September 2024

Human Capital Analytics

Addressing Employee Turnover Through Data-Driven Solutions

Joao Pedro Jacomossi

Table of Contents

Table of Figures

Table of Tables

Executive Summary

This project aimed to address high employee turnover at an engineering company by uncovering the reasons behind departures and predicting which current employees are at risk of leaving. The analysis provided actionable insights to guide strategic retention efforts and improve workforce stability.

Using a dataset provided by the HR department containing 14,999 observations and 10 variables, exploratory data analysis revealed that dissatisfaction, low salaries, limited career advancement opportunities, and excessive or insufficient workloads were significant contributors to employee turnover. Turnover peaked during mid-tenure periods, particularly in years 4–6, with year 5 showing the highest resignations.

A clustering model was developed to segment former employees into three distinct profiles to better understand the patterns of employee departures. Disengaged employees, characterized by low performance, minimal projects, and short tenures, likely left due to insufficient support or engagement. Overworked achievers, despite being high performers with heavy workloads, experienced burnout and unrecognized contributions. Stagnated employees, who were long-tenured, highly satisfied, and strong performers, ultimately left due to a lack of growth and promotion opportunities. While these profiles highlighted specific challenges faced by different groups, low salaries and infrequent promotions emerged as systemic issues affecting the entire workforce, particularly those who leave the organization. These foundational problems exacerbate the pain points within each cluster, driving turnover on a larger scale. Although employees in each profile respond to these challenges differently based on their unique circumstances, the systemic nature of these issues underscores their roots in organizational practices rather than individual factors.

After evaluating multiple machine learning algorithms, the Random Forest model demonstrated the best performance, achieving 99.07% accuracy, 96.47% sensitivity, and 99.39% precision on the test set using a 0.5 threshold. An additional analysis identified that satisfaction levels, a key turnover predictor, were primarily influenced by the number of projects and hours worked, reinforcing the importance of workload balance in retention.

The predictive model applied to the current employee dataset identified 8 employees at high risk of leaving, 19 at moderate risk, and 218 at low-moderate risk, with the remaining 11,183 employees classified as low risk. Employees from each risk category were ranked using an impact score, which combined salary, tenure, and performance evaluation to measure their organizational value. This ranking allows HR to prioritize retention efforts for employees whose departures would have the greatest impact on the company.

By focusing on these foundational issues and leveraging the predictive model and insights gained from this analysis, HR can implement a proactive, data-driven framework to enhance retention, improve satisfaction, and foster a more stable and engaged workforce.

Section 1: Introduction

In today's competitive business environment, a company's success relies not only on its products, services, or technology but also, and most importantly, on its human capital. Organizations operate as an integrated system where each component is interdependent and must function effectively to achieve the business's goals. Therefore, the effectiveness of a company is inherently linked to employee engagement and productivity. When the workforce lacks necessary resources and support, performance is likely to suffer, potentially leading to increased turnover. This, in turn, impairs the company's capacity to innovate and fulfill customer expectations, compromising its competitive edge and long-term viability. For this reason, organizations that retain their top talent gain a critical market advantage.

This project leverages analytics to assist a medium-sized engineering company in identifying the factors contributing to the resignation of experienced and high-performing employees. The examination of the Human Capital Analytics dataset, which contains information about each current and past employee's relationship with the company, will enable the identification of patterns and extraction of meaningful insights. This assessment will serve as the foundation for determining the most influential variables leading to turnover and developing a predictive model to estimate the likelihood of an employee leaving the company. Finally, the insights gained will enable the HR department and board of directors to implement strategies to improve retention and sustain organizational effectiveness.

Section 2: Business Goals

A business goal is defined as an accomplishment or target an organization aims to achieve to drive its overall success and growth. In this project, the business goals are the following:

1. **Understand the reasons for employee turnover:**
   Identify why employees, particularly the most skilled and experienced ones, are leaving the company.

2. **Reduce turnover among employees:**
   Implement strategies to retain critical talent and minimize attrition.

Section 3: Analytics Goals

The analytics goal refers to the outcome expected to be achieved using analytics. This is the guiding purpose behind the analytical efforts and ensures the project remains focused on supporting the business goal. In this project, the analytics goals are the following:

1. **Exploratory Data Analysis (EDA):**

   The first goal is to conduct a comprehensive exploratory data analysis to uncover patterns and factors contributing to employee turnover. This step provides insights into why employees leave the company, laying the foundation for further analysis.

2. **Turnover Profiling:**

   A specific focus will be placed on applying k-means clustering to the subset of employees who have left the company (left = 1). The clustering will segment these employees into distinct profiles based on shared characteristics. These profiles will help identify common patterns among those who leave, enabling the organization to understand the diverse reasons behind employee exits and develop more targeted retention strategies.

3. **Predictive Modeling for Turnover Reduction:**

   This phase consists of developing a predictive model to assess the risk of employee turnover. By analyzing historical data of employees who have resigned (left = 1) and those who have stayed (left = 0), the model will identify key factors influencing turnover. Once trained, it will be applied to current employees to predict their likelihood of leaving the company. This approach enables the organization to proactively identify at-risk employees and implement targeted interventions to improve retention and reduce turnover.

4. **Retention Optimization:**

   The final goal focuses on using the findings from clustering and predictive modeling to identify optimal working conditions and practices that enhance employee retention.

Section 4: Dataset Assumptions and Considerations

1. The dataset includes only employees who have been in the company for more than at least two complete years, therefore employees who have less than two years of tenure are not considered in this analysis.

2. The dataset presents employees who either left or stayed in the company within the past year.

3. Since the dataset does not include new hires, the total number of employees will remain constant. Consequently, the turnover ratio will be calculated based on the number of employees at the beginning of the year, rather than the average number of employees throughout the year.

Section 5: Data Preprocessing

The goal of data preprocessing is to ensure that the dataset is clean, consistent, and in a format that is suitable for analysis or modeling, ultimately leading to more accurate and reliable results. The key preprocessing steps include data cleaning, data transformation, feature engineering, and attributes definition.

Data Cleaning

- Ensuring that data types, formats, and variable names are appropriate across the dataset.
- Identifying and addressing outliers to prevent them from skewing the results.

Data Transformation

- Converting variables into appropriate data types for analysis and modeling.

Feature Engineering

- Deriving new variables from existing ones to better capture the underlying patterns in the data or to provide additional insights for analysis.

Attributes Definition

- Clearly defining each variable in the dataset, including its meaning, type, units of measurement, and the expected range of values.

5.1 Dataset

The original dataset comprises a data frame with 14,999 observations and 10 variables. These variables include 2 numeric, 6 integer and 2-character data types.

```
'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 5.1: Original Dataset Structure

To enhance the analysis and modeling of employee data, categorical variables such as “sales” and “salary,” as well as binary variables like “Work_accident,” “left,” and “promotion_last_5years,” have been converted into factors. This approach ensures that these variables are handled correctly in analysis, improves computational efficiency, and enhances interpretability.

The column “sales” has been renamed to “department” to better represent the predictor variable.

5.2 Impact Score

Given the company’s focus on understanding turnover among the “best” and most experienced employees, a new variable called “Impact Score” has been developed. This score, which ranges from 0 to 1, assesses the level of impact the loss of a specific employee would have on the company. The criteria for determining employee impact are:

1. **Time Spent in Company:**
    - This criterion measures employee tenure.
    - **Scoring:**
        - 0 for time_spend_company ≤ 3 years
        - 0.5 for time_spend_company between 4 and 6 years
        - 1 for time_spend_company > 6 years
    - **Weight:** 40% (0.4)
2. **Salary Level:**
    - This criterion measures the employee’s salary category.
    - **Scoring:**
        - 0 for salary == "low"

- 0.5 for salary == "medium"
- 1 for salary == "high"

- **Weight:** 20% (0.2)

3. **Last Evaluation Score:**

   - This criterion reflects the employee's performance based on the last evaluation.
   - **Scoring:**
     - 0 for last_evaluation < 0.6
     - 0.5 for last_evaluation between 0.6 and 0.8
     - 1 for last_evaluation > 0.8
   - **Weight:** 40% (0.4)

The impact score is calculated as a weighted sum of these individual scores, with the weights reflecting the relative importance of each criterion.

Formula:

$$Impact\ Score = (0.4\ x\ Score\ of\ time\_spend\_company) + (0.2\ x\ Score\ of\ salary) + (0.4\ x\ Score\ of\ last\ \_\ evaluation)$$

The reason for including 'salary' is that while the variables 'time_spend_company' and 'last_evaluation' provide valuable insights into "best" and "experienced" employees, there is still a possibility that some employees in less critical positions might be ranked similarly to those in more significant roles. To address this potential limitation, the salary variable has also been incorporated into the impact score. This helps differentiate between employees who are genuinely critical to the organization and those whose roles, despite having similar evaluation scores and tenures, may not be as impactful.

```
> str(employee.df)
'data.frame':   14999 obs. of  11 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        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ left                 : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ promotion_last_5years: Factor w/ 2 levels "0","1": 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 "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
 $ impact_score         : num  0 0.7 0.7 0.6 0 0 0.4 0.6 0.6 0 ...
```

Figure 5.2: Structure of Modified Dataset ("impact score" included).

5.3 Variable Definition

To achieve both business and analytics goals, it is imperative to understand and define the different variables present in the dataset. Their definitions can be observed in ***Table 5.1***.

Variable Name	Definition	Type
Satisfaction Level	This variable represents the satisfaction level of the employee. It ranges from 0 to 1, where 0 indicates the lowest level of satisfaction and 1 indicates the highest.	Numeric
last_evaluation	Last evaluation of the employee. It ranges from 0 to 1, with higher values indicating better performance evaluations.	Numeric
number_project	Number of projects the employee has been involved in. Higher values suggest greater involvement in projects.	Integer
average_montly_hours	Average number of hours the employee works per month.	Integer
time_spend_company	Number of years the employee has spent in the company.	Integer
work_accident	This variable indicates whether the employee has had a work-related accident (1= Had an accident and 0 = No accident).	Factor
left	This variable shows whether the employee has left the company (1= Left and 0 =Still with the company).	Factor
promotion_last_5years	This variable indicates whether the employee has been promoted in the last five years (1= Promoted and 0 = Not promoted).	Factor
department	Department to which the employee belongs.	Factor
salary	Salary level of the employee (High, Medium, or Low).	Factor

impact_score	The level of negative impact the loss of the employee would have on the company. It ranges from 0 to 1, with higher values indicating more significant impact.	Numeric

Table 5.1: Variable Definition

Section 6: Exploratory Data Analysis (EDA)

The exploratory data analysis phase involves investigating and understanding the data to find patterns, gain insights, detect anomalies, and test hypotheses with the support of graphical representations and summary statistics. This process is critical in forming a solid foundation for further analysis and model development.

6.1 Summary Statistics and Distributions

