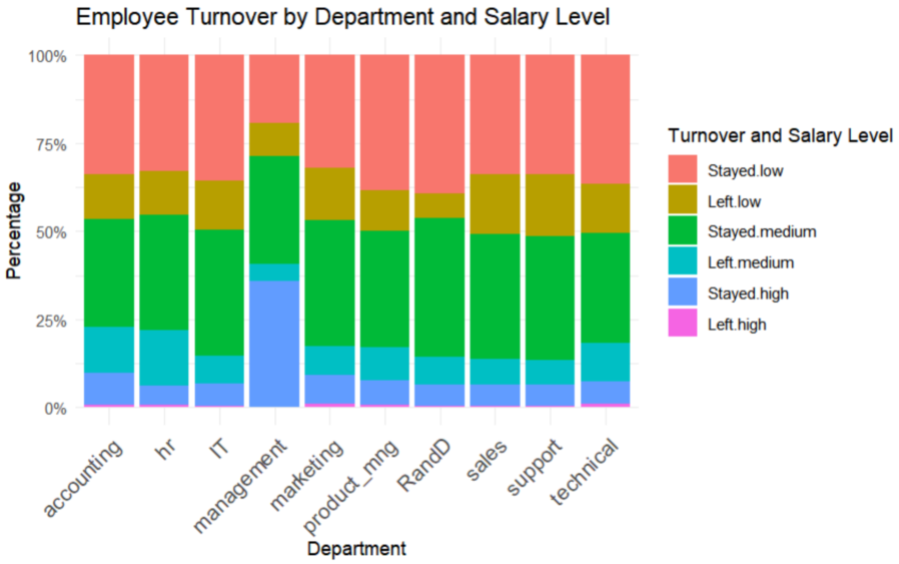
The plot shows the average monthly hours worked by employees who stayed versus those who left the company. Employees who left worked more hours on average, as shown by the taller blue box. There's more variation in hours for those who left, indicated by the larger box size. This suggests that working longer hours might be related to higher turnover rates.



<Figure 3.3.5: Average Monthly Hours by Turnover>

Employee Turnover by Department & Salary Level:

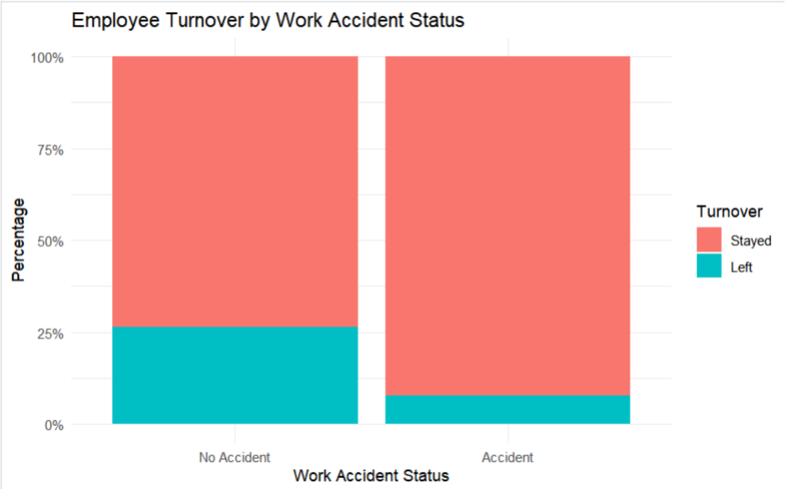
This plot shows employee turnover rates across different departments, broken down by salary levels. Each department is represented by a stacked bar, with colours indicating different combinations of turnover (stayed or left) and salary levels (low, medium, high). The proportions of these categories vary between departments, suggesting that turnover patterns and salary distributions differ across different areas of the company. This visualization helps identify which departments might have higher retention or turnover issues at specific salary levels.



<Figure 3.3.6: Employee Turnover by Department & Salary Level>

Employee Turnover by Work Accident:

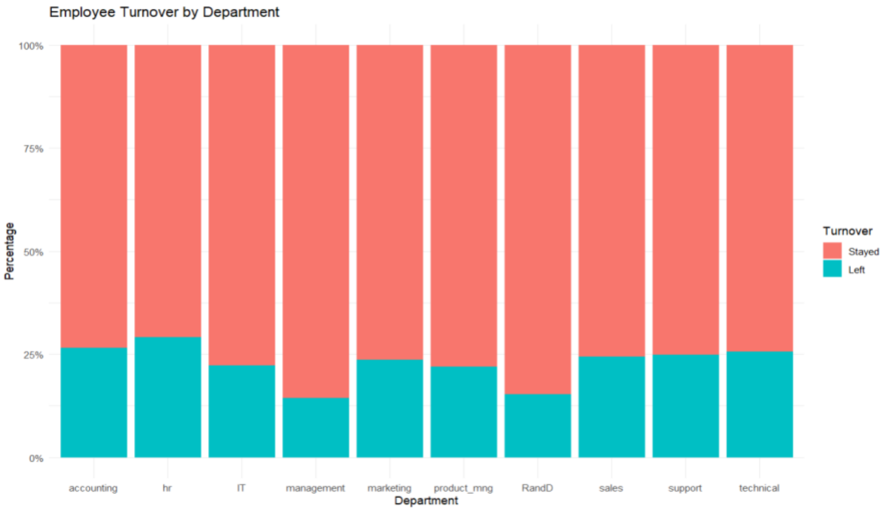
The plot shows employee turnover based on whether they had a work accident or not. Most employees, regardless of accident history, stayed with the company (shown by the large pink bars). However, there's a slightly higher percentage of employees who left among those who didn't have accidents compared to those who did. This suggests that work accidents might have a small impact on employee retention, but it's not a major factor in turnover.



<Figure 3.3.7: Employee Turnover by Work Accident Status>

Employee Turnover by Department:

As we see in the chart below, the proportion of employees who stayed versus those who left for each department. Each bar represents a department, with the red portion showing the percentage who left and the blue portion showing those who stayed. This allows for a quick comparison of turnover rates across different departments, helping to identify which areas might need attention in terms of employee retention.



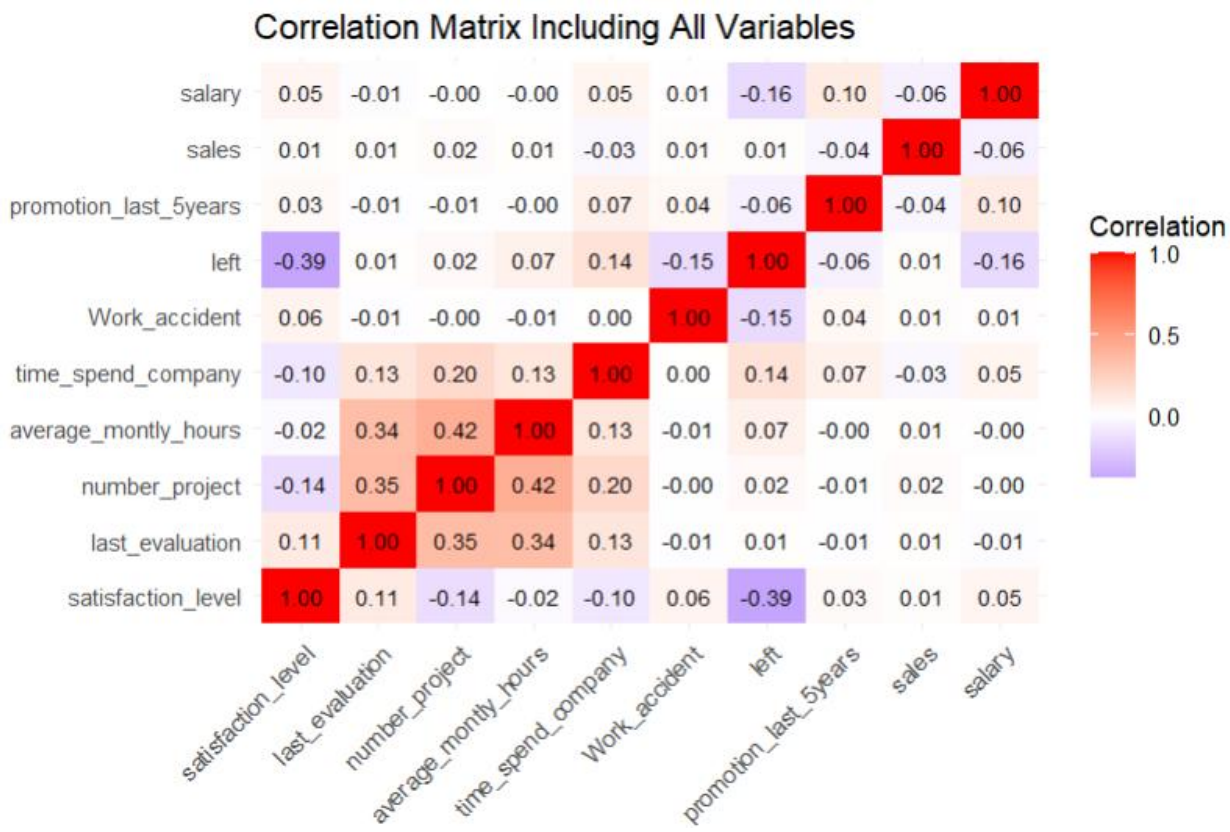
<Figure 3.3.8: Employee Turnover by Department>

3.4 Correlation Analysis:

The correlation matrix provides insights into key relationships in the employee dataset:

- **Satisfaction Level** is strongly negatively correlated with turnover (left) (-0.39), indicating that less satisfied employees are more likely to leave. It has a minor positive correlation with performance evaluations (0.11).
- **Last Evaluation** positively correlates with both the number of projects (0.35) and average monthly hours (0.34), suggesting that higher workloads and longer hours are linked to better evaluations.

- **Number of Projects** has a strong positive correlation with average monthly hours (0.42) and last evaluation (0.35), implying more projects lead to longer hours and better reviews.
- **Average Monthly Hours** is strongly correlated with the number of projects (0.42) and last evaluation (0.34), indicating longer hours are associated with more projects and higher performance ratings.
- **Time Spent at the Company** shows weak positive correlations with turnover (0.14) and performance evaluations (0.13), suggesting longer tenure slightly increases the likelihood of leaving and marginally improves evaluations.
- **Work Accident** has a weak negative correlation with turnover (-0.15), suggesting that employees with accidents are slightly less likely to leave.
- **Turnover (left)** is strongly negatively correlated with satisfaction (-0.39) and weakly with salary (-0.16), indicating low satisfaction and lower salaries are associated with higher turnover.
- **Promotion in Last 5 Years** shows a weak positive correlation with salary (0.10), implying that recent promotions slightly increase salary levels.
- **Sales** does not show strong correlations with other variables, indicating its minimal impact on other factors.
- **Salary** has weak correlations with both turnover (-0.16) and promotions (0.10), suggesting a minor effect on turnover and salary adjustments.



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
salary	0.05	-0.01	-0.00	-0.00	0.05	0.01	-0.16	0.10	-0.06	1.00
sales	0.01	0.01	0.02	0.01	-0.03	0.01	0.01	-0.04	1.00	-0.06
promotion_last_5years	0.03	-0.01	-0.01	-0.00	0.07	0.04	-0.06	1.00	-0.04	0.10
left	-0.39	0.01	0.02	0.07	0.14	-0.15	1.00	-0.06	0.01	-0.16
Work_accident	0.06	-0.01	-0.00	-0.01	0.00	1.00	-0.15	0.04	0.01	0.01
time_spend_company	-0.10	0.13	0.20	0.13	1.00	0.00	0.14	0.07	-0.03	0.05
average_montly_hours	-0.02	0.34	0.42	1.00	0.13	-0.01	0.07	-0.00	0.01	-0.00
number_project	-0.14	0.35	1.00	0.42	0.20	-0.00	0.02	-0.01	0.02	-0.00
last_evaluation	0.11	1.00	0.35	0.34	0.13	-0.01	0.01	-0.01	0.01	-0.01
satisfaction_level	1.00	0.11	-0.14	-0.02	-0.10	0.06	-0.39	0.03	0.01	0.05

<Figure 3.4.1: Correlation Matrix>

Key Insights

- **Satisfaction** is a major factor in reducing turnover, Strong negative correlation (-0.39) between satisfaction and turnover; logistic regression identifies it as the most important predictor (coefficient -4.05).
- **Workload** affects performance evaluations significantly; Projects and monthly hours strongly correlate with evaluation scores (0.35 and 0.34); decision tree uses these as key predictors.
- **Salary and Promotions** have a smaller impact on turnover, Logistic regression shows lower coefficient for salary (1.93) than satisfaction; most employees didn't receive promotions, regardless of staying or leaving.