```
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      impact_score
0:12830       0:11428    0:14680               sales      :4140   high  :1237   Min.   :0.0000
1: 2169       1: 3571    1:  319               technical  :2720   low   :7316   1st Qu.:0.2000
                                               support    :2229   medium:6446   Median :0.4000
                                               IT         :1227                 Mean   :0.3478
                                               product_mng: 902                 3rd Qu.:0.5000
                                               marketing  : 858                 Max.   :1.0000
                                               (Other)    :2923
```

Figure 6.1: Summary Statistics

By observing the summary statistics ***(Figure 6.1)***, it is possible to observe that the dataset contains **no missing values**. Missing data requires special attention since it can have a significant negative impact on the analysis and modeling if not handled properly. Failure to address this issue introduces risk of biases, information loss, distorted correlations, and poor model performance. Since all values are present in this dataset it is possible to proceed with the analysis without the need for data exclusion or imputation.

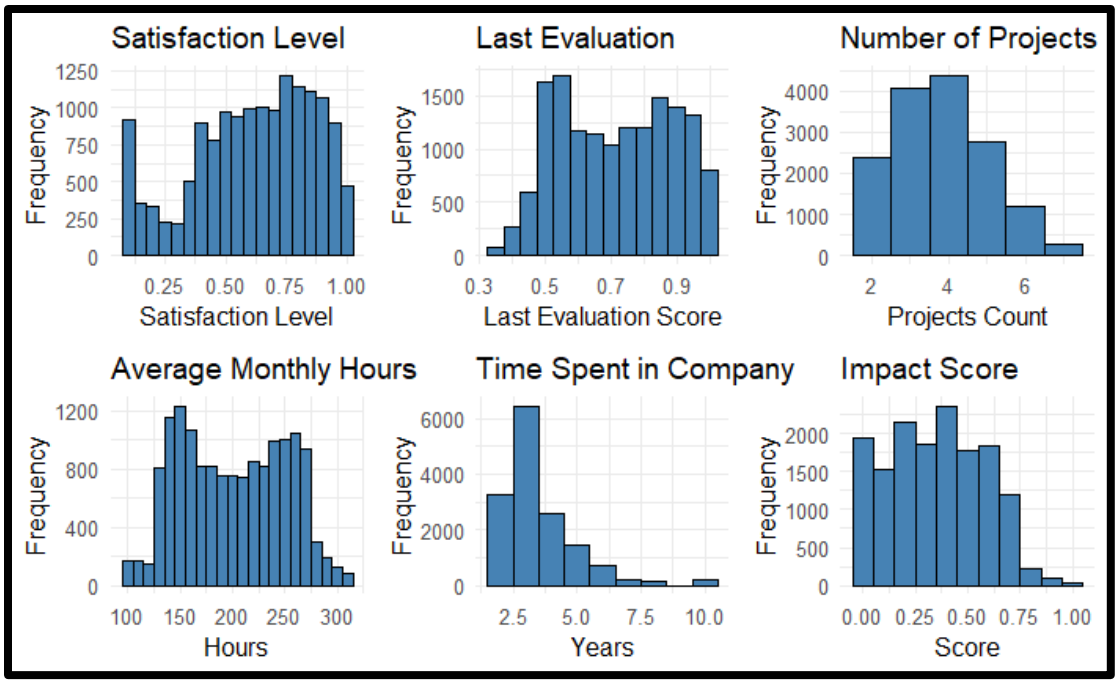

Figure 6.2: Histograms

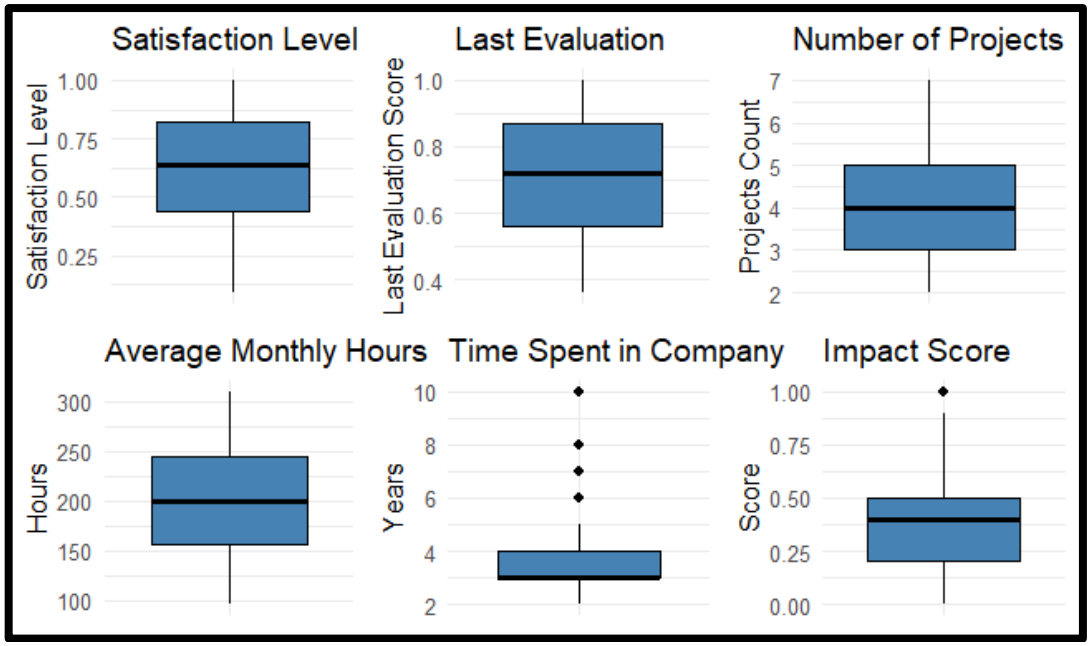

Figure 6.3: Boxplots

With the dataset complete, it is now possible to examine the key variables to gain insights into employee characteristics and trends.

Satisfaction Level:

The satisfaction level ranges from 0.09 to 1.00, indicating that while some employees are very dissatisfied, others are extremely satisfied. The mean satisfaction level is 0.6128, suggesting

that overall, employees tend to be satisfied. This is further reinforced by the histogram and boxplot, which shows most employees have a satisfaction level above 0.5, with the highest concentration around 0.75.

Last Evaluation:

The median of 0.72 and a mean of 0.7161 suggests that most employees have relatively high evaluation scores, indicating good performance. The higher quartiles show that many employees are evaluated very positively. The histogram shows relatively high concentrations of employees ranging from around 0.5 to 0.9.

Number of Projects:

Employees are typically involved in between 2 to 7 projects. The median value is 4, which indicates that most employees are involved in a moderate number of projects, with the majority falling between 2 and 4 projects. The mean value is slightly lower than the median, suggesting a slight tendency towards handling fewer projects overall. Only a small number of employees are involved in 6 or more projects, as indicated by the longer upper whisker in the box plot.

Average Monthly Hours:

The range of average monthly hours is relatively broad, from 96 to 310 hours. This range shows there is large variability in work hours. The median and mean are close, showing the average work hours is around 200 hours per month. The histogram shows a high concentration around 150 monthly hours and around 250 monthly hours and a fair concentration around 200 monthly hours. This suggests that most employees work full-time, with many having a regular workload, while others have a high workload. However extreme workloads on both ends of the spectrum are not widespread.

Time Spent in Company:

Most employees have been with the company for between 2 and 4 years, with the highest concentration of the workforce at 3 years, and the mean time spent being just under 3.5 years. This suggests that a significant portion of the workforce is relatively new. There are very few employees who have been in the company for over 7 years.

Work Accident:

About 85% of the employees have not had a work accident, while a smaller percentage (about 15%) have experienced one.

Left:

It can be observed that 11,428 employees are with the company and that 3,571 left. Based on the information provided, it is possible to conclude that the overall turnover rate for the organization is 23.8%, which is considerably high.

Promotion in Last 5 Years:

Based on the information provided, it is possible to conclude that only 2.12% of the employees have been promoted within the last 5 years. This is a very low percentage, suggesting that the company may offer limited career advancement opportunities.

Department:

The largest department is Sales, followed by Technical and Support. The smallest is Marketing, followed by Product Management. The distribution shows diverse roles within the company, with some departments having a significantly larger number of employees than others.

Salary:

Most employees (48.77%) have a low salary, 42.97% earn a medium salary, and only 8.25% earn a high salary.

Impact Score:

The mean impact score is 0.3478 and the median is 0.4 with the middle 50% falling within 0.2 and 0.5. This indicates that a large portion falls in the lower to middle range of importance.

6.2 Visual Exploration of Relationships

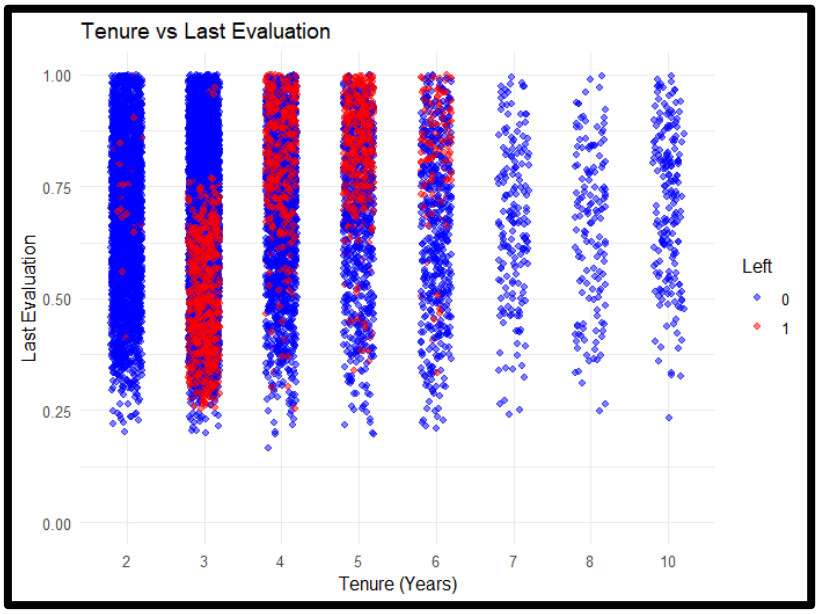

Figure 6.4: Scatterplot of Tenure vs Last Evaluation

Based the scatterplot ***(Figure 6.4)***, a high concentration of low and medium performers leaving the company within 3 years of employment is visible. It is also possible to observe a high concentration of high performers leaving within 4 to 6 years of employment. On the other hand, those who have been in the company for over 7 years are the ones with the least number of resignations.

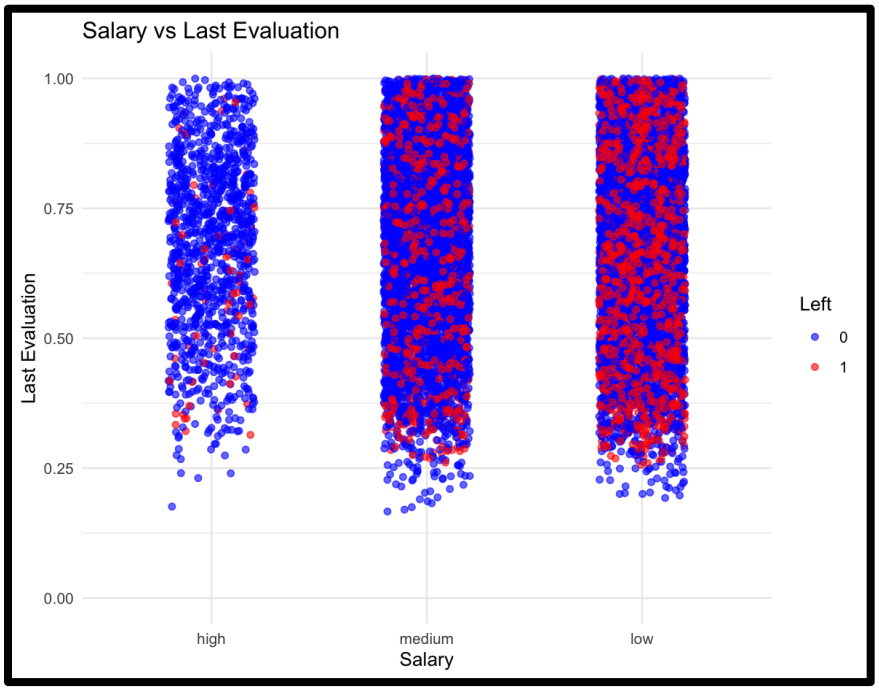

Figure 6.5: Scatterplot of Salary vs Last Evaluation

According to the scatterplot ***(Figure 6.5)***, it is visible that the highest concentration of employees who left are those with low salaries, followed by those with medium salaries and lastly those with high salaries, where there are very few resignations. The resignations are well distributed across all performance levels within each salary group. This suggests that salary has a significant impact in the employee's decision to leave, regardless of their performance level.

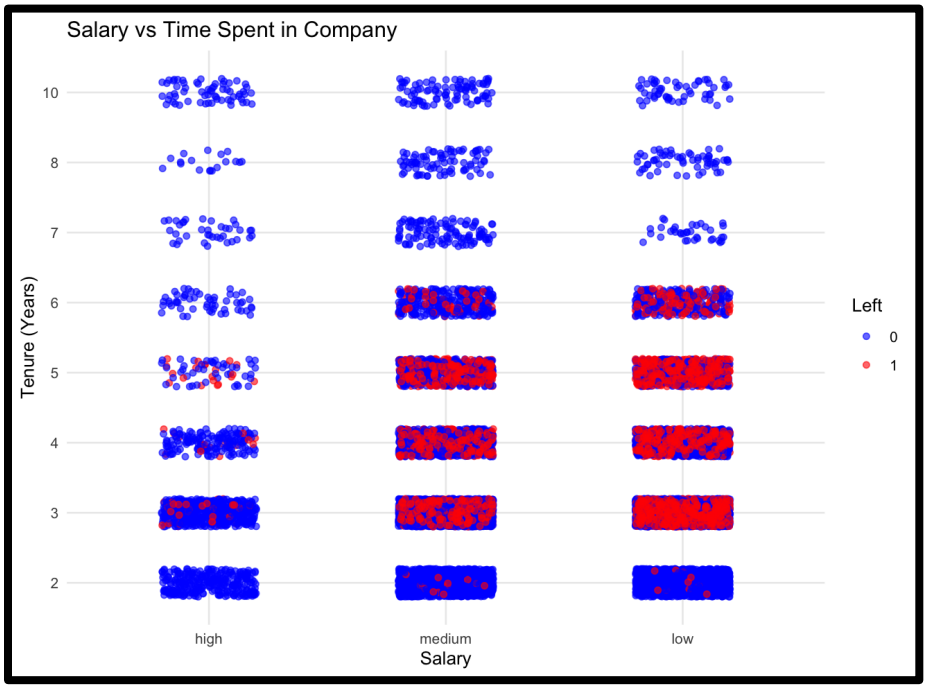

Figure 6.6: Scatterplot of Salary vs Time Spent in Company

The image ***(Figure 6.6)*** shows that resignations become significant at year 3 for both medium and low salary categories, and it continues until year 6. It is also possible to see that resignations for the period occur more frequently among employees earning a low salary followed by those in the medium salary category. There are very few resignations for employees earning high salaries.

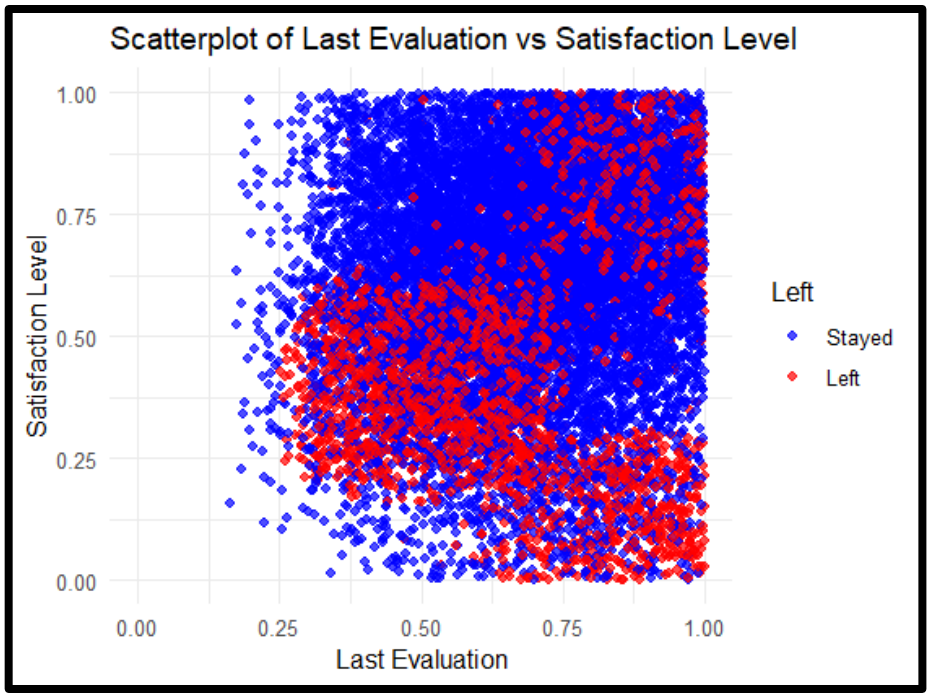

Figure 6.7: Scatterplot of Last Evaluation vs Satisfaction Level

The scatterplot in ***Figure 6.7*** reveals important insights into employee turnover patterns. Notably, there is a higher concentration of data points at higher satisfaction levels and performance evaluations. From this, three distinct groups of employees who are leaving the organization can be identified.

First, there are employees who are both satisfied and performing well, yet still decide to leave. This suggests that other factors besides job satisfaction and performance are leading to their exit. Second, there is a group of high-performing employees who have strong performance but are very dissatisfied. Finally, the third group consists of employees who are moderately dissatisfied and not performing at their best.

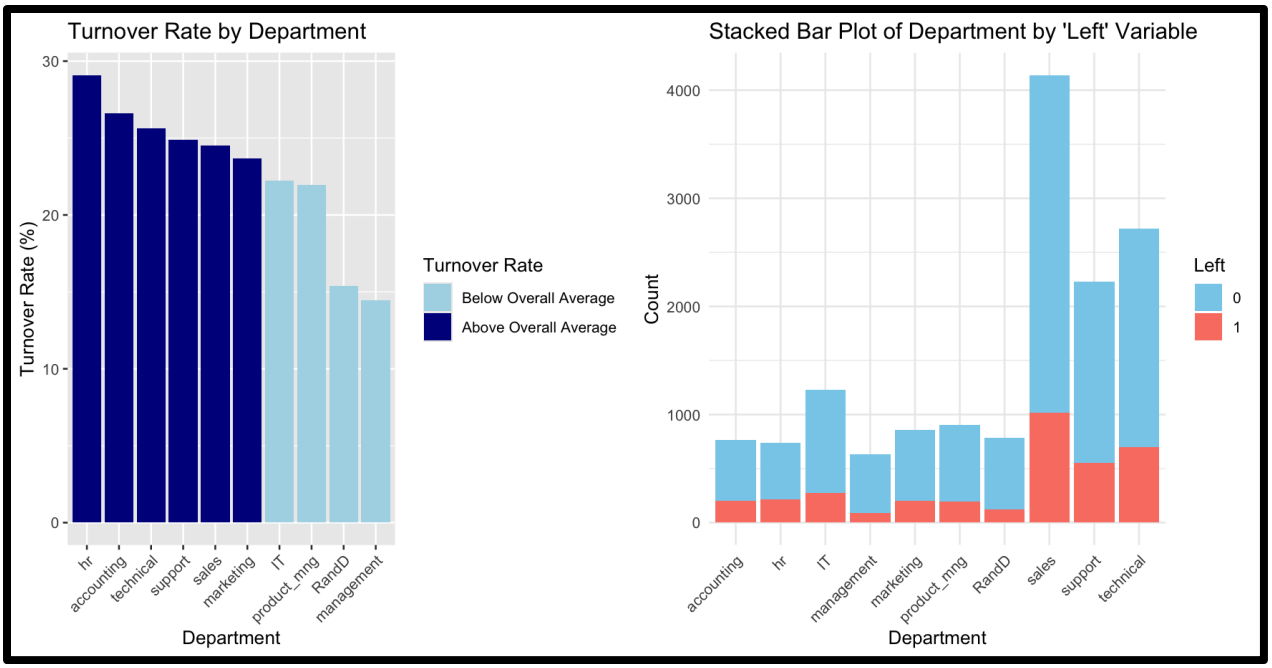

Figure 6.8: Turnover Rates and Resignation Counts by Department

The "Turnover Rate by Department" plot in ***Figure 6.8*** represents the proportion of resignations across different departments. In this plot, departments with turnover rates above the company's overall average are highlighted in dark blue, while those with rates below the average are shown in light blue. This distinction is important because it shows departments with higher turnover rates that may need interventions.

In contrast, the "Stacked Bar Plot of Department by ‘Left’ Variable" in ***Figure 6.8*** presents a different perspective by showing the distribution and proportion of employees who left versus those who stayed within each department. This plot reveals that HR and Accounting departments have a higher proportion of resignations compared to other departments, such as Management and Research & Development (R&D), which have lower turnover rates.

Together, these plots complement each other: the "Turnover Rate by Department" plot identifies which departments are experiencing higher turnover relative to the company’s average, while the "Stacked Bar Plot" depicts in detail the distribution of resignations and retention within each department. By analyzing both plots, it is possible to better understand where turnover occurs the most and how it changes across departments.

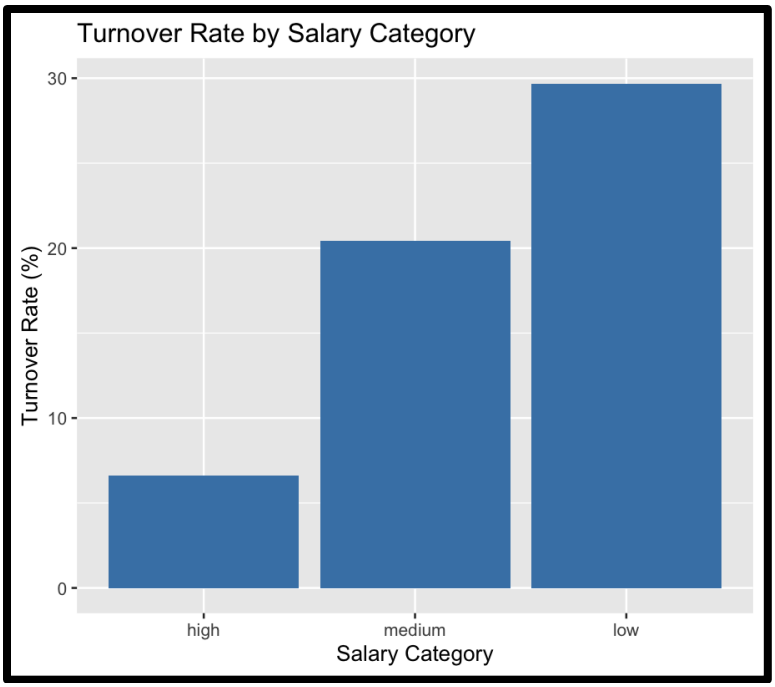

Figure 6.9: Turnover Rate by Salary Category

In ***Figure 6.9***, it is possible to observe an inverse relationship between turnover rate and salary. In other words, the lower the salary the higher the turnover rate. This suggests that employees earning lower salaries are significantly more likely to leave the company.

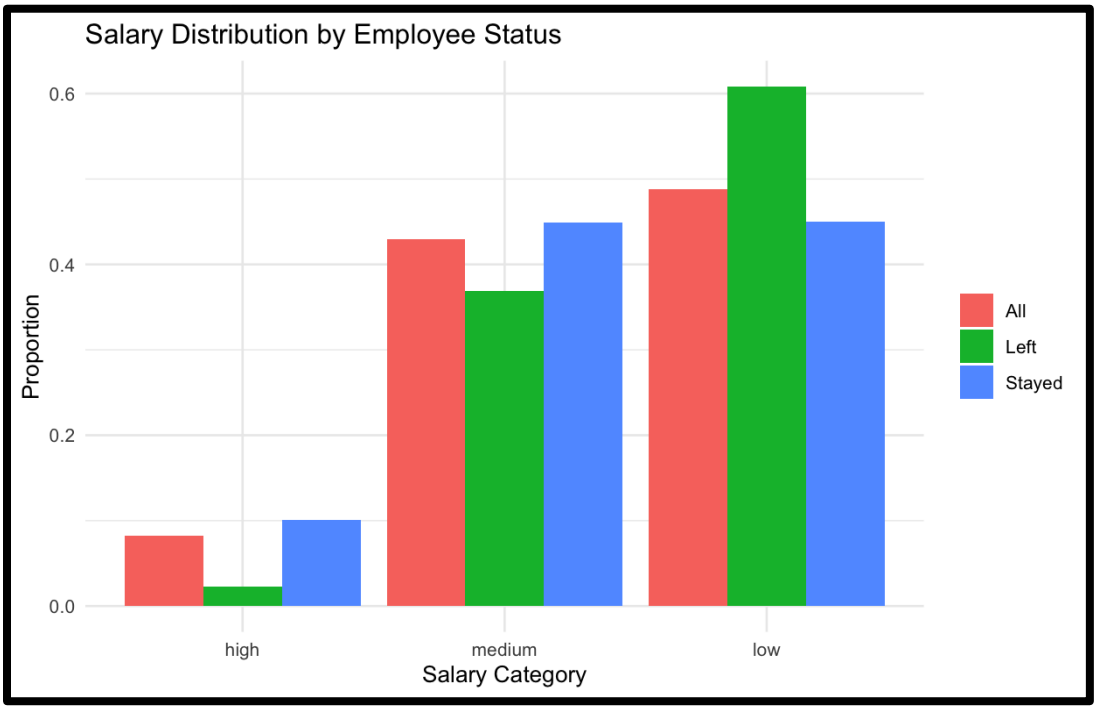

Figure 6.10: Salary Distribution by Employee Status

The plot ***(Figure 6.10)*** reinforces the findings by showing the differences in salary distribution between employees who left, stayed, and the overall workforce. Across all employees, 48.8% are in the low salary group, 42.9% in the medium group, and only 8.2% in the high group. Among those who stayed, the distribution is more balanced, with 45% in the low salary group, 44.9% in the medium group, and 10.1% in the high group. For those who left, the difference is clear, where 60.8% are in the low salary group, compared to 36.9% in the medium group and just 2.3% in the high group.

This analysis highlights that the high turnover rate in the low-salary category is not merely a result of the larger number of employees in this group. Instead, low salaries are a clear driver of employee departures, as individuals in this category are disproportionately represented among those who left. Addressing the needs of low-salary employees through improved compensation could play a key role in reducing turnover and fostering a more stable and engaged workforce.

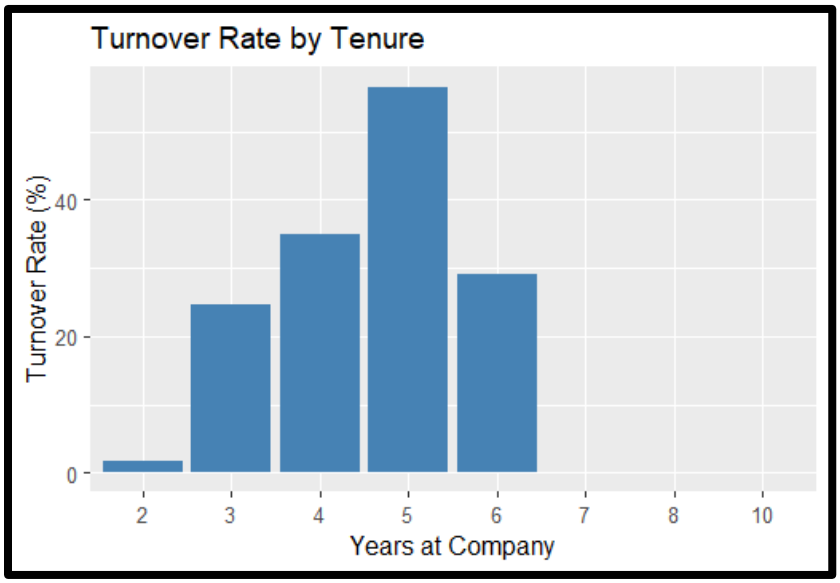

Figure 6.11: Turnover Rate by Tenure

The plot in ***Figure 6.11*** shows that the turnover rate starts increasing over the years within the company until it reaches its peak at year 5. Following this peak, the turnover rate declines in year 6 and becomes negligible in year 7 and beyond. This suggests that a higher number of people leaving at year 5 may be related to lack of promotion offers since only 2.12% of the observations have been promoted within the last 5 years.

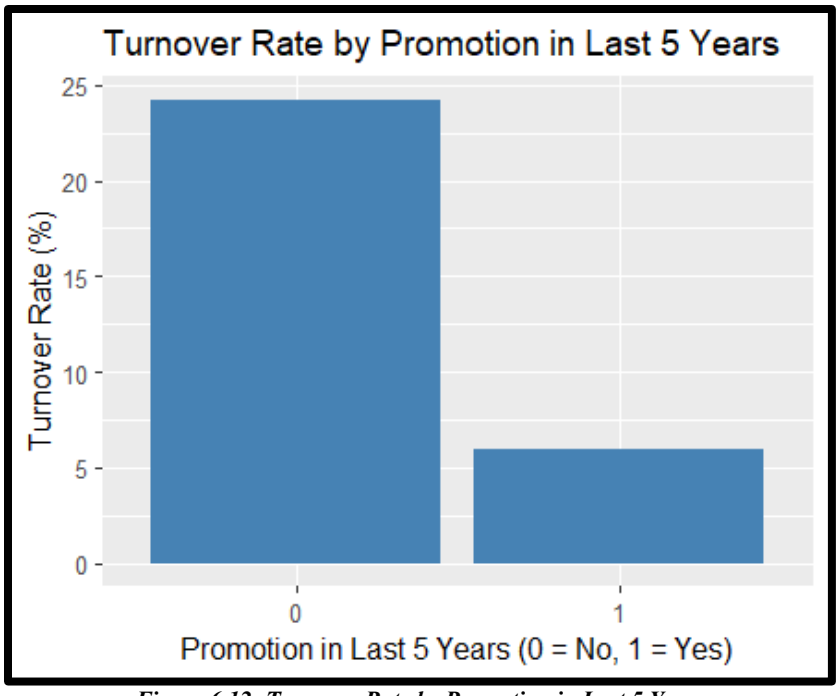

Figure 6.12: Turnover Rate by Promotion in Last 5 Years

The plot ***(Figure 6.12)*** shows the turnover rate across promotion status. According to the information provided, it is possible to observe that employees that haven't been promoted within the last 5 years have approximately 4 times higher turnover rate compared to employees that have. This indicates that promotions are a significant factor in reducing employee turnover, as they likely provide recognition and career growth opportunities.

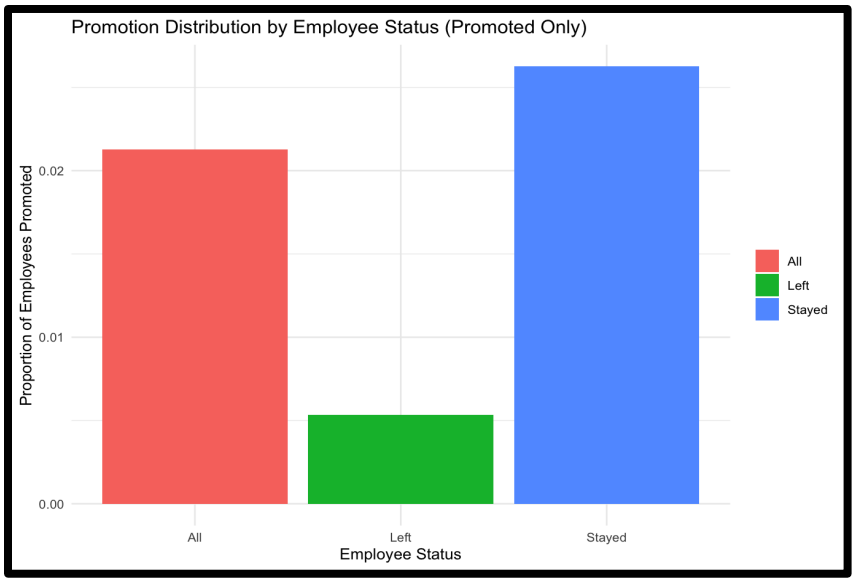

Figure 6.13: Promotion Distribution by Employee Status

The chart ***(Figure 6.13)*** highlights the distribution of promotions among all employees, those who stayed, and those who left. Among all employees, only 2.13% were promoted. Breaking this down, 2.63% of employees who stayed were promoted, while only 0.53% of employees who left received a promotion. This stark difference suggests that the lack of promotion opportunities is closely tied to employee turnover, as those who left were far less likely to have experienced career advancement.

To further investigate the hypothesis that one of the reasons for turnover is that employees are not being promoted, a chi-squared test between variables 'left' and 'promotion_last_5years' was performed.

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

data:  table_promotion_turnover
X-squared = 56.262, df = 1, p-value = 6.344e-14
```