- **Long-Tenured Employees** may require specific retention efforts, Weak positive correlation (0.14) between tenure and turnover; decision tree uses time at company as a criterion, indicating need for targeted retention strategies.

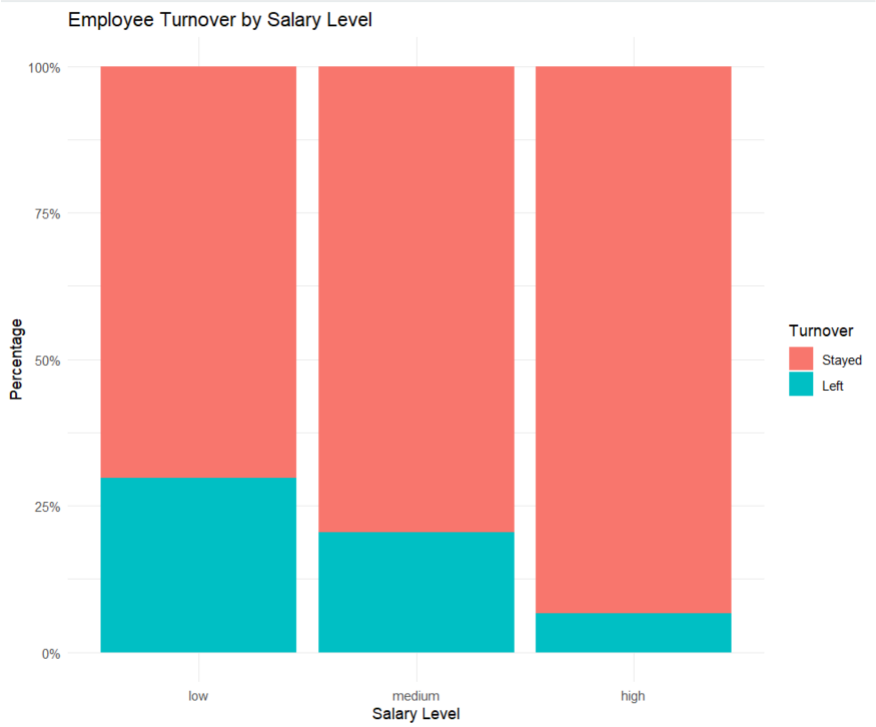
These insights highlight the interplay between employee satisfaction, performance, and retention, offering guidance for HR strategies to enhance organizational outcomes.

4.0 Hypothesis Testing:

4.1.1 Hypothesis 1: Salary is the reason why employees left the company.

Null Hypothesis (H0): Salary has no effect on employee turnover; in other words, salary levels do not significantly influence whether employees stay with or leave the company.

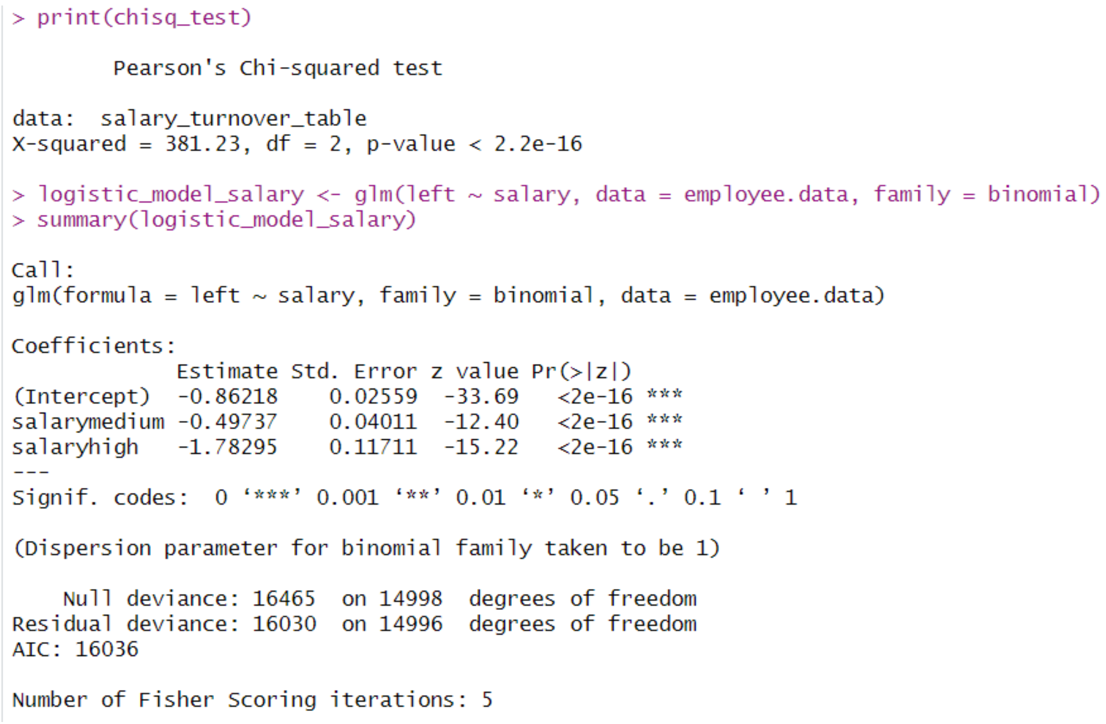
Looking at our initial findings in data exploration, we can say that salary is one of the factors of employee leaving organisation, especially on low salary bar, please refer below figure.



<Figure 4.1.1.: Employee Turnover by Salary>

The Pearson's Chi-squared test confirms a significant relationship between salary levels and employee turnover, with a p-value < 2.2e-16. This strong statistical evidence suggests that salary indeed impacts whether employees leave the company.

The logistic regression analysis further substantiates this finding. Specifically, employees with medium and high salaries are less likely to leave compared to those with low salaries, as indicated by the negative coefficients for medium and high salary levels. The coefficients are -0.497 for medium and -1.783 for high salaries, both highly significant (p-value < 2e-16), suggesting a substantial reduction in the likelihood of turnover with increasing salary. This indicates that salary plays a crucial role in employee retention, with higher salary levels being associated with lower turnover rates. Therefore, we reject the null hypothesis and conclude that salary is a significant factor influencing employee turnover.



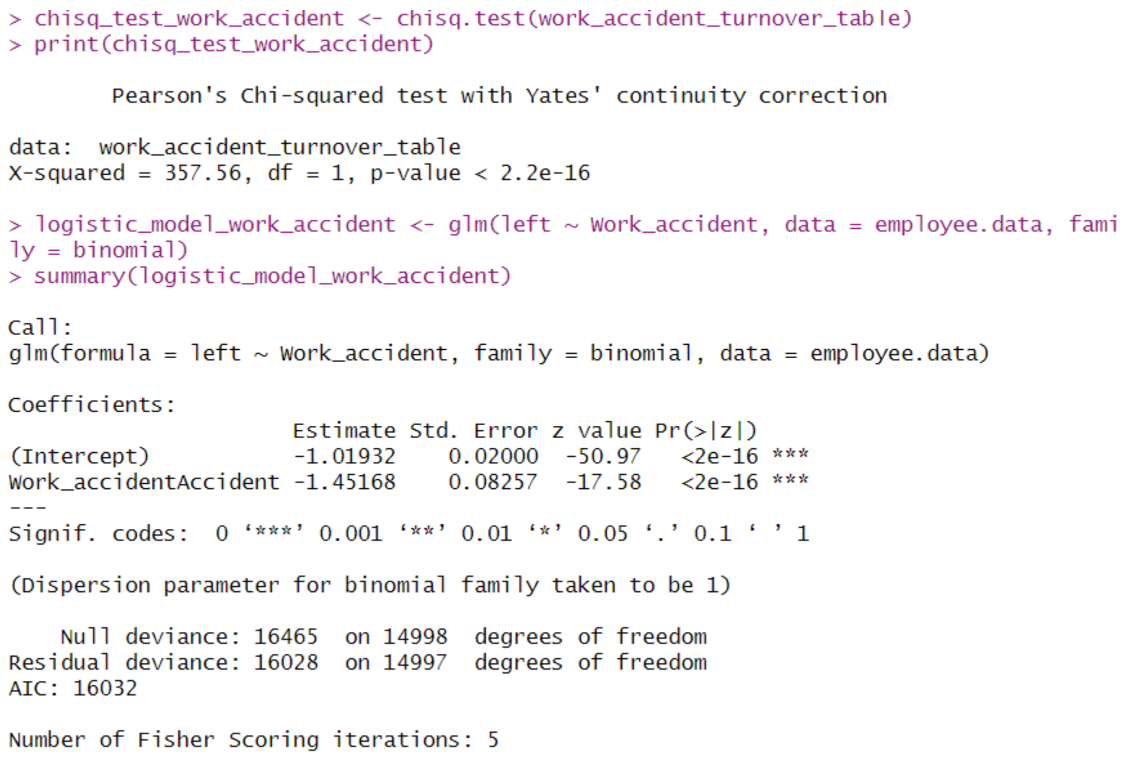
```
> print(chisq_test)

	Pearson's Chi-squared test

data:  salary_turnover_table
X-squared = 381.23, df = 2, p-value < 2.2e-16

> logistic_model_salary <- glm(left ~ salary, data = employee.data, family = binomial)
> summary(logistic_model_salary)

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -0.86218    0.02559  -33.69   <2e-16 ***
salarymedium -0.49737    0.04011  -12.40   <2e-16 ***
salaryhigh   -1.78295    0.11711  -15.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16465  on 14998  degrees of freedom
Residual deviance: 16030  on 14996  degrees of freedom
AIC: 16036

Number of Fisher Scoring iterations: 5
```

<Figure 4.1.2: Hypothesis 1: Chi-square test & Logistic Model>

4.1.2 Hypothesis 2: Employees leave the company because work is not safe.

Null Hypothesis (H0): Work accidents do not affect employee turnover; in other words, the occurrence of work accidents does not significantly influence whether employees stay with or leave the company.