Figure 6.14: Chi-Squared Test Between Variables 'left' and 'promotion_last_5years'

According to the test **(Figure 6.14)**, which shows a very small p-value, there is a strong association between promotion status and whether an employee leaves the company.

Having established a significant association between promotion status and employee turnover, the next step is to delve deeper into turnover patterns. Therefore, the distribution of turnover rates among employees who have been in the company for less than 5 years in relation to their promotion status as well as of those who have been in the company for over 5 years will be analyzed.

Figure 6.15: Turnover Rate by Promotion Status for Employees with at Least 5 Years Tenure

Figure 6.15 shows that employees with less than five years of tenure experience a significantly higher turnover rate (20.9%) if they have not been promoted compared to those who have been promoted (8.07%). This suggests that a lack of promotions early in an employee's career strongly contributes to turnover, emphasizing the importance of providing clear growth opportunities for newer employees.

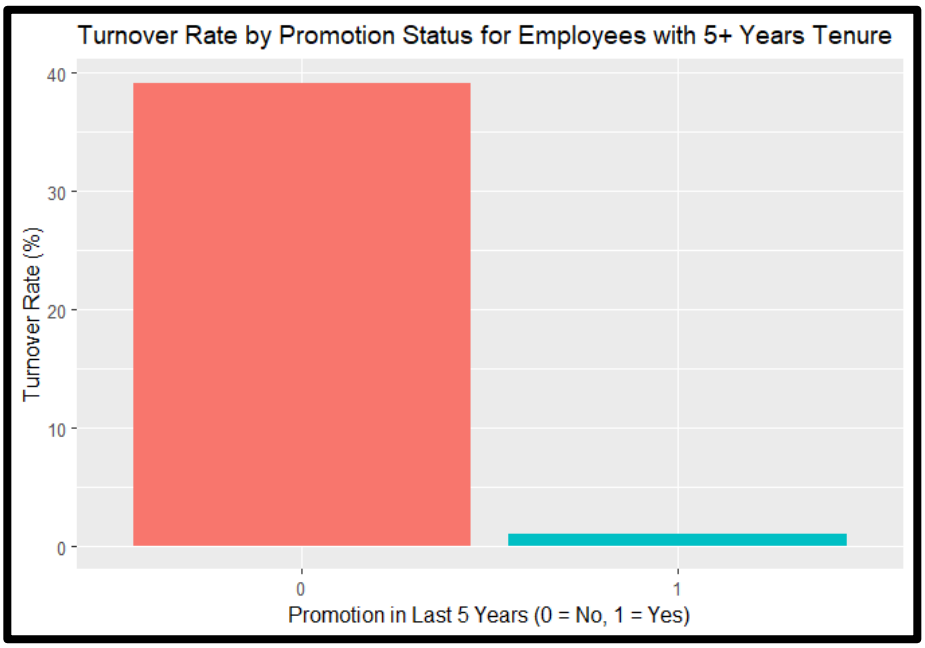

Figure 6.16: Turnover Rate by Promotion Status for Employees with More Than 5 Years Tenure

Figure 6.16 reveals an even sharper contrast for employees with five or more years of tenure. For those who have not been promoted, the turnover rate almost doubles compared to short-tenured employees, increasing from 20.9% to 39.2%. On the other hand, for employees who have been promoted, the turnover rate decreases drastically, from 8.07% to just 1.04%. This comparison highlights that while promotions significantly reduce turnover in both groups, their impact is particularly pronounced for long-tenured employees, where the absence of promotions leads to much higher attrition.

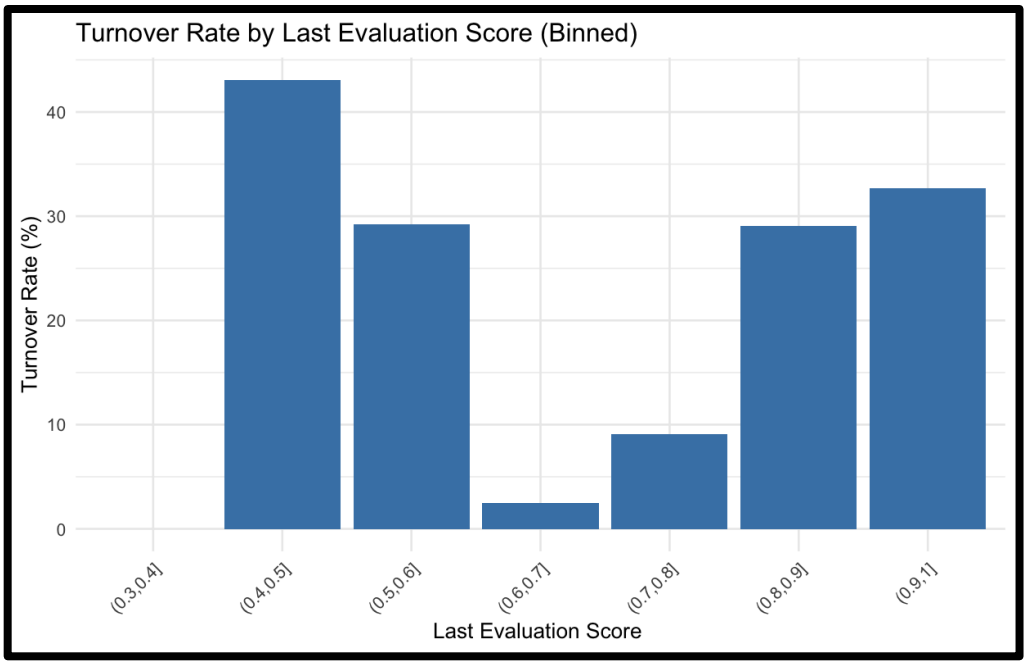

Figure 6.17: Turnover Rate by Last Evaluation Score

The area chart ***(Figure 6.17)*** shows that turnover ratio is the highest for low performers, lowest for medium performers, but also well above the overall average of 23.8% for a significant portion of high performers. This suggests that employees who perform poorly tend to leave due to their underperformance while those with strong evaluation scores may leave due to seeking better opportunities in terms of salary and professional growth.

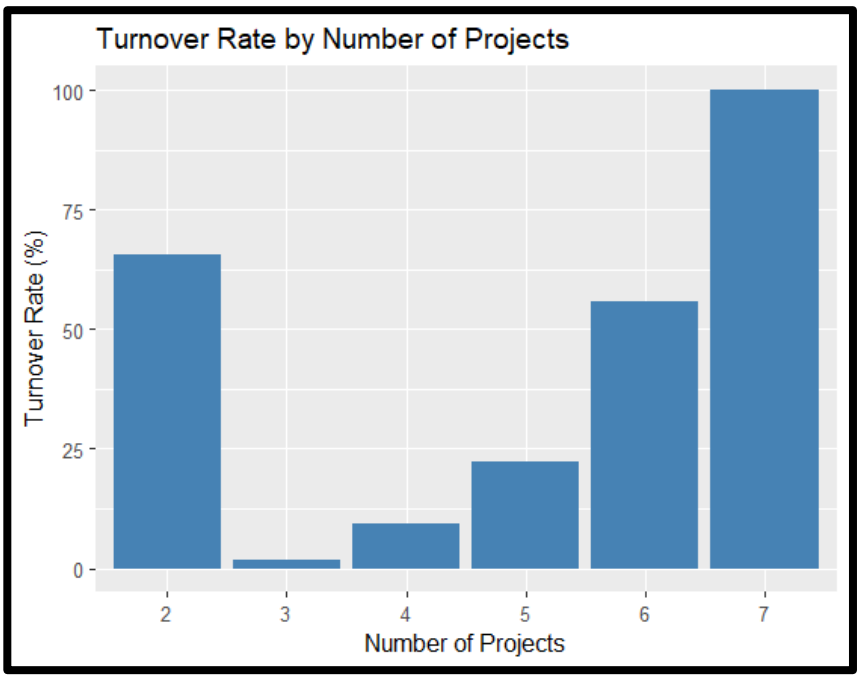

Figure 6.18: Turnover Rate by Number of Projects

Figure 6.18 shows a direct relationship between the number of projects assigned and turnover rates for employees managing between 3 and 7 projects. On the other hand, employees who were assigned the minimal number of projects also have a considerably high turnover ratio. This suggests that employees who handle only 2 projects likely feel undervalued and disengaged. In contrast, those who handle 6 and 7 projects, likely feel overworked.

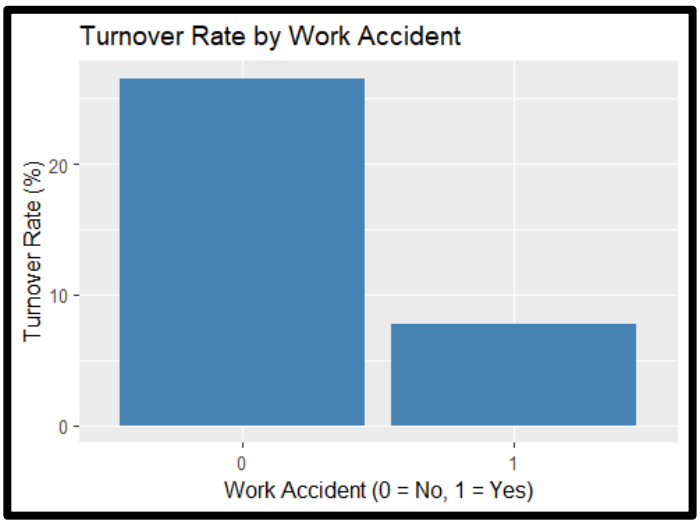

Figure 6.19: Turnover Rate by Work Accident

Figure 6.19 shows that workers that have had accidents have a much lower turnover ratio than those who haven't. Based on that, work accident may not be the main reason why employees leave.

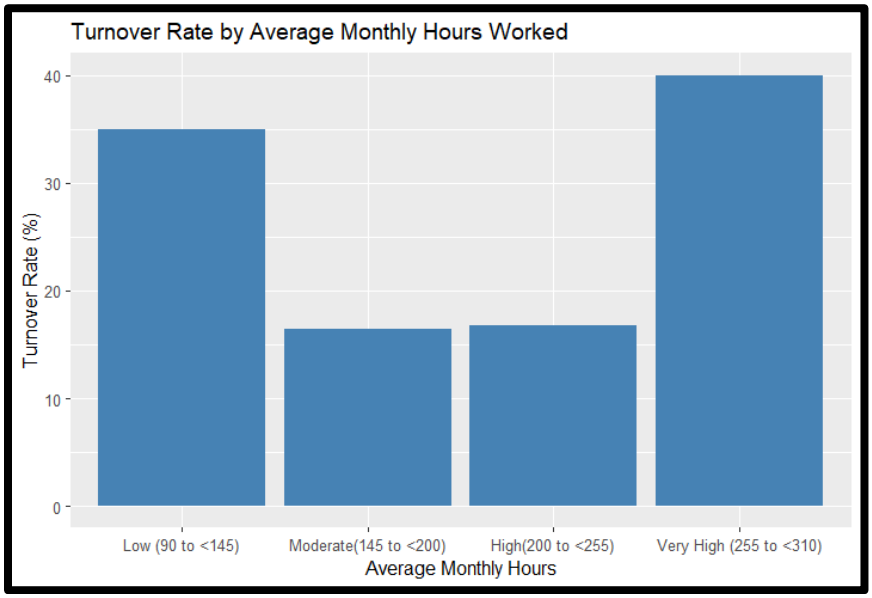

Figure 6.20: Turnover Rate by Average Monthly Hours Worked

The bar chart ***(Figure 6.20)*** provides a straightforward summary of turnover rates categorized by average monthly hours worked. It highlights that turnover rates are significantly higher for employees working either low hours (90–145) or very high hours (255–310). For these groups, the turnover rates exceed 30% and even approach 40% for very high hours, suggesting

that both underutilization and overwork play critical roles in employee resignations. In contrast, the moderate (145–200) and high (200–255) ranges show lower average turnover rates, around 17%, implying that these workloads are generally more sustainable for employees.

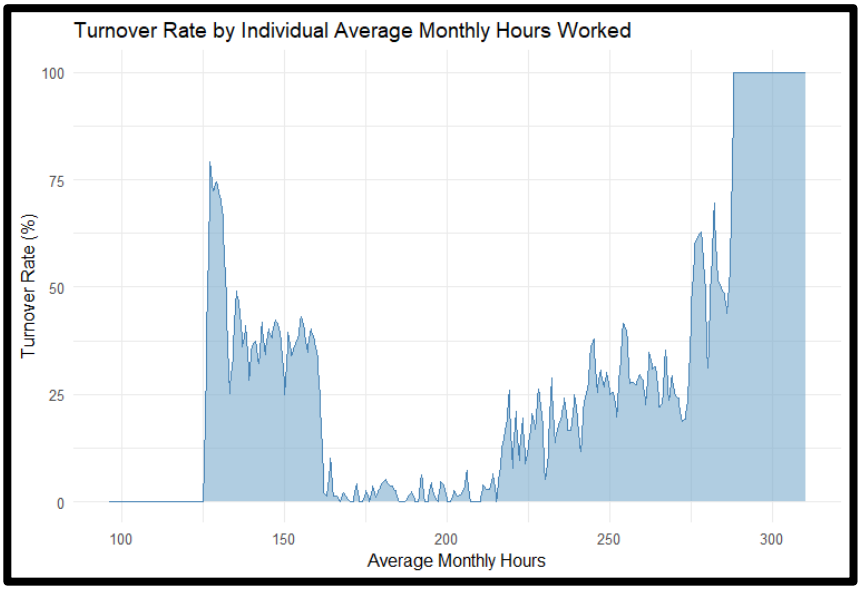

Figure 6.21: Turnover Rate by Individual Average Monthly Hours Worked

The area chart in ***Figure 6.21*** provides a more detailed summary of turnover rates by workload. It shows the highest turnover rates in the 288 to 310 hours range, where turnover reaches 100%, and in the 126 to 161 hours range, with an average turnover of 42.1%, indicating dissatisfaction among underutilized and overworked employees. In contrast, the 162 to 216 hours range stands out with the lowest turnover rate for full time employees, averaging only 2.11%, suggesting this as an optimal workload range for retention.

Beyond 217 hours, turnover rates begin to rise significantly, fluctuating between 5% and 69% in the 217–287 hour range, with an average of 28.8%, reflecting both instability and higher-than-average turnover. The 96 to 125 hours range, with a 0% turnover rate, indicates strong retention, likely driven by lower workload stress and better alignment between employee expectations and actual workloads.