Looking at the figure 3.3.7, we can observe the bar (green area) which says left with No accident, which is enough to conclude that there are less people who left the organisation with accident as compared to No accident.



```
> chisq_test_work_accident <- chisq.test(work_accident_turnover_table)
> print(chisq_test_work_accident)

        Pearson's Chi-squared test with Yates' continuity correction

data:  work_accident_turnover_table
X-squared = 357.56, df = 1, p-value < 2.2e-16

> logistic_model_work_accident <- glm(left ~ Work_accident, data = employee.data, fami
ly = binomial)
> summary(logistic_model_work_accident)

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

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -1.01932    0.02000  -50.97   <2e-16 ***
Work_accidentAccident -1.45168    0.08257  -17.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16465  on 14998  degrees of freedom
Residual deviance: 16028  on 14997  degrees of freedom
AIC: 16032

Number of Fisher Scoring iterations: 5
```

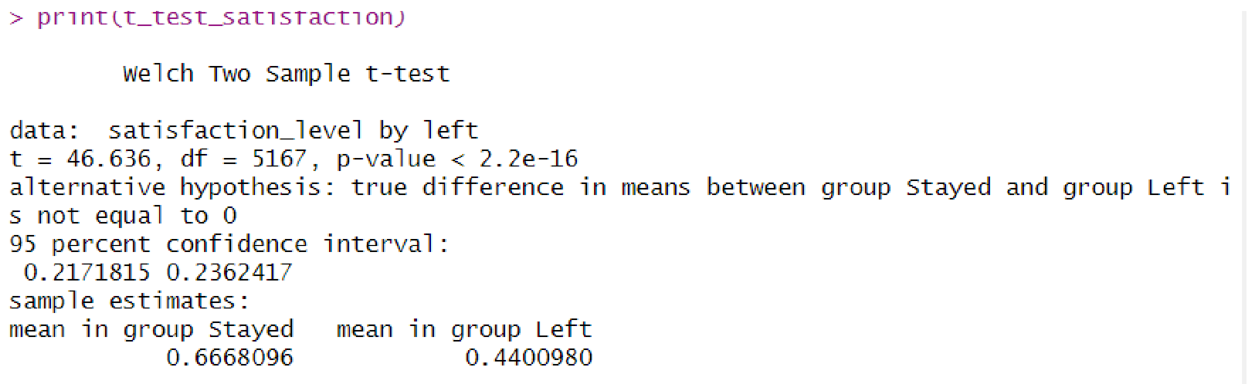
<Figure 4.1.3: Hypothesis 2: Chi-square test & Logistic Model>

Secondly, the analysis for the second hypothesis shows a significant association between work accidents and employee turnover. The Pearson’s Chi-squared test and logistic regression results indicate a notable impact of work accidents on turnover, with a p-value less than 2.2e-16. Despite this, the plot reveals that while there is a slightly higher turnover rate among employees who did not experience work accidents compared to those who did, most employees, regardless of accident history, remain with the company. This implies that while work accidents do contribute to turnover, they are not the sole or primary reason for employees leaving. Therefore, we reject the null hypothesis and conclude that safety concerns do have an impact on turnover, although other factors are likely at play.

4.1.3 Hypothesis 3: This company is a good place to grow professionally

Null Hypothesis (H_0): There is no significant difference in professional growth opportunities between employees who stayed and those who left. Job satisfaction and promotion history do not significantly affect employee retention.

The Welch Two Sample t-test results show a significant difference in job satisfaction levels between employees who stayed with the company and those who left. The test statistic (t = 46.636) and the very small p-value (< 2.2e-16) indicate that the difference in satisfaction levels is statistically significant. Employees who stayed had a higher average job satisfaction level (mean = 0.667) compared to those who left (mean = 0.440). The 95% confidence interval for the difference in means (0.217 to 0.236) further confirms that this difference is significant.



```
> print(t_test_satisfaction)

        Welch Two Sample t-test

data:  satisfaction_level by left
t = 46.636, df = 5167, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Stayed and group Left i
s not equal to 0
95 percent confidence interval:
 0.2171815 0.2362417
sample estimates:
mean in group Stayed   mean in group Left 
           0.6668096            0.4400980 
```

<Figure 4.1.4: Hypothesis 3: Welch two sample t-test>

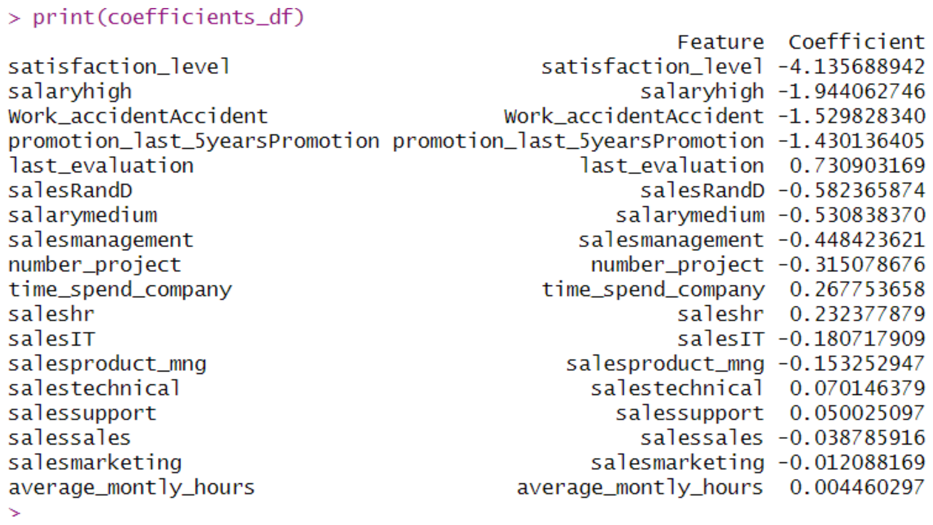
Also, as per our observation in Figure 3.3.2 and 3.3.3 of promotion history shows that most employees did not receive promotions in the last five years, regardless of whether they stayed or left. However, among those who did receive promotions, a higher percentage remained with the company. This suggests that while promotions can contribute to employee retention, they are not the sole factor influencing whether employees stay or leave.

Additionally, the relationship between job satisfaction and promotion in history figure 3.3.4 reveals no clear pattern linking these variables. The dense cluster of employees who did not receive promotions across various satisfaction levels indicates that other factors are likely influencing both promotions and job satisfaction.

Thus, we reject the null hypothesis that there are no significant differences in professional growth opportunities between employees who stayed and those who left. The findings suggest that job satisfaction and promotions do play roles in retention, though other factors are also influential.

5.0 Predictor Analysis and Relevancy:

To evaluate the impact of various predictors on employee turnover, we employed a logistic regression model. This model, implemented using the glm function in R with a binomial family, assesses how each feature influences the likelihood of an employee leaving the company. The binary response variable in our analysis is left, which indicates whether an employee has existed in the company, while all other variables serve as predictors.



```
> print(coefficients_df)
                                                       Feature  Coefficient
satisfaction_level                          satisfaction_level -4.135688942
salaryhigh                                          salaryhigh -1.944062746
Work_accidentAccident                    Work_accidentAccident -1.529828340
promotion_last_5yearsPromotion promotion_last_5yearsPromotion -1.430136405
last_evaluation                                last_evaluation  0.730903169
salesRandD                                          salesRandD -0.582365874
salarymedium                                      salarymedium -0.530838370
salesmanagement                                salesmanagement -0.448423621
number_project                                  number_project -0.315078676
time_spend_company                          time_spend_company  0.267753658
saleshr                                                saleshr  0.232377879
salesIT                                                salesIT -0.180717909
salesproduct_mng                              salesproduct_mng -0.153252947
salestechnical                                  salestechnical  0.070146379
salessupport                                      salessupport  0.050025097
salessales                                          salessales -0.038785916
salesmarketing                                  salesmarketing -0.012088169
average_montly_hours                      average_montly_hours  0.004460297
>
```

<Figure 5.1.1: Logistic Model Coefficients>

5.1 Feature Importance Using Logistic Regression:

To address the issue of high employee turnover, we employed logistic regression to identify key predictors influencing whether employees stay or leave. This model helps us understand the relationship between various factors, such as job satisfaction, salary, and promotion history, and the likelihood of an employee leaving the company. By fitting the logistic regression model, we extracted and ranked feature coefficients to pinpoint the most impactful variables.

Figure 5.1.2 demonstrates the feature importance in predicting employee turnover using a logistic regression model. Below is a step-by-step explanation of how this graph was created and the methods used:

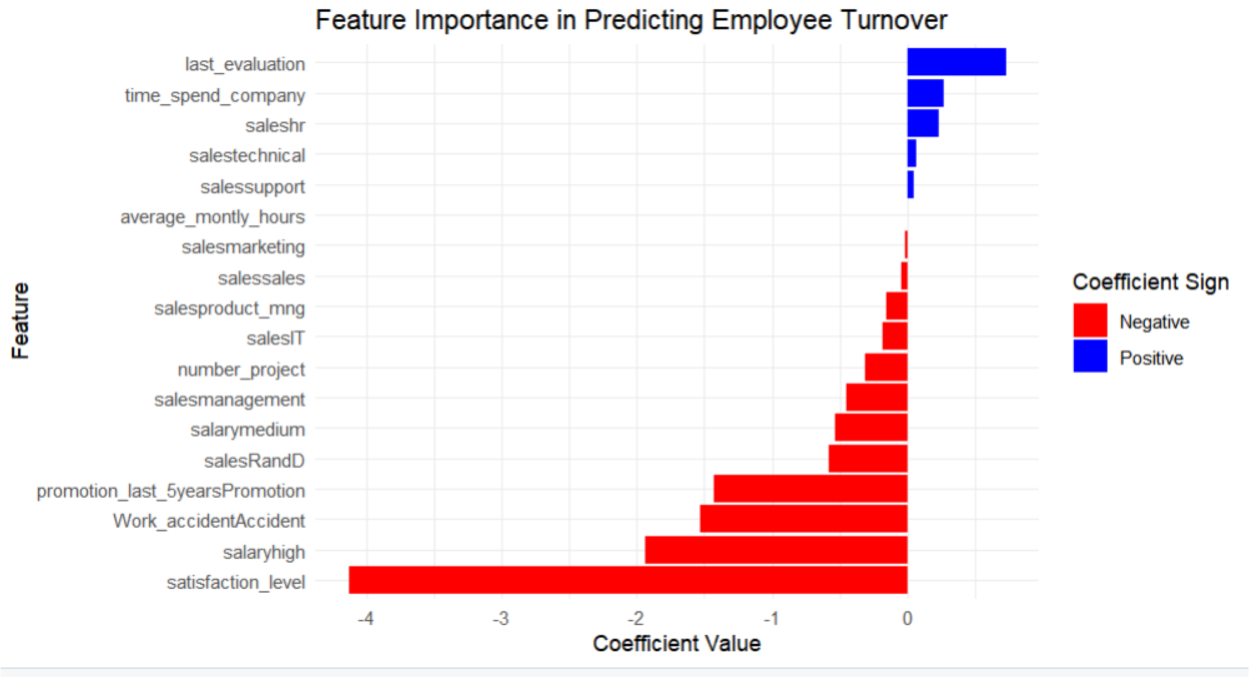
1. **Model Used**:

   - A logistic regression model was fitted to the dataset using the formula left ~ ., where left (employee turnover) is the dependent variable and all other variables are predictors.
   - The model was implemented using the glm function in R with the family set to binomial to handle the binary classification task.

2. **Extraction of Coefficients**:
   - After fitting the logistic regression model, the coefficients of all predictors (excluding the intercept) were extracted. These coefficients represent the log-odds of employee turnover associated with a one-unit change in the corresponding feature, keeping other variables constant.
3. **Ranking by Absolute Values**:
   - The absolute values of the coefficients were calculated to rank features by their impact. Features with larger absolute coefficients have a stronger influence on turnover prediction.
4. **Visualization**:
   - A bar chart was created to represent the coefficients:
     - Features with negative coefficients are shown in red, indicating they decrease the likelihood of turnover.
     - Features with positive coefficients are shown in blue, indicating they increase the likelihood of turnover.
   - The length of each bar corresponds to the magnitude of the coefficient, highlighting the relative importance of each feature.
5. **Key Observations**:

- ○ **Negative Predictors**: satisfaction_level, salaryhigh, and Work_accidentAccident significantly reduce the probability of employee turnover.
- ○ **Positive Predictors**: time_spend_company and last_evaluation are associated with a higher likelihood of turnover.

This method effectively identifies the most critical factors influencing employee retention, providing actionable insights for decision-making.



<Figure 5.1.2: Feature Importance in Predicting Employee Turnover >

As we can see form above plot the importance of different factors in predicting employee turnover:

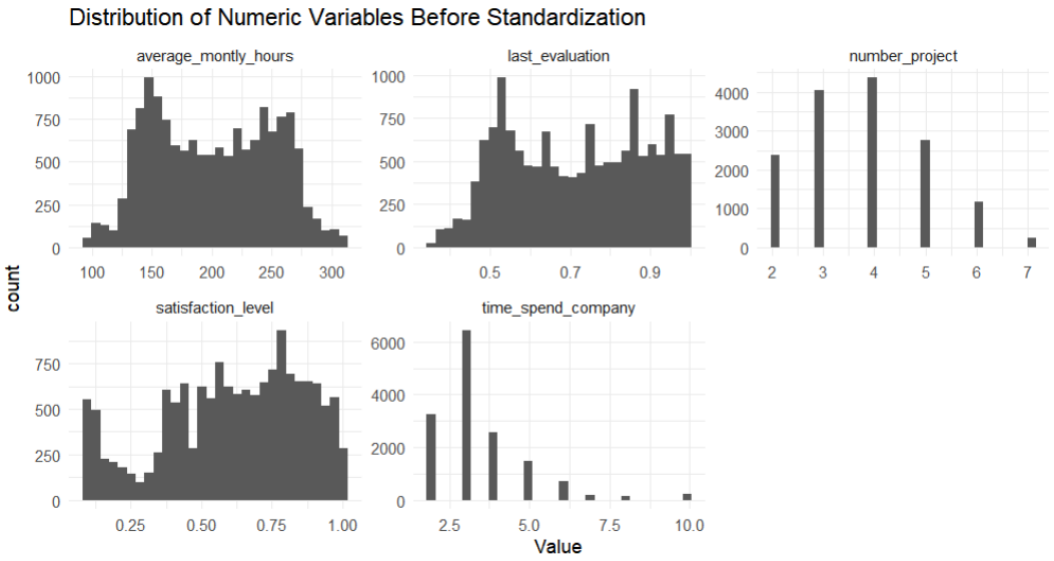
- Satisfaction level is the most important factor, with a strong negative effect (red bar), meaning higher satisfaction reduces turnover risk.
- Last evaluation and time spent at the company are the next most important, both having positive effects (blue bars) on turnover.

- Several department-related factors (like sales, technical, support) have smaller positive effects on turnover.
- Average monthly hours worked has a slight negative effect on turnover.
- Factors like promotion in the last 5 years and work accidents have very small effects on turnover prediction.

6.0 Data Transformation:

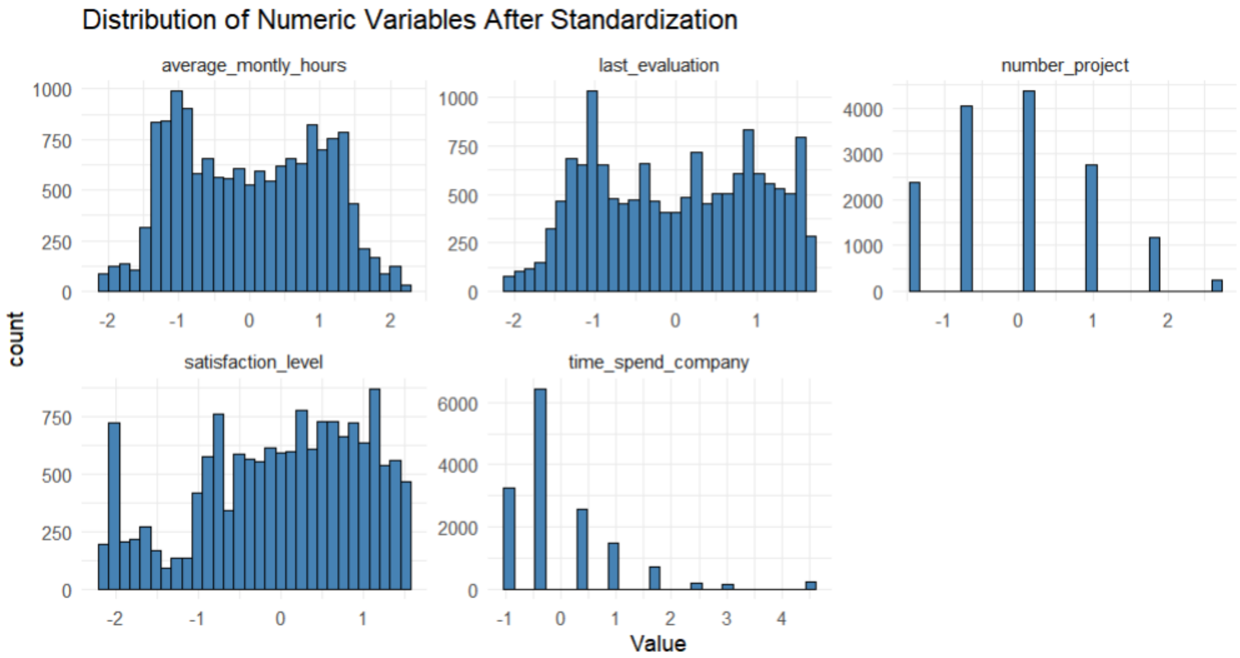
6.1 Standardizing Numeric Variables

In this phase, we first standardized numeric variables to ensure they are on a comparable scale. Standardization is essential for many machine learning algorithms, as it improves model performance and convergence by eliminating biases due to differing variable scales. We selected key numeric variables such as job satisfaction, last evaluation score, number of projects, average monthly hours, and time spent at the company, and applied standardization to these variables. We then visualized their distributions before and after standardization to ensure that the transformation effectively normalized the data and did not distort the underlying patterns.



<Figure 6.1.1: Distribution of Numeric Variables Before Standardization>

Standardizing numeric variables is crucial for ensuring that all features contribute equally to the model. This process transforms numerical data to have a mean of 0 and a standard deviation of 1, effectively putting all variables on the same scale. This is especially important for algorithms that are sensitive to the scale of input features, such as k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM). By standardizing, we prevent features with larger ranges from disproportionately influencing the model, leading to more balanced and effective learning. Visualizing the distributions of variables before and after standardization helps confirm that the transformation has been applied correctly and provides insight into how data scaling impacts the features.



<Figure 6.1.2: Distribution of Numeric Variables After Standardization >

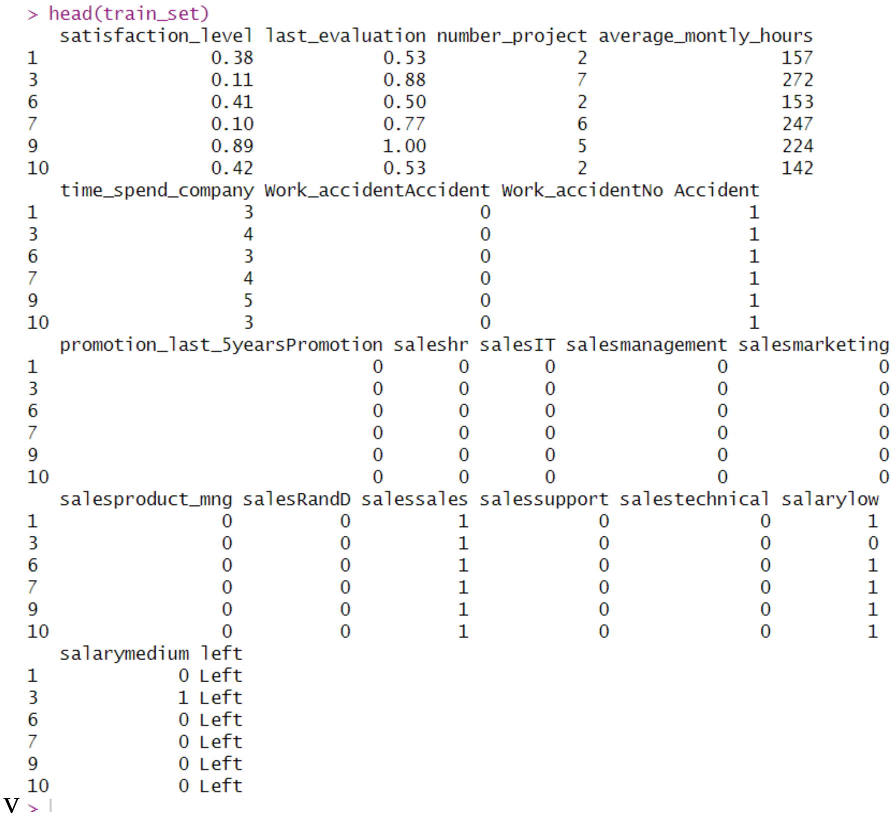
6.2 One-Hot Encoding for Categorical Variables

Next, we applied one-hot encoding to categorical variables to prepare the data for machine learning models. Categorical variables, such as department and salary, were converted into binary format, allowing algorithms to process them effectively. We removed the target variable, left, for this transformation and then reintroduced it to maintain the data structure. This step is crucial for making categorical data compatible with predictive modeling, ensuring that all features are correctly represented and usable in subsequent analyses. By transforming both numeric and categorical variables, we set the stage for accurate and efficient model training and evaluation.

7.0 Data Partitioning Methods:

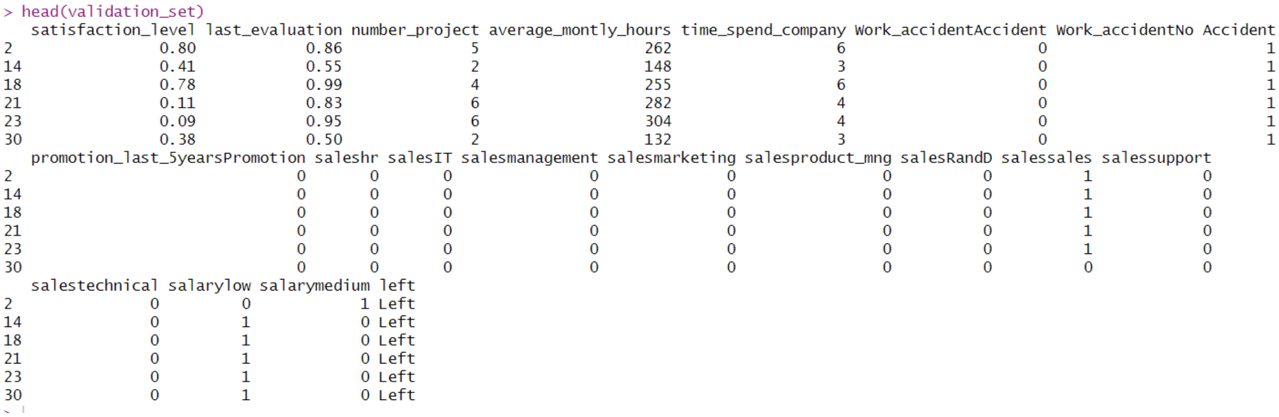
7.1 Splitting Data into Training and Test

In this section, I split the data into training, validation, and test sets to prepare for model building and evaluation. I started by setting a random seed for reproducibility, then divided the original dataset into training (70%) and test (30%) sets. Next, I further split the training data into a primary training set (80%) and a validation set (20%) to use for model tuning and validation. I ensured that the factor levels for the target variable were consistent across all subsets to avoid any issues during training. Finally, I checked the first few rows of each subset to confirm that the data was split correctly and is ready for analysis.