These insights emphasize the importance of maintaining workloads within the 162 to 216 hours range while addressing the high turnover seen in the underutilized 126 to 161 range and the extreme workloads beyond 288 hours.

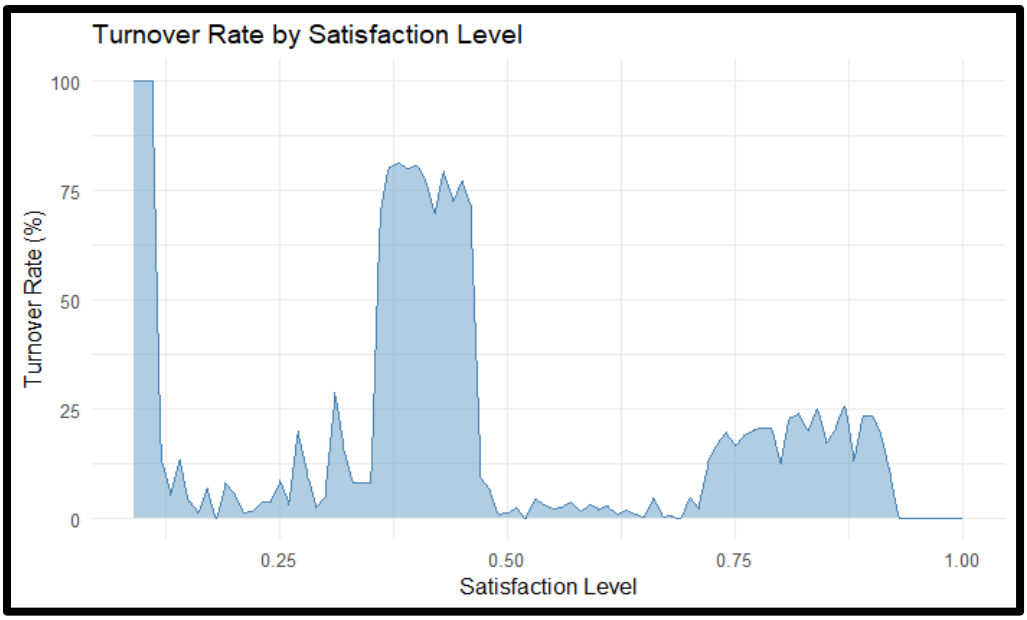

Figure 6.22: Turnover Rate by Satisfaction Level

The turnover rate by satisfaction level area chart ***(Figure 6.22)*** reveals stark differences in employee retention. Turnover reaches 100% among the most dissatisfied employees, highlighting a critical retention issue. Moderately dissatisfied employees also exhibit a high turnover rate, exceeding 70%. In contrast, turnover rates for generally satisfied employees' range between 12.5% and 25%, indicating significantly better retention in this group but still not considered good, indicating that even among those who are somewhat content, there is a notable risk of turnover. This suggests that while satisfaction improves retention, it may not be sufficient on its own to ensure strong employee loyalty. Addressing other factors beyond satisfaction might be necessary to further reduce turnover.

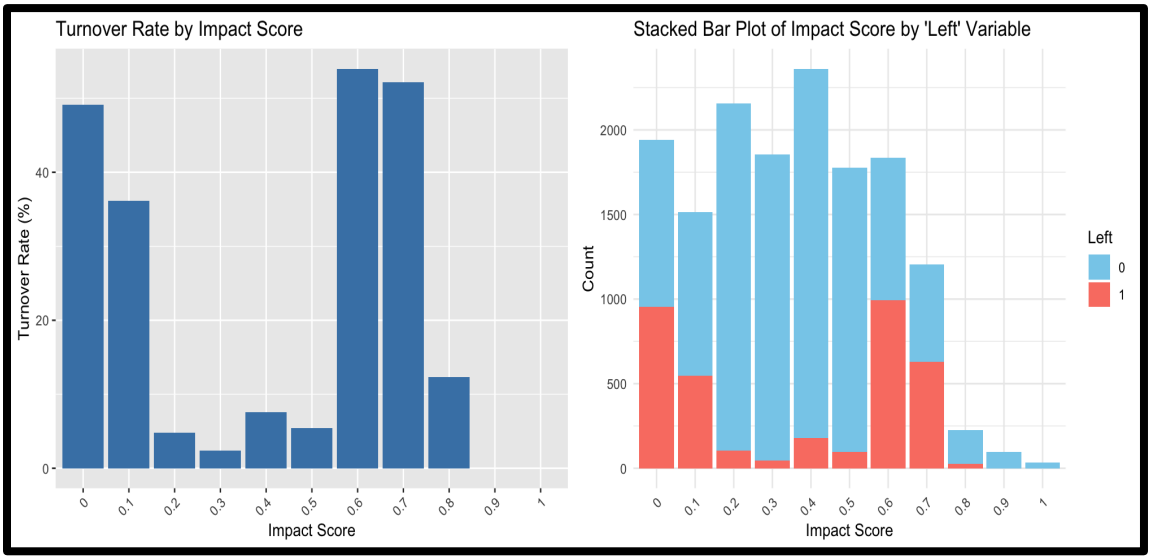

Figure 6.23: Turnover Rates and Resignation Counts by Impact Score

In ***Figure 6.23*** the plot titled "Turnover Rate by Impact Score" represents turnover rates across different impact scores attributed to employees. From this plot, it is evident that employees with the lowest impact scores tend to have significantly higher turnover rates. Additionally, employees considered to have moderately high importance (with scores of 0.6 and 0.7) show the highest turnover rates.

The "Stacked Bar Plot of Impact by 'Left' Variable" displays the data in a similar manner but with a slightly different perspective. This stacked bar plot aggregates the data by stacking the proportions of employees who left (indicated by the 'left' variable) across different impact scores. It visually emphasizes the proportion of employees who left relative to their impact scores, making it easier to compare the distribution of turnover across various levels of importance.

As previously observed in ***Figure 6.11***, employees who have been with the company for 4 to 6 years, particularly around year 5, experience higher-than-average turnover rates. Similarly, those with low salaries tend to have the highest turnover rates ***(Figure 6.9)***. To explore this further, a stacked bar plot was created to illustrate how salary distribution varies with tenure.

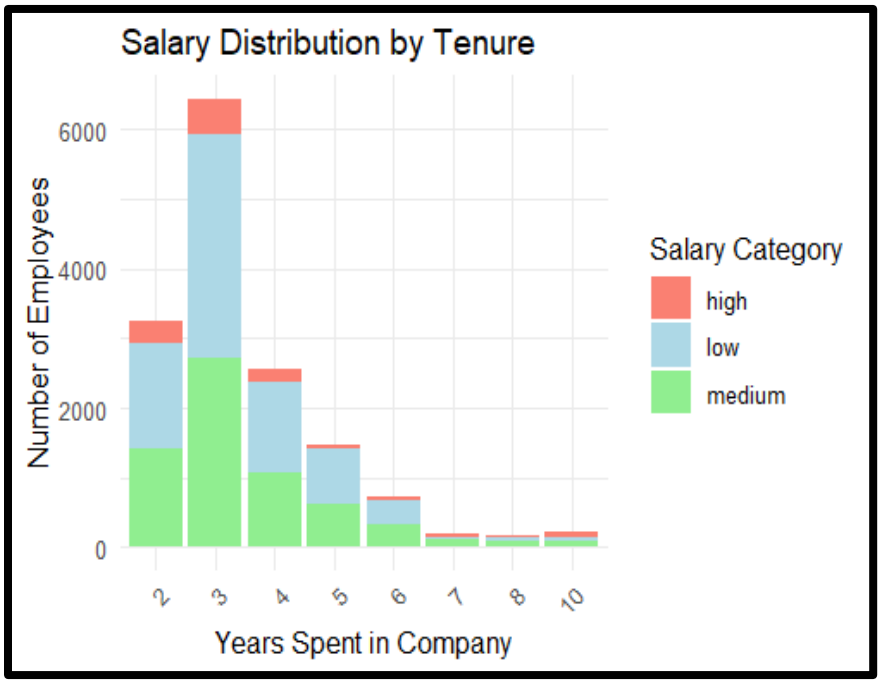

Figure 6.24: Salary Distribution by Tenure

The plot ***(Figure 6.24)*** shows that for employees with tenures of 7, 8, and 10 years (where turnover is 0%) the proportion of low-salary employees is smaller compared to other years. This suggests that, in these periods, low-salary employees are not predominant.

In contrast, for tenures ranging from 6 years down to 2 years, the percentage of low-salary employees increases, making them the majority in these groups. For tenures of 4 to 6 years, where turnover rates are highest, the percentage of high-salary employees decreases compared to those with shorter tenures. Specifically, in years 4 and 5, which see the highest

turnover rates, the proportion of low-salary employees is at its peak. For example, 54% of employees with 5 years of tenure earn low salaries.

These observations suggest a significant link between low salaries in certain years and high turnover rates. The elevated turnover rates in years 4 and 5 may be largely influenced by the high percentage of low-salary employees during these periods.

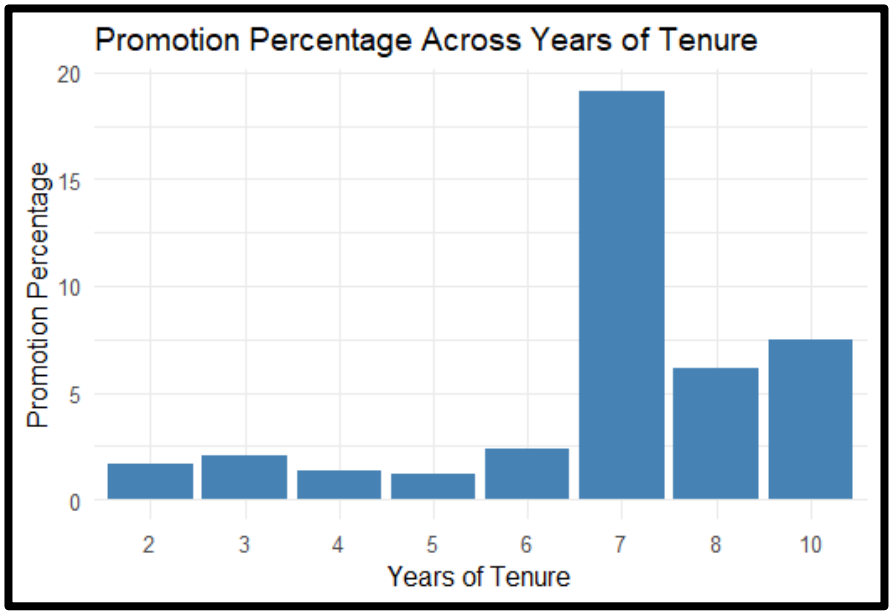

Figure 6.25: Promotion Percentage Across Years of Tenure

Based on the plot in ***Figure 6.25*** it is also clear that employees with tenures of 4 and 5 years have the lowest promotion rates, with figures of 1.37% and 1.15%, respectively. This low promotion rate likely contributes to their high percentage of low salaries within the company. Consequently, this factor could be a significant reason for the elevated turnover rates observed among employees in these tenure groups. In contrast, employees with a tenure of 7 years or more have the highest promotion rates. Interestingly, these groups experience zero turnover.

As it was found earlier, a higher percentage of employees leave between 3 and 6 years, with the highest turnover at year 5. Promotion rates are also the lowest between years 4 and 5, while they are the highest after year 7. Therefore, by reviewing variables like satisfaction levels, last evaluations, number of projects, and average monthly hours across different tenure groups, it is possible to draw meaningful conclusions.

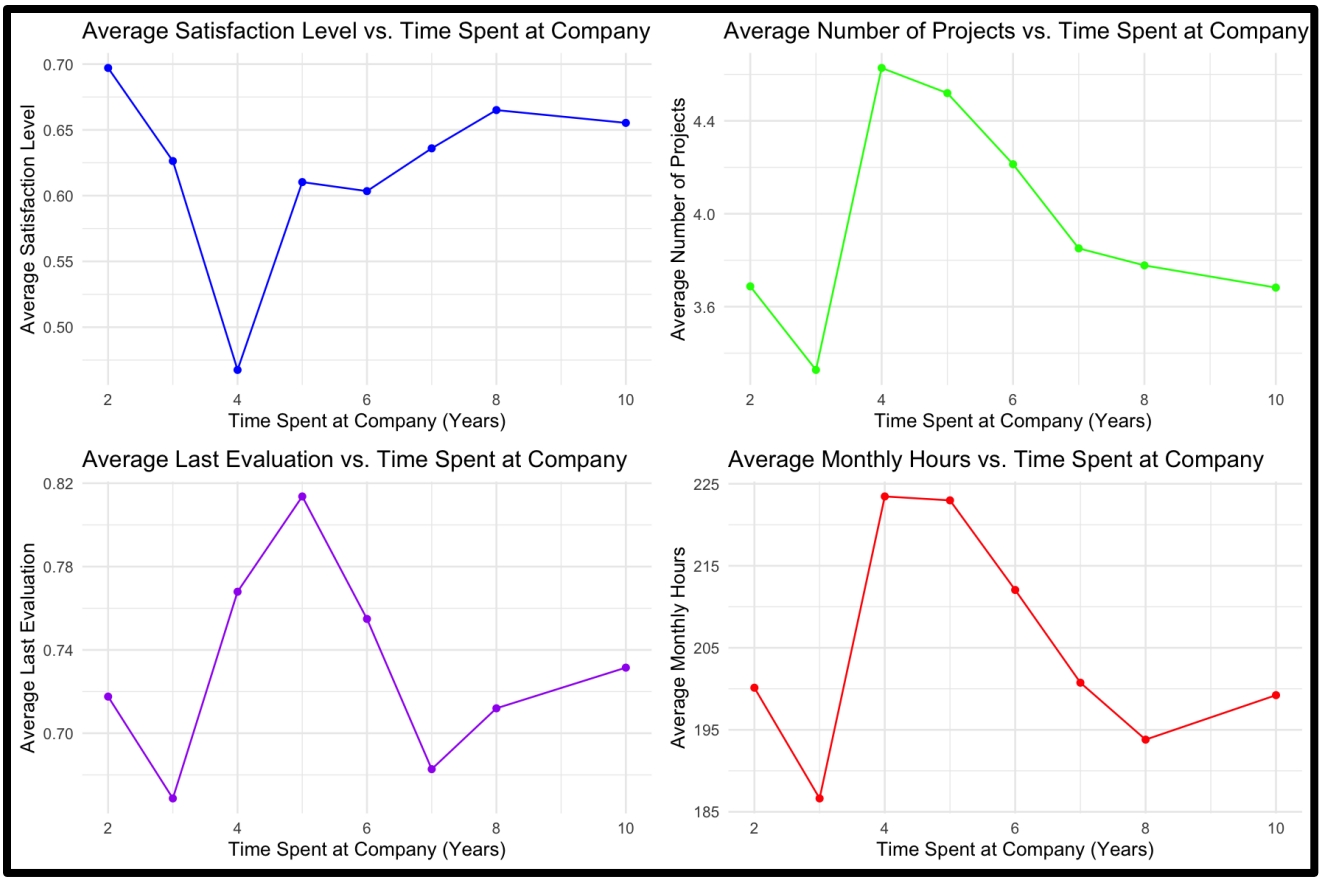

Figure 6.26: Trends in Employee Metrics Over Time Spent at the Company

Figure 6.26 shows that employees generally start with high satisfaction levels, but this begins to decline, reaching its lowest at year 4, where the worst satisfaction levels across the company are observed. Although satisfaction improves slightly in years 5 and 6, it remains low compared to other groups. During years 4 to 6, employees also have the highest evaluation scores and take on the most projects, with both peaking in year 4, along with the highest number of hours worked. These findings suggest that the workload might be too high and not well recognized, which, again, explains a possible reason for turnover.