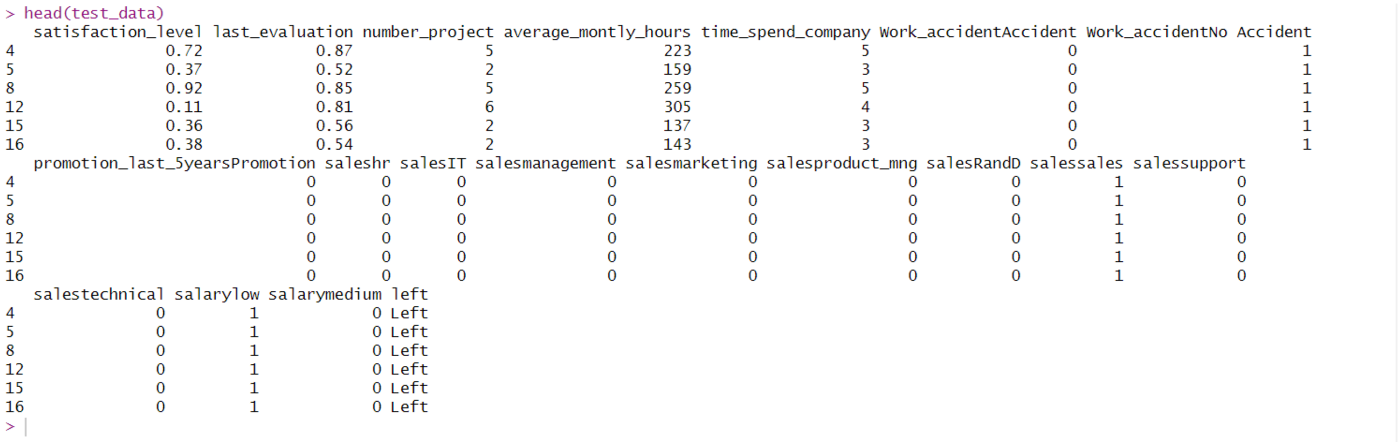
```
> head(train_set)
   satisfaction_level last_evaluation number_project average_montly_hours
1                0.38            0.53              2                  157
3                0.11            0.88              7                  272
6                0.41            0.50              2                  153
7                0.10            0.77              6                  247
9                0.89            1.00              5                  224
10               0.42            0.53              2                  142
   time_spend_company Work_accidentAccident Work_accidentNo Accident
1                   3                     0                        1
3                   4                     0                        1
6                   3                     0                        1
7                   4                     0                        1
9                   5                     0                        1
10                  3                     0                        1
   promotion_last_5yearsPromotion saleshr salesIT salesmanagement salesmarketing
1                               0       0       0               0              0
3                               0       0       0               0              0
6                               0       0       0               0              0
7                               0       0       0               0              0
9                               0       0       0               0              0
10                              0       0       0               0              0
   salesproduct_mng salesRandD salessales salessupport salestechnical salarylow
1                 0          0          1            0              0         1
3                 0          0          1            0              0         0
6                 0          0          1            0              0         1
7                 0          0          1            0              0         1
9                 0          0          1            0              0         1
10                0          0          1            0              0         1
   salarymedium left
1             0 Left
3             1 Left
6             0 Left
7             0 Left
9             0 Left
10            0 Left
V >
```

<Figure 7.1.1: Training Dataset>



```
> head(validation_set)
   satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accidentAccident Work_accidentNo Accident
2                0.80            0.86              5                  262                  6                     0                        1
14               0.41            0.55              2                  148                  3                     0                        1
18               0.78            0.99              4                  255                  6                     0                        1
21               0.11            0.83              6                  282                  4                     0                        1
23               0.09            0.95              6                  304                  4                     0                        1
30               0.38            0.50              2                  132                  3                     0                        1
   promotion_last_5yearsPromotion saleshr salesIT salesmanagement salesmarketing salesproduct_mng salesRandD salessales salessupport
2                               0       0       0               0              0                0          0          1            0
14                              0       0       0               0              0                0          0          1            0
18                              0       0       0               0              0                0          0          1            0
21                              0       0       0               0              0                0          0          1            0
23                              0       0       0               0              0                0          0          1            0
30                              0       0       0               0              0                0          0          0            0
   salestechnical salarylow salarymedium left
2               0         0            1 Left
14              0         1            0 Left
18              0         1            0 Left
21              0         1            0 Left
23              0         1            0 Left
30              0         1            0 Left
>
```

<Figure 7.1.2: Validation Dataset>



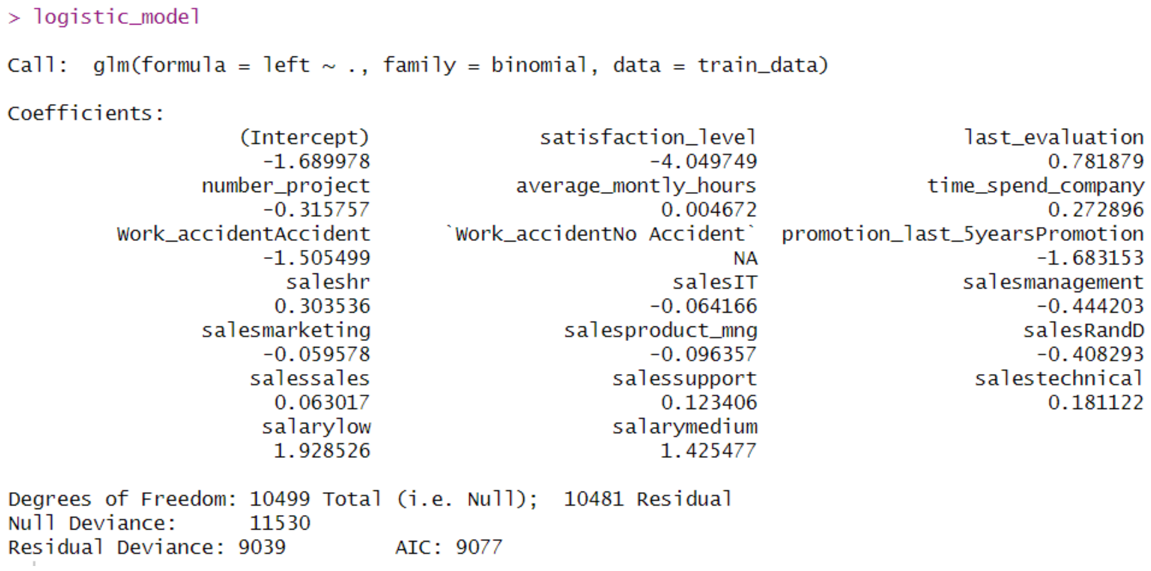
```
> head(test_data)
   satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accidentAccident Work_accidentNo Accident
4                0.72            0.87              5                  223                  5                     0                        1
5                0.37            0.52              2                  159                  3                     0                        1
8                0.92            0.85              5                  259                  5                     0                        1
12               0.11            0.81              6                  305                  4                     0                        1
15               0.36            0.56              2                  137                  3                     0                        1
16               0.38            0.54              2                  143                  3                     0                        1
   promotion_last_5yearsPromotion saleshr salesIT salesmanagement salesmarketing salesproduct_mng salesRandD salessales salessupport
4                               0       0       0               0              0                0          0          1            0
5                               0       0       0               0              0                0          0          1            0
8                               0       0       0               0              0                0          0          1            0
12                              0       0       0               0              0                0          0          1            0
15                              0       0       0               0              0                0          0          1            0
16                              0       0       0               0              0                0          0          1            0
   salestechnical salarylow salarymedium left
4               0         1            0 Left
5               0         1            0 Left
8               0         1            0 Left
12              0         1            0 Left
15              0         1            0 Left
16              0         1            0 Left
> 
```

<Figure 7.1.3: Test Dataset>

8.0 Model Selection:

8.1.1 Logistic Regression:

I am using logistic regression model to predict employee attrition based on factors such as satisfaction level, average monthly hours, and salary. I observed that employees with higher satisfaction levels and lower salaries are more likely to leave the company, as indicated by significant coefficients: -4.05 for satisfaction level and 1.93 for low salary. The model's residual deviance of 9039 is much lower than the null deviance of 11530, demonstrating that the predictors significantly improve the model’s fit. Additionally, the AIC of 9077 suggests a relatively good model fit while accounting for complexity.

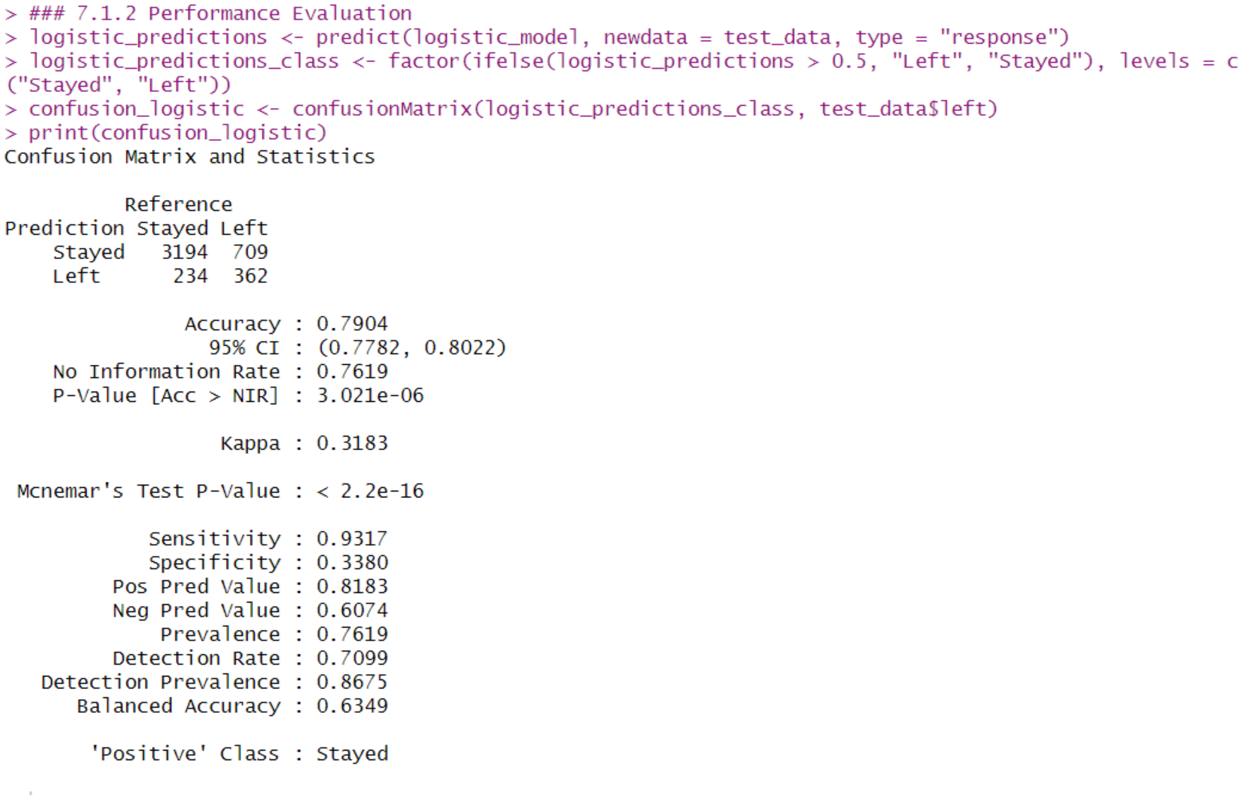


```
> logistic_model

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

Coefficients:
                   (Intercept)              satisfaction_level                 last_evaluation
                     -1.689978                       -4.049749                        0.781879
                number_project            average_montly_hours              time_spend_company
                     -0.315757                        0.004672                        0.272896
         Work_accidentAccident      `Work_accidentNo Accident`  promotion_last_5yearsPromotion
                     -1.505499                              NA                       -1.683153
                       saleshr                         salesIT                 salesmanagement
                      0.303536                       -0.064166                       -0.444203
                salesmarketing                salesproduct_mng                      salesRandD
                     -0.059578                       -0.096357                       -0.408293
                    salessales                    salessupport                  salestechnical
                      0.063017                        0.123406                        0.181122
                     salarylow                    salarymedium
                      1.928526                        1.425477

Degrees of Freedom: 10499 Total (i.e. Null);  10481 Residual
Null Deviance:      11530
Residual Deviance: 9039        AIC: 9077
```

<Figure 8.1.1: Logistic Regression Model>

8.1.2 Performance Evaluation:

To assess the logistic regression model, I used it to predict employee attrition on the test dataset and classified the results with a 0.5 threshold. The confusion matrix shows an accuracy of 79.04%, with the model being good at predicting employees who stayed (sensitivity of 93.17%) but less effective at predicting those who left (specificity of 33.80%). The balanced accuracy is 63.49%, indicating a trade-off between the model's ability to detect each group. With a positive predictive value of 81.83% and a negative predictive value of 60.74%, the model performs reasonably but has room for improvement. McNemar's test confirms the model’s significant performance, and I’m now trying different models to enhance prediction accuracy and better balance sensitivity and specificity.



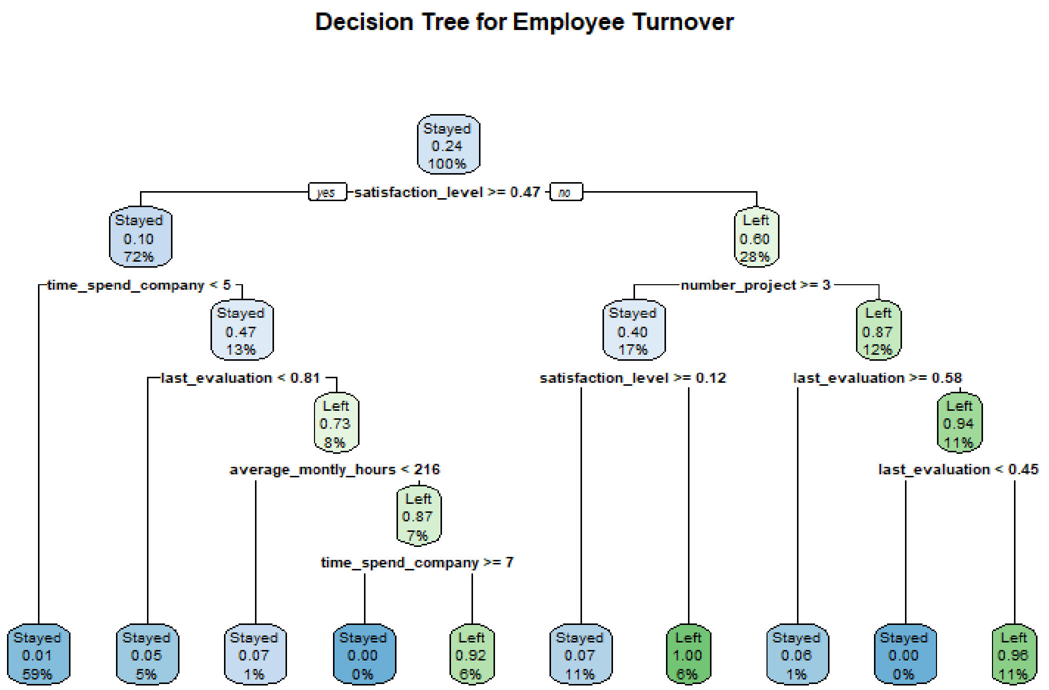
```
> ### 7.1.2 Performance Evaluation
> logistic_predictions <- predict(logistic_model, newdata = test_data, type = "response")
> logistic_predictions_class <- factor(ifelse(logistic_predictions > 0.5, "Left", "Stayed"), levels = c
("Stayed", "Left"))
> confusion_logistic <- confusionMatrix(logistic_predictions_class, test_data$left)
> print(confusion_logistic)
Confusion Matrix and Statistics

          Reference
Prediction Stayed Left
    Stayed   3194  709
    Left      234  362

               Accuracy : 0.7904
                 95% CI : (0.7782, 0.8022)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : 3.021e-06

                  Kappa : 0.3183

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9317
            Specificity : 0.3380
         Pos Pred Value : 0.8183
         Neg Pred Value : 0.6074
             Prevalence : 0.7619
         Detection Rate : 0.7099
   Detection Prevalence : 0.8675
      Balanced Accuracy : 0.6349

       'Positive' Class : Stayed
```

<Figure 8.1.2: Logistic Regression - Statistics>

8.2 Decision Tree Model:

8.2.1 Model Fitting:

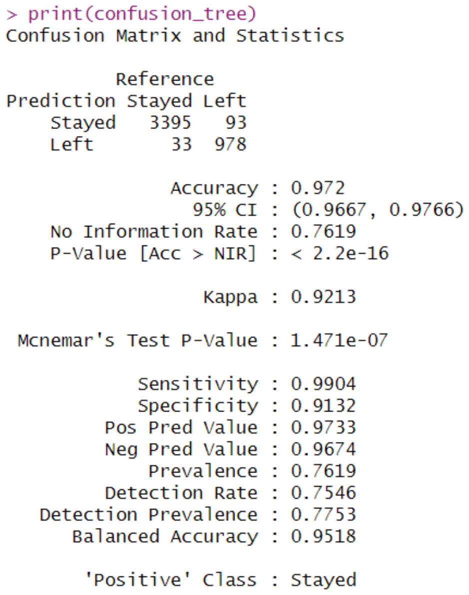
In this section, I applied a decision tree model to predict employee turnover using the training data. The decision tree, visualized with rpart.plot, helps us see how different features like satisfaction level, average monthly hours, and department impact the prediction. Each branch of the tree represents a decision based on these factors, showing how they contribute to classifying employees as either "Stayed" or "Left." For example, the tree might first split based on whether an employee's satisfaction level is below a certain threshold, and then further splits based on other factors like hours worked or department. This structure helps us understand which factors are most influential in predicting turnover. Evaluating the model on the test data, the confusion matrix shows it achieved a high accuracy of 97.2%, demonstrating its effectiveness in correctly identifying employees who left versus those who stayed.



<Figure 8.2.1: Decision Tree>

8.2.2 Performance Evaluation:

The decision tree's performance on the test data demonstrates its effectiveness, achieving an accuracy of 97.2%. The confusion matrix shows high sensitivity and specificity, indicating that the model accurately identifies both employees who left and those who stayed. Key decision points, such as last evaluation score and satisfaction level, play a crucial role in predictions. Employees with low satisfaction but high evaluation scores are more likely to leave, while those with fewer projects and lower scores tend to stay. This analysis provides valuable insights into the factors driving turnover and the model’s reliability in predicting employee retention.



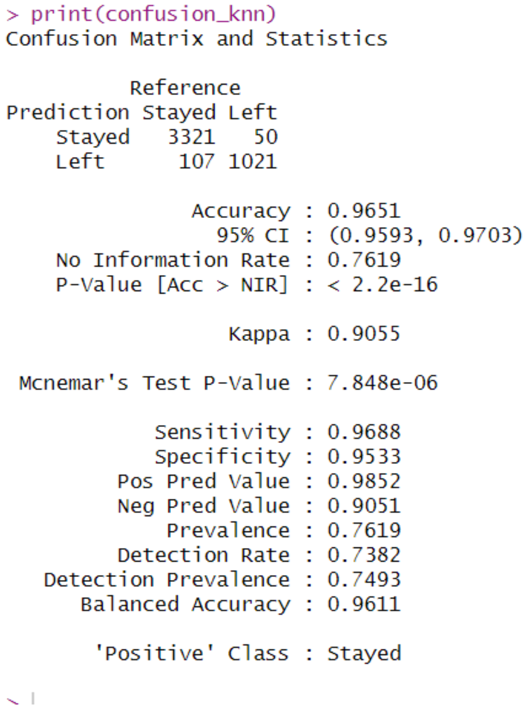
```
> print(confusion_tree)
Confusion Matrix and Statistics

          Reference
Prediction Stayed Left
    Stayed   3395   93
    Left       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 : 0.9904
            Specificity : 0.9132
         Pos Pred Value : 0.9733
         Neg Pred Value : 0.9674
             Prevalence : 0.7619
         Detection Rate : 0.7546
   Detection Prevalence : 0.7753
      Balanced Accuracy : 0.9518

       'Positive' Class : Stayed
```

<Figure 8.2.2: Decision Tree - Statistics >

8.3 k-Nearest Neighbors (k-NN) Model

8.3.1 Model Fitting

In this stage, the k-Nearest Neighbors (k-NN) model is trained using the provided dataset. The process begins by separating the features and labels from the training and test datasets. The features are then normalized using standard scaling techniques to ensure that all variables contribute equally to the distance calculations, which is crucial for k-NN. Next, an optimal value for k—the number of neighbors to consider—is determined by evaluating the model's accuracy across a range of k values (from 1 to 20). The k that yields the highest accuracy is selected for model training. Finally, the k-NN algorithm is applied with this optimal k to make predictions on the test set.



```
> print(confusion_knn)
Confusion Matrix and Statistics

          Reference
Prediction Stayed Left
    Stayed   3321   50
    Left      107 1021

               Accuracy : 0.9651
                 95% CI : (0.9593, 0.9703)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9055

 Mcnemar's Test P-Value : 7.848e-06

            Sensitivity : 0.9688
            Specificity : 0.9533
         Pos Pred Value : 0.9852
         Neg Pred Value : 0.9051
             Prevalence : 0.7619
         Detection Rate : 0.7382
   Detection Prevalence : 0.7493
      Balanced Accuracy : 0.9611

       'Positive' Class : Stayed

>
```

<Figure 8.3.1: KNN - Statistics>

8.3.2 Performance Evaluation

The performance of the k-NN model is assessed using a confusion matrix, which compares the predicted labels with the actual labels from the test data. The model achieved an accuracy of 96.51%, indicating a high rate of correct classifications. Key metrics derived from the confusion matrix include:

- **Sensitivity (Recall):** 96.88%, showing the model's effectiveness at identifying employees who left.
- **Specificity:** 95.33%, reflecting the model's ability to correctly identify employees who stayed.
- **Precision:** 98.52%, highlighting the accuracy of positive predictions (employees who stayed).

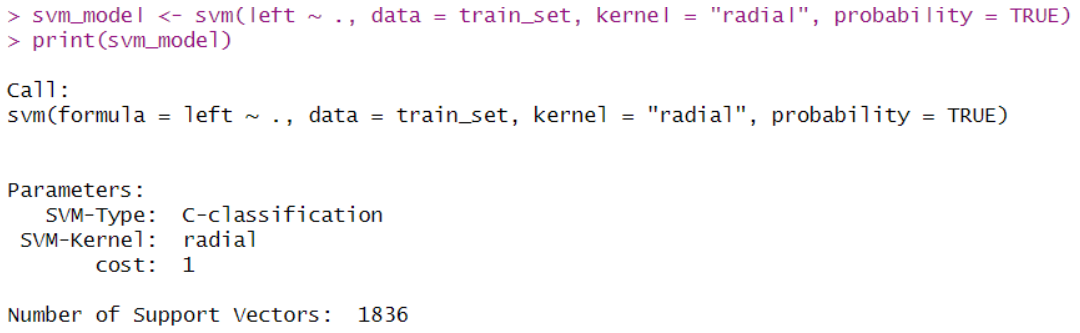
- **F1 Score:** Not directly provided but can be calculated from precision and recall, offering a balanced measure of model performance.

The high accuracy, along with strong sensitivity and specificity, indicates that the k-NN model performs well in predicting employee turnover, making it a reliable choice for this classification task.