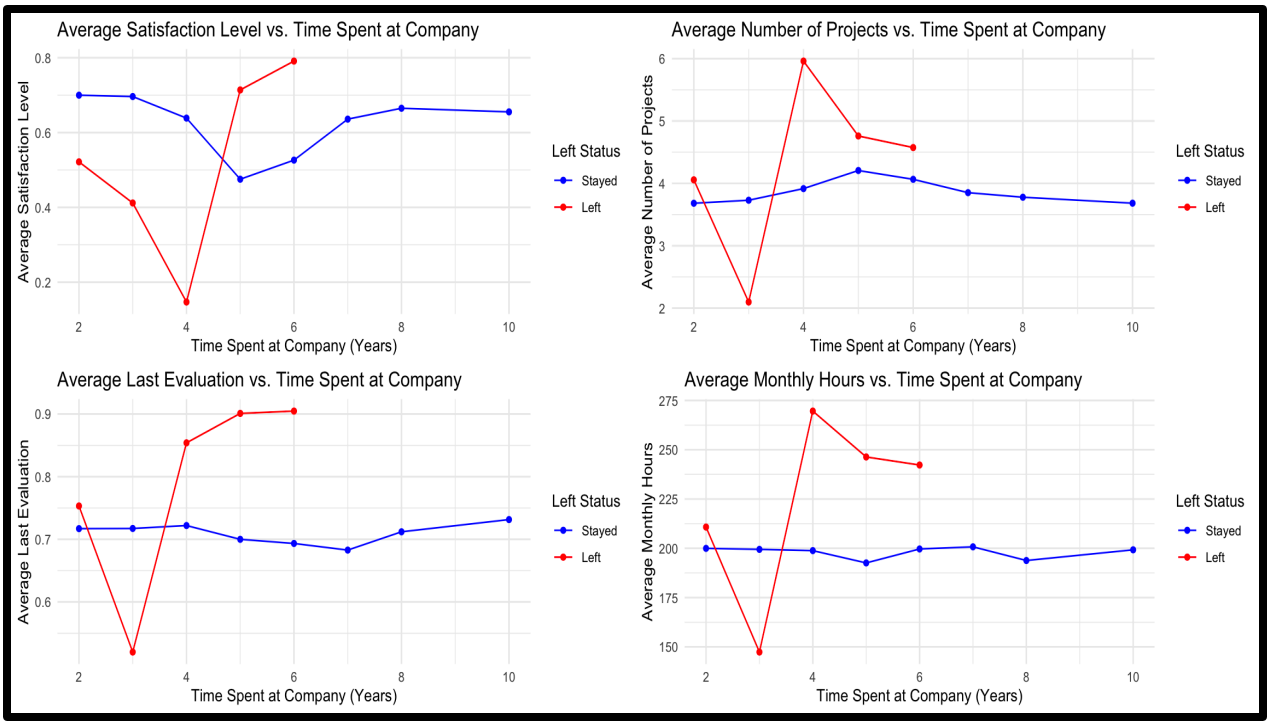

Figure 6.27: Trends in Employee Metrics Over Time Spent at the Company

To better understand and compare these trends, the data was split into two groups ***(Figure 6.27)***: employees who stayed and those who left. Employees who stayed generally showed steady and similar values across all years, suggesting they experienced consistent work conditions and satisfaction. However, employees who left had more varied and often extreme values, indicating either too much work or too little involvement.

In detail, employees in their fourth year felt the most unhappy and overworked but still performed well. Those in their third year were somewhat unhappy and scored the lowest in terms of performance and involvement in projects, suggesting they were less engaged at work.

Interestingly, despite year five showing the highest number of employees leaving, these individuals were satisfied. There reasons for leaving could be due to lower pay and fewer chances for promotion, leading them to look for better opportunities elsewhere.

6.3 Exploratory Data Analysis Summary

The exploratory data analysis reveals potential key drivers of employee turnover. Dissatisfaction is one of these factors, with lower satisfaction levels strongly linked to higher turnover. However, even satisfied employees leave, indicating that other factors also play a role. Compensation also has a significant influence, as employees with low salaries show the highest turnover, while competitive pay is associated with better retention.

Turnover is more pronounced during the mid-tenure period (4–6 years), peaking in year 5. This period is characterized by limited promotions, increased workloads, and lower satisfaction, suggesting a lack of career growth and recognition as contributing factors. In

addition, extreme work hours and number of projects are associated with higher attrition, while moderate workloads support better retention.

Performance evaluations provide additional insights. While low performers tend to leave likely due to underperformance, many high performers also resign, possibly in search of better compensation or growth opportunities. Departmental variations in turnover further highlight specific retention challenges and the need for interventions, while work accidents appear to have minimal impact on resignations.

In summary, the analysis shows the importance of addressing satisfaction, recognition, workload balance, and compensation to reduce turnover.

Section 7: Hypothesis Testing

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

The insights obtained during the data exploration phase show that the lower the salary, the lower the turnover rate is. Additionally, it was learned that employees with low salaries that have been working in the company for between 3 to 6 years (especially 5 years) have the highest turnover ratios in the company.