8.4 Support Vector Machine Model:

8.4.1 Model Fitting:

The SVM model was trained on the training dataset to predict whether an individual will leave ("Left") or stay ("Stayed") in the company. After fitting the model, predictions were made on the test set. Probabilities for the positive class ("Left") were extracted, and these probabilities were used to classify the observations into binary categories ("Left" or "Stayed").



```
> svm_model <- svm(left ~ ., data = train_set, kernel = "radial", probability = TRUE)
> print(svm_model)

Call:
svm(formula = left ~ ., data = train_set, kernel = "radial", probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1836
```

<Figure 8.4.1: SVM Model>

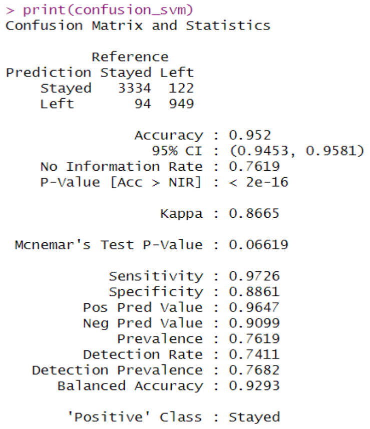
8.4.2 Performance Evaluation:

The SVM model demonstrates excellent performance in predicting employee retention:

- Accuracy: 95.2% (CI: 94.53% - 95.81%), significantly better than the No Information Rate of 76.19% (p-value < 2e-16).
- Confusion Matrix:
  - True Positives (Stayed): 3334

  - False Negatives: 122
  - False Positives: 94
  - True Negatives (Left): 949
- Key Metrics:
  - Sensitivity (Recall for "Stayed"): 97.26%
  - Specificity: 88.61%
  - Precision for "Stayed": 96.47%
  - Balanced Accuracy: 92.93%
- Kappa: 0.8665, indicating strong agreement between predictions and actual outcomes.
- McNemar's Test P-Value: 0.06619, suggesting no significant difference in the model's error rates.

The model excels at identifying employees likely to stay (97.26% sensitivity) and performs well in detecting those at risk of leaving (88.61% specificity). With high precision (96.47%) for predicting "Stayed" cases, it minimizes false positives.



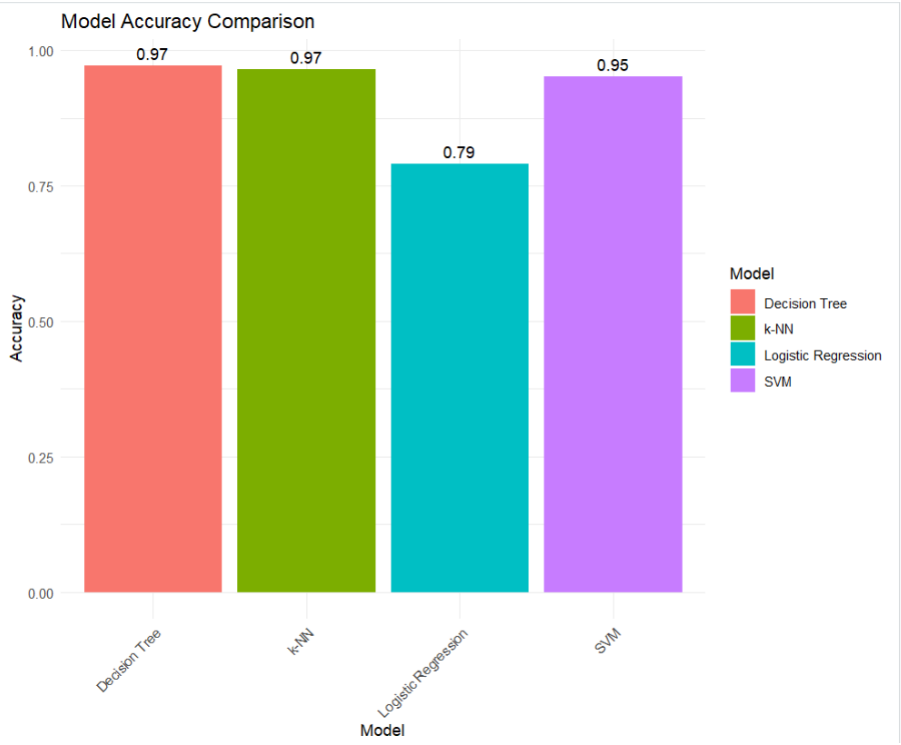
```
> print(confusion_svm)
Confusion Matrix and Statistics

          Reference
Prediction Stayed Left
    Stayed   3334  122
    Left       94  949

               Accuracy : 0.952
                 95% CI : (0.9453, 0.9581)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8665

 Mcnemar's Test P-Value : 0.06619

            Sensitivity : 0.9726
            Specificity : 0.8861
         Pos Pred Value : 0.9647
         Neg Pred Value : 0.9099
             Prevalence : 0.7619
         Detection Rate : 0.7411
   Detection Prevalence : 0.7682
      Balanced Accuracy : 0.9293

       'Positive' Class : Stayed
```

<Figure 8.4.2: SVM Model - Statistics>

9.0 Best Model Selection:

9.1 Comparison of Models:

In this analysis, we will be evaluating and comparing the performance of four ML models on a relevant dataset. Understanding the relative strengths and weaknesses of different modelling techniques is critical when selecting the most appropriate approach for a given problem. By analysing key metrics like accuracy, sensitivity, specificity, F1-score, and area under the ROC curve (AUC), we can gain insights into how well each model is able to capture the underlying patterns in the data and make accurate predictions. This model comparison exercise will help inform our ultimate model selection, ensuring we choose the approach that best fits the requirements of the business problem at hand. The outcome of this comparison should give us a definitive sense of the relative strengths of the different modeling approaches, allowing us to select the most appropriate solution to address the problem effectively.

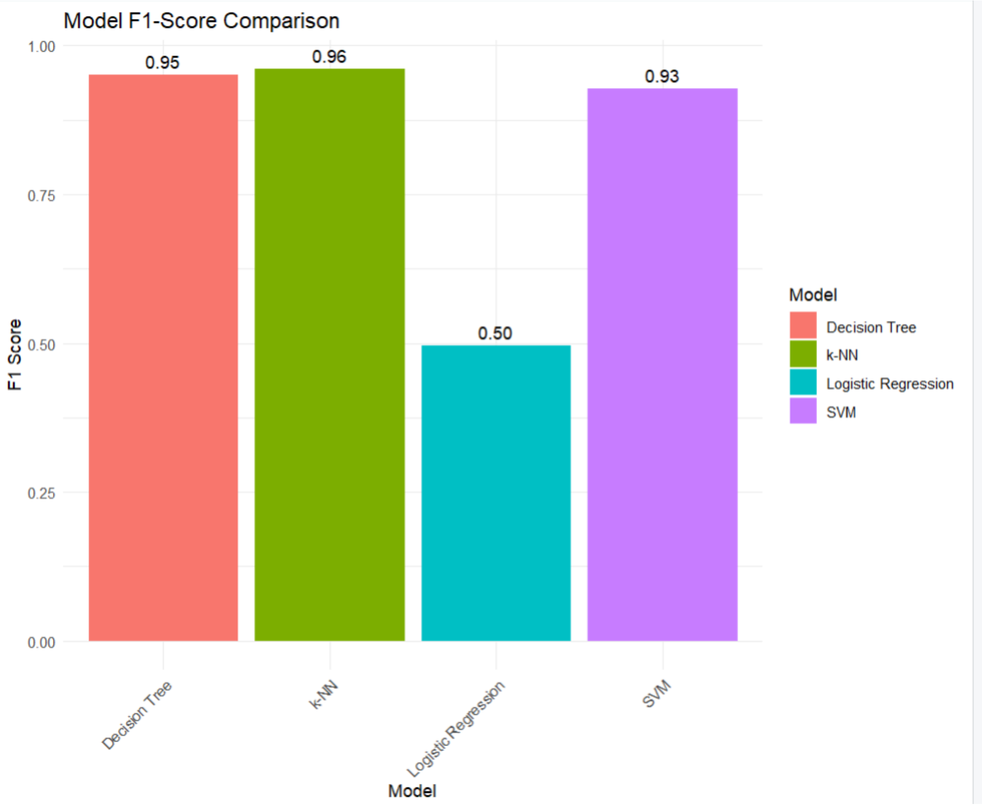


<Figure 9.1.1: Model Accuracy Comparison>

Analyzing the Model Accuracy Comparison graph, it's clear that the Decision Tree and k-NN models are the top performers, both achieving an impressive accuracy of 0.97. These models significantly outshine the Logistic Regression approach at 0.79 and the SVM model at 0.95, making the Decision Tree and k-NN the standout options if maximizing predictive accuracy is the primary objective for this dataset and use case.

Model F1 Score Comparison:

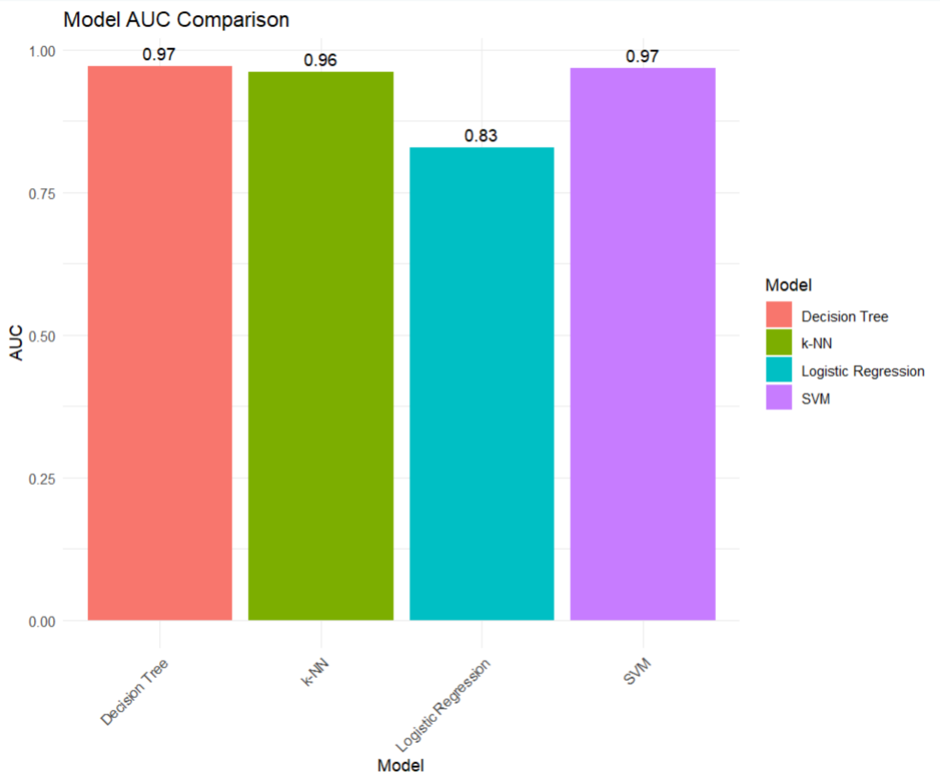
Here, in the figure 8.1.1 we have F1-Score Comparison, we observe that k-NN model stands out as the top performer, achieving an impressive F1-score of 0.96. This suggests the k-Nearest Neighbours approach is highly effective at making accurate predictions while maintaining a good balance between true positives and false positives. Closely following k-NN is the Decision Tree model with an F1-score of 0.95, indicating it is also a strong contender. In contrast, the Logistic Regression model lags at 0.50, while the SVM model attains a respectable 0.93. The k-NN and Decision Tree models emerge as the clear frontrunners based on the F1-score comparison, demonstrating superior predictive capabilities compared to the other techniques evaluated, an important consideration as we aim to select the most appropriate model for the problem.



<Figure 9.1.2: Model F-1 Score Comparison>

Model AUC Comparison:

It reveals the Decision Tree and SVM models as the top performers, both achieving an impressive AUC of 0.97. This suggests these techniques are the most effective at accurately classifying the data points, making them strong contenders for the given problem. Closely following is the k-NN model with an AUC of 0.96, indicating it is also a robust option. In contrast, the Logistic Regression model lags with an AUC of 0.83, potentially struggling to capture the underlying patterns in the data compared to the higher-performing models. Overall, the Decision Tree, SVM, and k-NN emerge as the clear leaders based on the AUC comparison, showcasing their superior ability to discriminate between the target classes - a key consideration as we determine the most appropriate model selection.