To the confirm the hypothesis that salary and resignations are statistically associated, a Chi-Squared test was performed.

```
        Pearson's Chi-squared test

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

Figure 7.1: Chi-Squared Test on Salary and Turnover

The results of the test show an extremely small p-value, suggesting a statistically significant association between salary levels and employee turnover. Also, the high chi-squared value indicates that observed frequencies (actual counts in the data) differ significantly from the expected frequencies (the counts to expect if there were no association between the variables). This means that salary is an important factor in determining whether employees stay or leave.

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

As observed during the initial data exploration, the turnover rate among employees who experienced a work accident is significantly higher than among those who did not. To confirm this observation, a Chi-Squared test was conducted to examine the association between work accidents and employee resignations.

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

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

Figure 7.2: Chi-Squared Test on Work Accidents and Turnover

The test results indicate a strong statistical association between these variables. However, while the Chi-Square test confirms the presence of a relationship, it does not provide specific insights into the direction or magnitude of this association. Contrary to the initial hypothesis, the significantly lower turnover rate among employees who did not suffer accidents suggests that the hypothesis should be reconsidered.

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

Based on the exploratory data analysis the company is not the best place to grow professionally, since there is very low promotion rates and high performers who have been in the company for many years have not been supported in terms of promotions.

Section 8: Predictor Analysis and Relevancy

The goal of this phase is to identify the key predictors influencing the target variable. By evaluating the relevance of each available variable, the most impactful ones are selected to enhance the model's performance. This process allows for developing an approach that is both effective and aligned with stakeholder expectations, setting a solid foundation for the subsequent stages of the project.

Category	Variable Names
Target Variable	left
Available Predictors	satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, Work_accident, promotion_last_5years, department, salary, impact_score.

Table 8.1: Target Variable and Available Predictors

8.1 Correlation Analysis

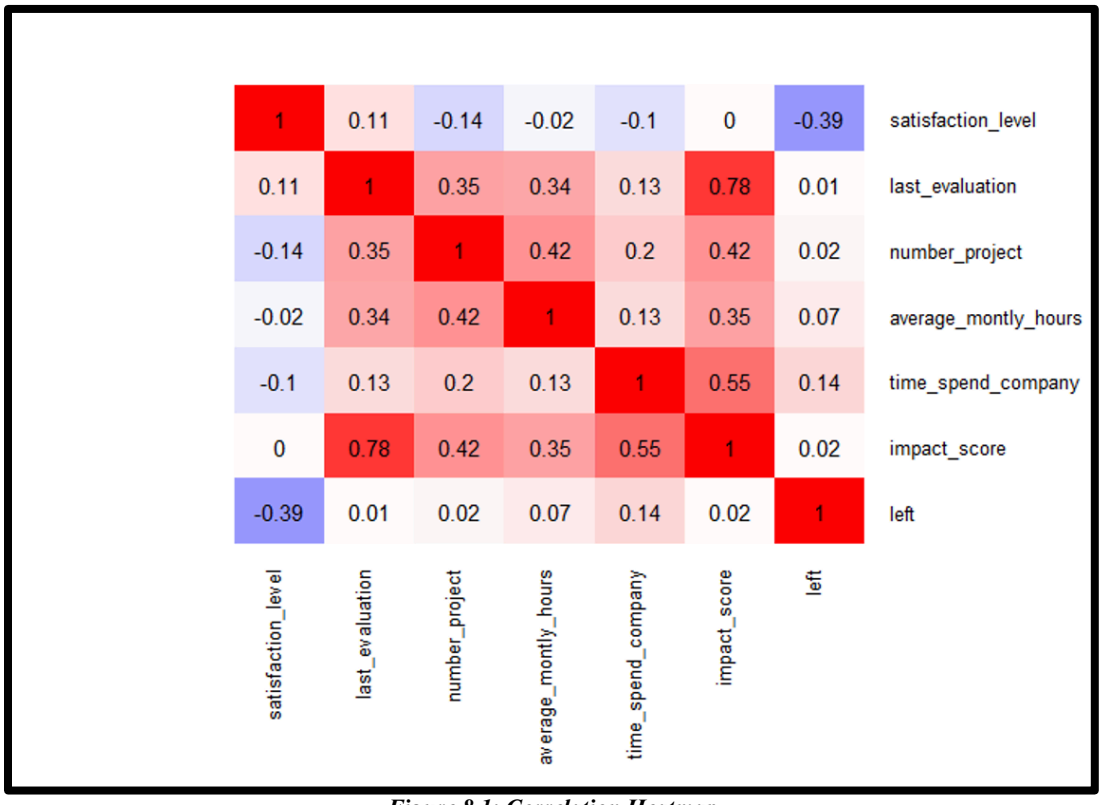

Figure 8.1: Correlation Heatmap

8.1.1 Correlations Between Target Variable 'left' and Predictors

Strong Correlations (|r| > 0.5):

- None

Moderate Correlations (0.3 < |r| ≤ 0.5):

- Satisfaction Level: -0.388 (Moderate negative correlation)

Weak Correlations (|r| ≤ 0.3):

- Time Spent at the Company: 0.145 (Weak positive correlation)
- Average Monthly Hours: 0.071 (Weak positive correlation)
- Number of Projects: 0.024 (Weak positive correlation)
- Impact Score: 0.016 (Weak positive correlation)
- Last Evaluation: 0.007 (Very weak positive correlation)

Satisfaction level has the most significant relationship with the outcome, showing a moderate negative correlation. This suggests that lower satisfaction levels may be meaningfully linked to the dependent variable. Other variables, such as **time spent at the company**, **average monthly hours**, **number of projects**, **impact score**, and **last evaluation**, show weak or negligible correlations, indicating they have little to no meaningful impact on the outcome.

8.1.2 Correlations Between Predictors

Strong Correlations (|r| > 0.5):

- Impact Score and Last Evaluation: 0.778 (Strong positive correlation)
- Impact Score and Time Spent at the Company: 0.549 (Strong positive correlation)

Moderate Correlations (0.3 < |r| ≤ 0.5):

- Number of Projects and Average Monthly Hours: 0.417 (Moderate positive correlation)
- Impact Score and Number of Projects: 0.420 (Moderate positive correlation)
- Impact Score and Average Monthly Hours: 0.353 (Moderate positive correlation)
- Last Evaluation and Number of Projects: 0.349 (Moderate positive correlation)
- Last Evaluation and Average Monthly Hours: 0.340 (Moderate positive correlation)

Weak Correlations (|r| ≤ 0.3):

- Time Spent at the Company and Number of Projects: 0.197 (Weak positive correlation)
- Time Spent at the Company and Average Monthly Hours: 0.128 (Weak positive correlation)
- Satisfaction Level and Last Evaluation: 0.105 (Weak positive correlation)
- Satisfaction Level and Time Spent at the Company: -0.101 (Weak negative correlation)
- Satisfaction Level and Number of Projects: -0.143 (Weak negative correlation)
- Satisfaction Level and Average Monthly Hours: -0.020 (Very weak negative correlation)

8.1.3 Handling Redundancy

A key concern in data analytics is **multicollinearity**, where features are highly correlated with one another. Multicollinearity can inflate the variance of model coefficients, making it difficult to interpret the model and reducing its generalization performance. Dimensionality reduction techniques address both the curse of dimensionality and multicollinearity by simplifying the feature space and ensuring that features are more independent.

In this dataset, the derived variable impact_score was found to be highly correlated with the variables used to create it. Including such a variable could introduce multicollinearity and redundancy, potentially biasing the model. To mitigate this, impact_score will be removed. This step ensures that the dataset remains clean and avoids potential overrepresentation of correlated information.

8.1.4 Chi-Square Test for Categorical Variables

The Chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. It compares the observed frequencies in each category to the frequencies expected if there were no association between the variables.

Categorical/Binary Variables	X-squared	p-value
work_accident	357.56	< 2.2e-16
promotion_last5years	56.262	6.344e-14
department	86.825	7.042e-15
salary	381.23	< 2.2e-16

Table 8.2: Chi-Squared Test on Categorical Variables and Predictor

Work Accident:

Employees who have had work accidents are significantly more or less likely to leave.

Promotions:

The likelihood of leaving is significantly related to whether an employee has been promoted in the last 5 years.

Department:

Different departments have a significant effect on the likelihood of leaving.

Salary:

The level of salary has a strong association with whether employees leave.

All the categorical/binary variables tested show significant associations with the target variable, as indicated by the very small p-values. This suggests that these variables are important factors in the model and could be useful predictors for whether an employee has left the company.

8.2 Dimensionality Reduction

Dimensionality reduction is an essential step in preparing data for modeling, particularly when dealing with datasets that have a high number of features relative to observations. High-dimensional datasets can lead to the **curse of dimensionality**, where the sparsity of data in high-dimensional spaces diminishes the effectiveness of many machine learning algorithms. This issue can result in:

- Increased computational complexity.
- Overfitting due to excessive model flexibility.
- Difficulty in identifying meaningful patterns.

Reducing the dimensionality of a dataset helps mitigate these issues by simplifying the feature space while retaining as much relevant information as possible. A common technique for dimensionality reduction is **Principal Component Analysis (PCA)**, which transforms the original variables into a smaller set of uncorrelated components called principal components. These components are linear combinations of the original variables and are ordered to capture the maximum variance in the data, allowing for efficient representation while reducing redundancy.

8.2.1 Dimensionality Reduction in This Dataset

The dataset contains 10 variables and 14,999 records, providing a favorable ratio of observations to features. After addressing multicollinearity by removing impact_score, the remaining features are sufficiently independent, and the feature space is already manageable. Therefore, advanced dimensionality reduction techniques, such as Principal Component Analysis (PCA), are not necessary in this case.

Later the focus will shift toward partitioning the dataset and applying feature selection during model training to identify the most predictive variables and optimize performance.

Section 9: K-Means Clustering

In this section, K-means clustering is applied to understand the different profiles of employees who leave the company. Identifying these profiles is crucial for uncovering patterns in employee turnover, enabling the organization to implement targeted strategies to improve retention and address systemic issues.

9.1 K-Means Algorithm

K-means clustering is an unsupervised learning algorithm that groups data points into distinct clusters based on their similarity. The algorithm iteratively assigns each data point to the nearest cluster center and updates the centers to minimize within-cluster variance. This process continues until stable clusters are formed.

9.1.1 Initialization

The algorithm begins by selecting the number of clusters (k), which defines how many groups the data will be divided into. Centroids, for each cluster are then initialized randomly within the dataset.

9.1.2 Assigning Data Points to Clusters

Each data point is assigned to the cluster with the nearest centroid, based on Euclidean Distance. This ensures that data points within a cluster share similarities, while points in different clusters remain distinct.

9.1.3 Updating Cluster Centroids

After assigning all data points, the centroids of each cluster are recalculated as the average position of all points within that cluster. This step adjusts the clusters to better represent the grouped data.

9.1.4 Iteration

The assignment and centroid recalculation steps are repeated until the clusters stabilize. Stabilization occurs when data points stop switching between clusters or when the changes in centroids are minimal.

9.1.5 Output

The algorithm outputs kk clusters, each defined by a centroid and a set of assigned data points. These clusters reveal meaningful groupings in the data.

9.2 Variable Selection

For this analysis, numeric variables such as satisfaction level, last evaluation score, number of projects, average monthly hours, and time spent at the company were used. These variables were chosen because k-means clustering relies on Euclidean distance to measure similarity between data points, which performs better with numeric variables. While it is possible to adapt k-means for categorical data, the decision to use only numeric variables was made to ensure better clustering performance. Additionally, the numeric variables were standardized to ensure equal weighting and prevent variables with larger scales from disproportionately influencing the clustering results.

9.3 Determining the Value of K

Selecting the right number of clusters (k) is essential for meaningful K-means clustering. In this analysis, the **Elbow Method** was used to determine the optimal k by evaluating the point where adding more clusters no longer significantly reduces within-cluster variance.

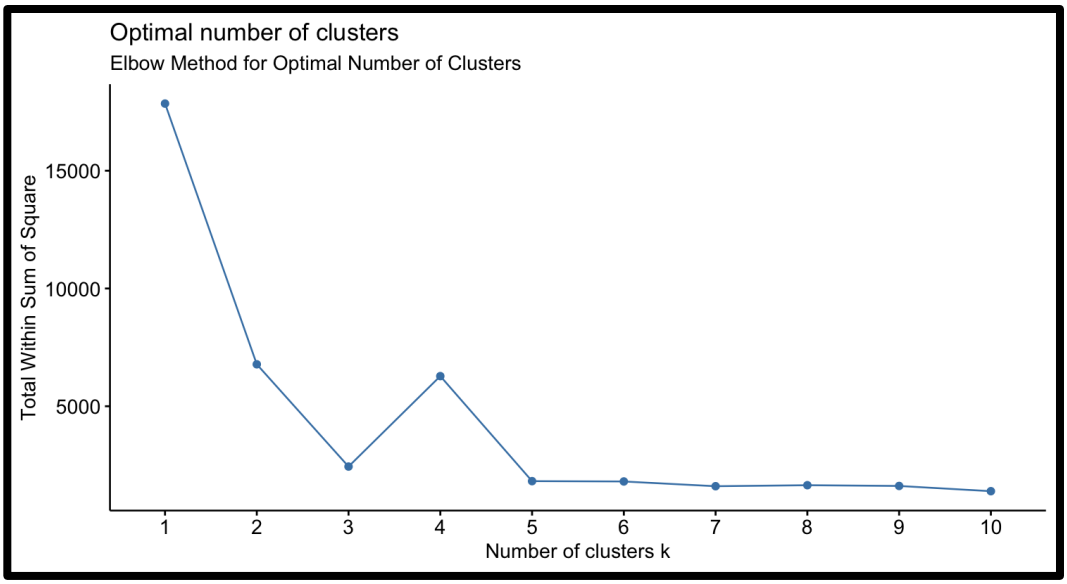

Figure 9.1: Elbow Method for Optimal Number of Clusters

The Elbow Method plot ***(Figure 9.1)*** shows a significant drop in within-cluster variance up to k=3, after which the improvement becomes minimal. This "elbow" at k=3 indicates that three clusters are sufficient to capture the key patterns in the data without overcomplicating the model. Thus, k=3 was chosen as the optimal number of clusters.

9.4 Performance Evaluation

The silhouette plot evaluates the quality of clustering by analyzing how well each data point fits within its assigned cluster compared to other clusters. It provides a visual and numerical way to assess the cohesiveness and separation of clusters.

9.4.1 Key Concepts

1. **Cohesion (ai)**:
    - The average distance between a point and all other points in the same cluster. A lower value indicates the point is tightly grouped within its cluster.
2. **Separation (bi)**:
    - The average distance between a point and points in the nearest cluster. A higher value indicates the point is far from other clusters.

9.4.2 Silhouette Score Formula

The silhouette score for a point is calculated by:

$$si = \frac{bi - ai}{\max(ai, bi)}$$

This standardizes the score between −1 and 1.

9.4.3 Interpretation of the Silhouette Plot

- **Values close to 1**: Points are well-matched to their own cluster and poorly matched to others.
- **Values close to 0**: Points are near the boundary between clusters.
- **Negative values**: Points may have been misclassified, fitting better in a different cluster.

9.4.4 Silhouette Analysis for Employee Turnover Clusters

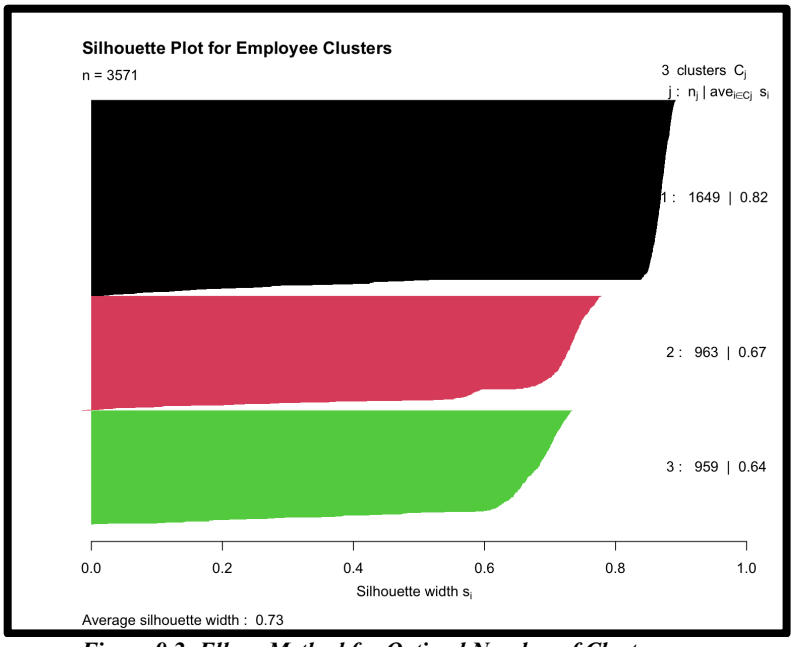

Figure 9.2: Elbow Method for Optimal Number of Clusters

The overall average silhouette width is **0.73**, indicating a strong clustering structure where most points are well-matched to their respective clusters and separated from others.

Cluster 1:

Cluster 1 stands out with the highest silhouette width of **0.82**, representing 1,649 employees. This suggests that the data points in this cluster are highly cohesive and well-separated from the other clusters. Employees in this group likely share very distinct characteristics, making this the most well-defined cluster.

Cluster 2:

Cluster 2 has a silhouette width of **0.67**, with 963 employees. While the clustering is still strong, it indicates slightly less separation from other clusters compared to Cluster 1. Some points in this cluster may lie closer to the boundary with neighboring clusters, but overall, the cluster is still distinct.

Cluster 3:

Cluster 3 has the lowest silhouette width of **0.64**, with 959 employees. Although this score is lower than the other clusters, it still reflects strong clustering quality.

9.5 Cluster Profile Analysis

The analysis focus on understanding the three distinct groups of employees who left the company, each with unique patterns of satisfaction, performance, workload, and tenure. These clusters show different reasons for turnover and provide insights into employee behavior.

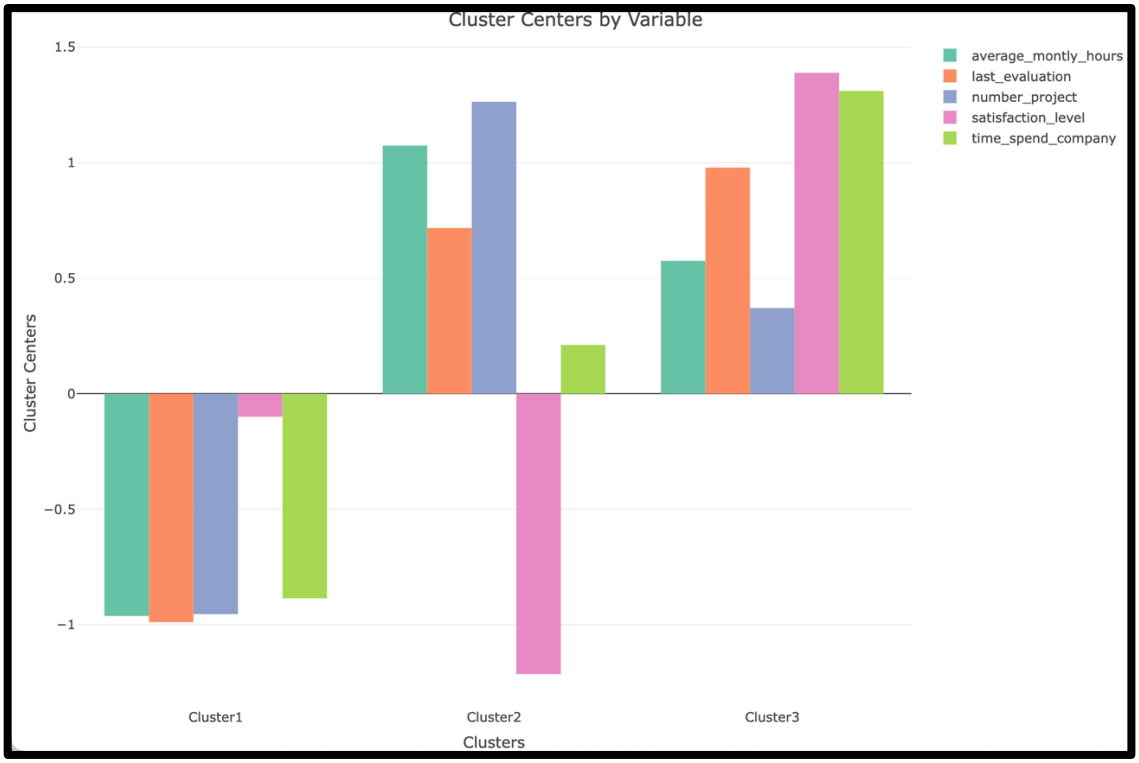

Figure 9.3: Profiles of Employees Who Left the Company

Cluster 1: Disengaged Employees

This group consists of employees who appear to have had minimal engagement during their time at the company. Their performance evaluations were notably low, suggesting difficulties in effectively fulfilling their roles. These employees also handled fewer projects and worked fewer hours than others, indicating a lack of involvement or underutilization. They typically left early in their tenure, hinting that they lacked support, motivation, and their experience at the company failed to meet their expectations.

Cluster 2: Overworked Achievers

Employees in this group stand out for their high performance and heavy workloads. Despite receiving above-average evaluations and handling the highest number of projects, their satisfaction levels are significantly low, suggesting frustration or burnout. These employees worked considerably more hours than others, and their tenure was moderate, indicating that while they contributed significantly, they left after enduring high stress and unmanageable workloads.

Cluster 3: Stagnated Employees

This cluster represents employees who stayed at the company the longest before leaving. They were highly satisfied and strong performers, indicating engagement and value as contributors. Their consistent workload suggests stability rather than burnout. However, despite their loyalty, they likely left in search of better opportunities for career growth and advancement that were not available within the company.

Section 10: The Impact of Salary and Promotions on Employee Turnover Profiles

In this section, the findings from the exploratory data analysis (EDA) and the clustering analysis are brought together to provide a comprehensive understanding of employee turnover. While the EDA identified salary and lack of promotions as key drivers of turnover, the clustering analysis revealed distinct profiles of employees who left the company, each with unique characteristics and reasons for leaving.

The approach involves examining how salary and promotions influence turnover across the clusters. First, their distributions will be analyzed within each cluster to determine whether these factors are consistent across all groups or specific to certain profiles. Then, their role in amplifying dissatisfaction or contributing to the unique turnover reasons identified for each cluster will be discussed.

10.1 Reviewing EDA Findings

It was previously observed that employees with lower salaries and fewer promotions experience significantly higher turnover rates. Turnover rates are 29.7% for low salaries, compared to 20.4% for medium salaries and 6.63% for high salaries. Similarly, employees without promotions have a turnover rate of 24.2%, compared to 5.96% for those who were promoted.

Additionally, employees who leave the company are disproportionately concentrated in the low salary group (60.8% of those who left versus 45% of those who stayed) and among those with fewer promotions (0.53% of those who left compared to 2.62% of those who stayed). These trends highlight the significant role that salary and promotion opportunities play in driving turnover.

Salary	Turnover Rate (%)
High	6.63
Medium	20.4
Low	29.7

Table 10.1: Turnover Rate Across Salary Categories

Salary	Proportion Stayed (%)	Proportion Left (%)
High	10.11	2.30
Medium	44.88	36.88
Low	45.01	60.82

Table 10.2: Salary Category Distribution for Employees Who Left vs. Stayed

Promotion in Last 5 Years	Turnover Rate (%)
No (0)	24.2
Yes (1)	5.96

Table 10.3: Turnover Rate by Promotions

Employee Status	Promotion Proportion (%)
All	2.13
Stayed	2.63
Left	0.53

Table 10.4: Promotion Proportion by Employee Status

These findings suggest that salary and promotion opportunities play a significant role in employee turnover. However, their influence may vary across the different profiles of employees who left, as identified through clustering analysis. The next section will explore how these factors relate to the unique characteristics of each cluster.

10.2 Distribution of Salary and Promotions Across Clusters

Cluster	High Salary (%)	Medium Salary (%)	Low Salary (%)
1	2.85	35.90	61.25
2	1.77	38.21	60.02
3	1.88	37.23	60.90

Table 10.5: Salary Distribution Across Clusters

As shown in the ***Table 10.5***, employees across all clusters are predominantly in the low-salary category, with proportions ranging from 60.02% to 61.25%. Medium-salary employees are the second-largest group, ranging from 35.90% to 38.21%. High-salary employees are a small minority, with no cluster exceeding 2.85%.

These distributions indicate that low salary is a consistent issue across all clusters, suggesting it is a significant factor contributing to turnover, regardless of the specific characteristics of each group.

Cluster	Not Promoted (0) (%)	Promoted (1) (%)
1	99.09	0.91
2	99.69	0.31
3	99.90	0.10

Table 10.6: Promotion Distribution Across Clusters

The analysis of promotion status ***(Table 10.6)*** emphasizes the lack of upward mobility for employees across all clusters. The vast majority of employees in each cluster were not promoted in the last five years.

Employees who were promoted represent less than 1% in all clusters, with the lowest proportion in Cluster 3 (0.10%). This stark disparity suggests that limited promotion opportunities are a systemic challenge that affects all groups of employees who left the company.

10.3 Conclusion

Together, these findings reveal that salary and promotions are not isolated issues but broader, foundational challenges that affect all employees who leave the organization. These systemic problems amplify the specific pain points identified within each cluster, contributing to turnover on a larger scale. While these broader issues form a "big picture" concern, employees in each profile react to them differently based on their unique circumstances, ultimately driving their decision to leave. These challenges are deeply rooted in organizational practices rather than the specific factors highlighted within individual clusters.

For disengaged employees (Cluster 1), low salaries and the lack of promotions intensify their feelings of being undervalued and disconnected, increasing their motivation to leave. For

overworked achievers (Cluster 2), the absence of promotions or better pay leaves these hardworking employees feeling unappreciated, further diminishing their satisfaction and driving them to seek opportunities elsewhere. Lastly, stagnated employees, despite being satisfied and having many years of experience in the company, eventually leave due to stagnant pay and limited growth opportunities, prompting them to pursue better career prospects.

Addressing these systematic issues, along with meeting the specific needs of each cluster, is crucial to lowering turnover and improving the work environment.

Section 11: Data Partitioning

Data partitioning is an essential part of a predictive model development process. Performing an appropriate data partition is crucial for ensuring that the model learns effectively and is able to generalize well to unseen data. In this phase of the project, the dataset will be divided into three different subsets: training, validation, and testing sets. Each of them serves a specific purpose and plays a critical role in the creation and deployment of a reliable model.

11.1 Training Set

The training set is the foundation of model development. This subset of the data is used for training the model; it is where underlying patterns and relationships between predictors and target variable are learned. Through exposure to a variety of records, the model's parameters are adjusted to improve its predictive accuracy.

Purpose: To train the model and enable it to learn patterns from the data.

Proportion: 70% of the dataset (10,499 records).

11.2 Validation Set

The validation set is a crucial component in the model development process, used to evaluate a model’s performance after training. This subset, consisting of unseen data, enables the comparison of different models and the fine-tuning of hyperparameters. By assessing the model on the validation set, it is possible to select the most appropriate model for the given task, ensuring it performs well under the intended conditions.

Additionally, the validation set helps in addressing overfitting, where a model learns not only the underlying patterns in the training data but also the noise and random fluctuations, which leads to exceptional performance in the training data but poor performance on new data.

Purpose: To evaluate model performance, fine-tune hyperparameters, and address overfitting.

Proportion: 20% of the dataset (3,000 records).

11.3 Testing Set

After training, validating, and comparing multiple models, the one that best aligns with the analytical goals is selected for further evaluation. The testing set, reserved exclusively for this purpose, is used to perform the final assessment of the model's performance. This step provides an unbiased estimate of how well the model generalizes to new, unseen data. Such evaluation is critical for determining the model's effectiveness and reliability in real-world scenarios prior to deployment.

Purpose: To provide an unbiased assessment of the final model's performance.

Proportion: 10% of the dataset (1,500 records).

11.4 Random Sampling

To ensure that each subset of the data is representative of the overall dataset, random sampling techniques are employed. This method involves randomly selecting data points for each set, which helps prevent biases and ensures that each subset reflects the diversity of the entire dataset. In some cases, stratified sampling may be used to maintain the same distribution of target classes across the subsets, particularly in imbalanced datasets.

Importance: Random sampling minimizes bias and ensures that the training, validation, and testing sets are representative, contributing to a more reliable assessment of the model's performance.

11.5 Partition Summary

For this project, the dataset was partitioned into the following proportions:

- **Training Set:** 70% (10,499 records)
- **Validation Set:** 20% (3,000 records)
- **Testing Set:** 10% (1,500 records)

By adhering to this structured approach, the resulting model is robust, generalizable, and capable of accurate predictions in real-world scenarios.

Section 12: Feature Selection

12.1 Feature Selection: Boruta

In this analysis, the Boruta feature selection algorithm was applied to identify the most relevant predictors for the target variable 'left'. Boruta, through Random Forests, compares the importance of actual features against randomly shuffled shadow features. This approach helps in determining which features genuinely contribute to the prediction of the target variable.

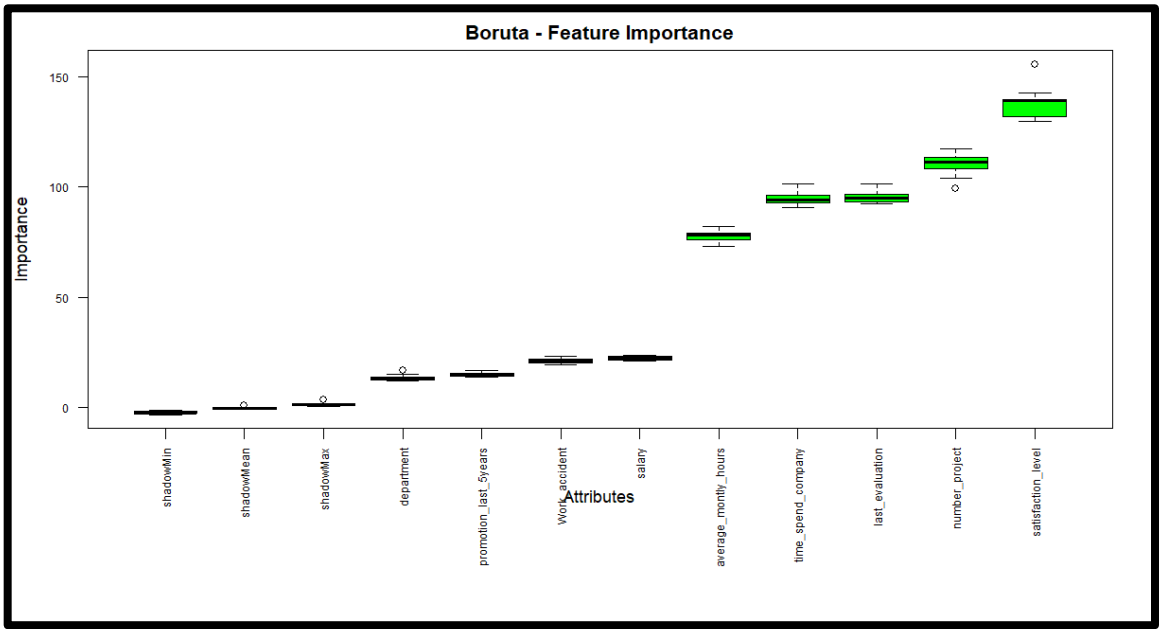

Figure 12.1: Boruta Feature Importance

In the graph ***(Figure 12.1)***, each variable is represented by a boxplot. This boxplot shows the distribution of importance over multiple iterations of the Random Forest model.