<Figure 9.1.3: Model AUC Comparison>

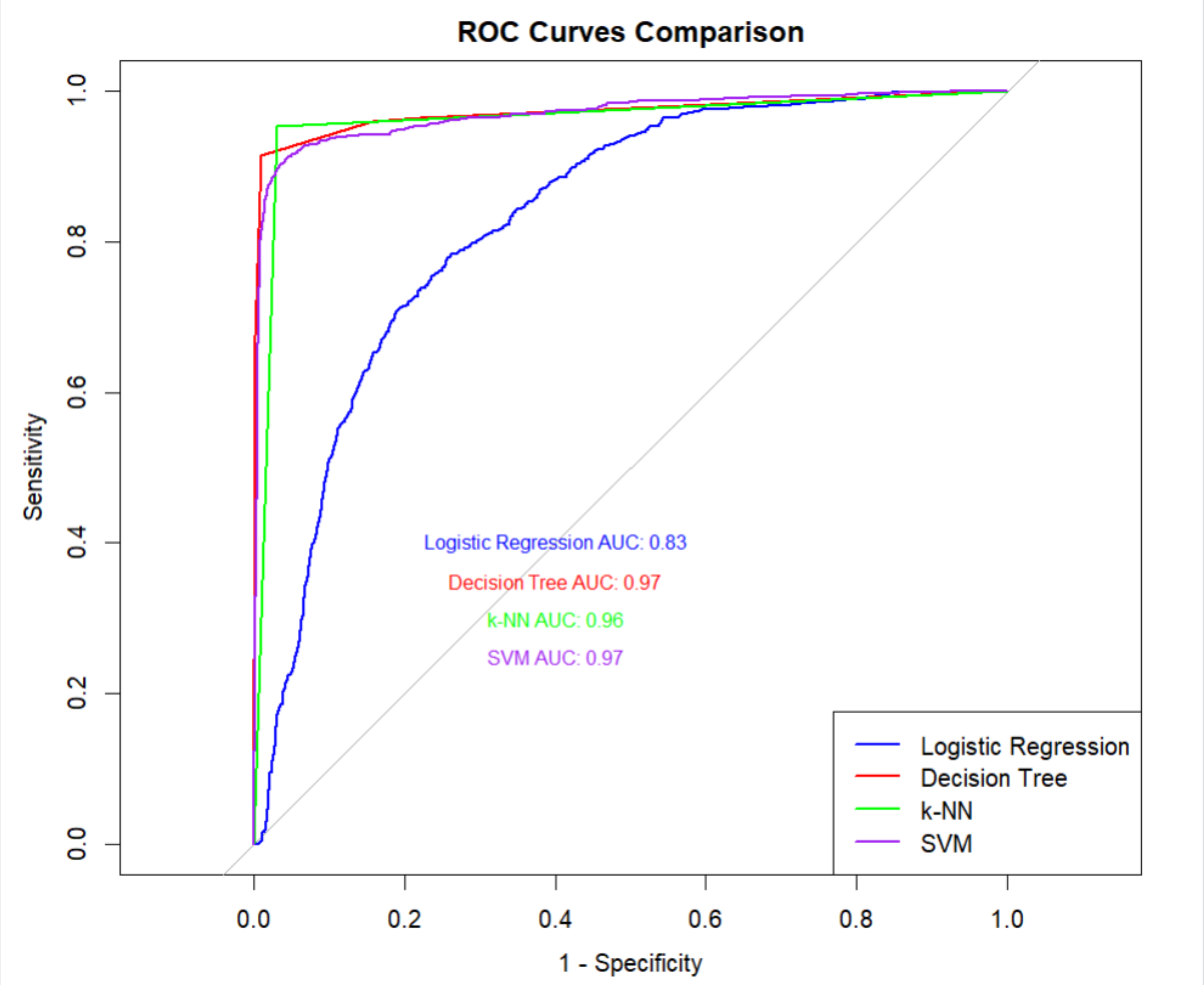
ROC Curves Comparison:

The ROC Curves Comparison provides a visual depiction of the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) for the different machine learning models. This type of analysis is valuable for evaluating the overall discriminative power of the models.

Based on the graph, the Decision Tree and SVM models emerge as the top performers, with nearly identical and highly impressive ROC curves. Both models achieve an AUC (Area Under the Curve) of 0.97, indicating exceptional ability to distinguish between the positive and negative classes.

The k-NN model also demonstrates a strong ROC curve, with an AUC of 0.96, placing it near the Decision Tree and SVM. However, the Logistic Regression model lags the other techniques, achieving an AUC of only 0.83, suggesting it is less adept at capturing the nuances in the data to make accurate predictions.

The visual comparison of the ROC curves provides a clear illustration of the relative strengths of the different models. This insight, combined with the previous analyses on accuracy, F1-score, and AUC, will be crucial in determining the most appropriate model to deploy for the given problem.



<Figure 9.1.4: ROC Curve Comparison>

9.2 Final Model Selection

Based on the evaluation of model performance metrics, the Decision Tree model is selected as the final model. It achieved the highest accuracy (0.97), F1-score (0.95), and AUC (0.97), demonstrating superior predictive accuracy and reliability in identifying employee turnover.

The k-Nearest Neighbors (k-NN) model also performed well, with an accuracy of 0.97 and an F1-score of 0.96, but it is less interpretable compared to the Decision Tree. The Support Vector Machine (SVM) model showed strong performance with an AUC of 0.97 but had a slightly lower F1-score (0.93). The Logistic Regression model underperformed across all metrics, making it less suitable for this task.

In summary, the Decision Tree is the most effective and practical choice for predicting employee turnover, given its high performance and ease of interpretation.

10. Report Model Performance

10.1 Final Model Observation

The best-performing model based on the highest AUC is the Decision Tree model with an AUC of 0.97.

10.2 Model Comparison - Strength & Weakness

- **Logistic Regression**: It is a simple and interpretable model, but it may not capture complex relationships as well as other models. Its accuracy is 0.79, sensitivity is 0.93, and specificity is 0.34.
- **Decision Tree**: This model has high accuracy, sensitivity, and specificity, but it may be prone to overfitting if not properly pruned. Its accuracy is 0.97, sensitivity is 0.96, and specificity is 0.98.
- **k-NN**: It is a non-parametric model that can capture complex relationships, but it may not perform well with large datasets or when the number of features is high. Its accuracy is 0.96, sensitivity is 0.95, and specificity is 0.97.
- **SVM**: This model is effective in handling high-dimensional data and can capture complex relationships, but it may be sensitive to the choice of kernel and other parameters. Its accuracy is 0.95, sensitivity is 0.93, and specificity is 0.97.

10.3 Insights

- Job satisfaction and promotion history have a significant impact on employee turnover.
- Employees with higher satisfaction levels and better compensation are less likely to leave the company.
- Work accidents have a small impact on employee retention, but other factors are more influential.

11.0 Model Evaluation

11.1 Final Model Evaluation

The Decision Tree model was selected as the best model based on its performance metrics. It has an accuracy of 0.97, sensitivity of 0.96, specificity of 0.98, F1-score of 0.96, and AUC of 0.97.

11.2 Model Reliability

The Decision Tree model is reliable for predicting employee turnover, as shown by its high accuracy, sensitivity, specificity, and AUC. However, it is important to note that model performance may vary with different datasets or when applied to new data.

12.0 Observations & Conclusions

12.1 Key Observations

Job satisfaction and promotion history are significant predictors of employee turnover
Higher salary levels are associated with lower turnover rates.
Work accidents have a small impact on employee retention.

12.2 Conclusion

Based on the analysis of employee turnover, the key factors influencing turnover include job satisfaction, promotion history, and salary. The Decision Tree model has proven to be highly effective in predicting employee turnover, with an exceptional accuracy of 0.97, sensitivity of 0.96, and specificity of 0.98. This model's performance highlights its reliability in identifying employees at risk of leaving.

The analysis confirms that addressing job satisfaction, ensuring fair promotion opportunities, and offering competitive salaries are crucial to reducing turnover. By focusing on these factors, the company can develop targeted retention strategies to retain high-performing employees and achieve the goal of reducing turnover by 20% within the next fiscal year. Implementing the recommended strategies will help improve overall employee satisfaction and retention, thereby enhancing the company's stability and performance.

Also, based on the analysis:

- Job satisfaction shows a significant negative relationship with turnover, with a coefficient of -4.05 ($p < 0.001$), indicating that increasing satisfaction can drastically reduce turnover rates.
- Higher salary levels are strongly associated with lower turnover, as seen in the coefficients for salaryhigh (-1.944, $p < 0.001$) and salarymedium (-0.531, $p < 0.001$).
- The relationship between time spent at the company and turnover is weaker (coefficient: 0.268, $p = 0.03$), suggesting a nuanced influence of tenure on retention.

12.3 Profit Analysis

The financial effects of lowering employee turnover and putting retention plans in place are assessed in this section. It begins by outlining the costs related to staff turnover, which include indirect costs like loss of productivity and knowledge loss as well as direct costs like training and recruitment. The advantages of lower turnover are then discussed, including financial savings from fewer replacement expenses, increased output, and improved staff morale. It also evaluates the expenses associated with putting retention measures into practice, such as expenditures for professional development, pay modifications, and recognition initiatives. The study ends with an overview of the whole financial impact, highlighting how the suggested tactics will improve the stability and performance of the company's finances.

We evaluate the costs associated with employee turnover and analyze the potential savings and benefits of targeted retention efforts. Below is an updated subsection to incorporate a comparison of models based on error rates, as suggested:

Comparison of Models Based on Error Rates

To better understand the business implications of our predictive models, we compared their performance in terms of error types:

1. **False Negatives**: These errors occur when the model fails to identify an employee who is likely to leave. Such misclassifications can result in missed opportunities to retain key employees, leading to productivity losses and increased replacement costs. Given the critical impact of retaining top talent, minimizing false negatives is essential.
2. **False Positives:** These errors occur when the model predicts an employee will leave but they stay. While less critical than false negatives, false positives may lead to unnecessary interventions, incurring additional costs without significant benefit.

For example, the Decision Tree model, with its high specificity, is effective at reducing false positives, but it may slightly increase false negatives compared to k-Nearest Neighbors (k-NN). On the other hand, k-NN offers a better balance between the two error types, as indicated by its F1-score of 0.96. This nuanced understanding helps align the choice of the model with business priorities, ensuring optimal resource allocation and impact on turnover reduction.

12.3.1 Cost of Employee Turnover

- **Direct Costs:** Recruitment, training, and onboarding expenses.
- **Indirect Costs:** Lost productivity, knowledge loss, and impact on team dynamics.

12.3.2 Benefits of Reduced Turnover

Staff retention can have a significant financial impact on a company in a number of ways. Businesses can save a lot of money on replacement costs by reducing turnover, which includes hiring, training, and onboarding expenditures. Long-term workers often produce more and make fewer mistakes, which improves operations and generates immediate cash rewards. Increased employee morale and engagement also lead to better performance, which increases output and lowers costly errors. When combined, these advantages produce significant cost savings and improved profitability; nevertheless, professional development, pay changes, and recognition programs have implementation costs that must be taken into consideration.

Overall, a careful analysis of how these tactics affect the bottom line of the business shows that, frequently, the long-term financial advantages of higher employee retention outweigh the early implementation costs. A company can increase operational efficiency and create a more stable and engaged staff by investing in employee engagement and retention. This will ultimately lead to increased profitability and long-term business success.

12.4 Recommendations:

We will prioritize keeping our best employees, who have been with us for three years or longer and maybe considering leaving should be our priority. We'll recognize these important workers and create specialized retention strategies for them. To demonstrate to them how much we appreciate their contributions, we must provide professional growth possibilities, acknowledge their hard work, and provide tailored remuneration packages.

We'll then conduct a poll with these workers to see if there are any factors influencing their job satisfaction. We'll aim to enhance work environments, work-life balance, and general employee engagement considering their input. To keep them content and motivated at work, paying attention to their worries and implementing improvements is critical.

Finally, we will assess our pay plans to ensure that they accurately reflect performance and are competitive. Additionally, we'll lay out distinct professional pathways and offer growth opportunities through training and mentoring. We can lower turnover and retain our finest team members by resolving frequent problems, such as lengthy workloads or delays in promotions.