```
> print(important_features)
[1] "satisfaction_level"    "last_evaluation"       "number_project"
[4] "average_montly_hours"  "time_spend_company"    "Work_accident"
[7] "promotion_last_5years" "department"            "salary"
```

Figure 12.2: Important Features Selected by Boruta

The Boruta analysis indicated that all features in the dataset are important. This outcome suggests that each variable provides valuable information for predicting whether an employee will leave the company.

12.2 Feature Selection: Stepwise Regression

In this section, the goal is to create a logistic regression model in which the most influential variables are selected to predict the value of the target variable 'left'.

- **Method Used:** Backward elimination, where the model starts with all predictors and iteratively removes the least significant ones.
- **AIC (Akaike Information Criterion):** Measures model quality with a trade-off between goodness of fit and complexity. The lower the AIC, the better the model.

```
> summary(stepwise_model)

Call:
glm(formula = left ~ satisfaction_level + last_evaluation + number_project +
    average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years + department + salary, family = binomial,
    data = train_set)
```

Figure 12.3: Stepwise Logistic Regression Formula

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.3388681  0.2297670  -5.827 5.64e-09 ***
satisfaction_level    -4.1234486  0.1169649 -35.254  < 2e-16 ***
last_evaluation        0.7031485  0.1779951   3.950 7.80e-05 ***
number_project        -0.3178972  0.0253265 -12.552  < 2e-16 ***
average_montly_hours   0.0044971  0.0006129   7.338 2.17e-13 ***
time_spend_company     0.2761742  0.0187723  14.712  < 2e-16 ***
Work_accident1        -1.5365054  0.1070103 -14.358  < 2e-16 ***
promotion_last_5years1 -1.5610938  0.3227940  -4.836 1.32e-06 ***
departmenthr           0.1770270  0.1559702   1.135 0.256373
departmentIT          -0.2143942  0.1439415  -1.489 0.136368
departmentmanagement  -0.4727506  0.1860713  -2.541 0.011063 *
departmentmarketing   -0.1557508  0.1580524  -0.985 0.324409
departmentproduct_mng -0.2788059  0.1553330  -1.795 0.072671 .
departmentRandD       -0.5882774  0.1739140  -3.383 0.000718 ***
departmentsales       -0.0934090  0.1211768  -0.771 0.440797
departmentsupport     -0.0245670  0.1297454  -0.189 0.849820
departmenttechnical    0.0384488  0.1264317   0.304 0.761046
salarylow              1.8319235  0.1494253  12.260  < 2e-16 ***
salarymedium           1.3381037  0.1502077   8.908  < 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: 11495.5  on 10498  degrees of freedom
Residual deviance:  8990.2  on 10480  degrees of freedom
AIC: 9028.2
```

Figure 12.4: Stepwise Logistic Regression Model

12.2.1 Selection Results

Initial Model AIC: 9028.2
Final Model AIC: 9028.2

The stepwise process didn't remove any predictors, as dropping any of them resulted in a higher AIC.

12.2.2 Key Predictors

Highly significant: satisfaction_level, last_evaluation, number_project, average_montly_hours, time_spend_company, Work_accident, promotion_last_5years, salary.

Not significant: Some department levels, indicating these variables might have less impact on the outcome.

12.2.3 Conclusion

In conclusion, the final model retained all predictors, with most showing strong statistical significance. The predictors like satisfaction level, work accident, and salary are particularly influential in predicting employee turnover.

Section 13: Model Selection

Identifying the appropriate model to be used in the project is crucial to achieve the business goal. Since the aim is to predict a binary/categorical variable 'left', the algorithm to be used should be appropriate for performing classifications.

For this section, there will be a special focus on five algorithms: **Logistic Regression**, **Neural Networks**, **Decision Trees**, **K-Nearest Neighbors (KNN), and Random Forest**. Each of them has different advantages which makes them potential choices, however there are special considerations to consider.

13.1 Logistic Regression

Logistic Regression is a statistical algorithm used for binary classification problems. It models the probability of a binary outcome (e.g., employee leaving or staying) based on one or more predictor variables.

13.1.1 Logistic Regression Advantages

- **Interpretability:** It is possible to clearly observe the relationship between predictors and outcomes through coefficients.

- **Efficiency:** Computationally inexpensive
- **Probabilistic Output:** The results of the predictions are given as probability before making the classification, which for the task proposed in the project would be a great way of obtaining insights into the likelihood of a current employee leaving the company.

13.1.2 Logistic Regression Considerations

- **Linear Assumption**: Assumes a linear relationship between predictors and log-odds, which may not hold in all cases.
- **Imbalanced Data**: Can struggle with datasets where one class significantly outweighs the other.
- **Sensitivity to Outliers**: Outliers can disproportionately influence the model's predictions.
- **Limited Complexity**: May fail to capture complex, non-linear relationships in the data.

13.2 Decision Trees

Decision Trees are a non-linear model that splits the data into subsets based on the specific values of predictors, forming a tree-like structure of decisions. The goal of each split is to minimize impurity at each node.

13.2.1 Classification Trees Advantages

- **Interpretability:** Easy to visualize and understand, especially for non-technical shareholders, as they represent decisions in a hierarchical manner.
- **Flexibility:** Ability to model complex and nonlinear relationships.

13.2.2 Classification Trees Considerations

- **Overfitting:** Prone to overfitting, especially with deep trees, which can result in poor generalization to new data.
- **Stability:** Small changes in the data can lead to different tree structures.

13.3 Random Forest

Random Forest is a machine learning method that combines multiple decision trees to obtain a more accurate and stable prediction.

13.3.1 Random Forest Advantages

- **Performance:** Typically offers high predictive accuracy due to its ensemble approach, combining multiple decision trees.
- **Robustness:** Reduces overfitting compared to individual decision trees by averaging multiple trees.
- **Feature Importance:** Provides insights into the importance of different predictors.

13.3.2 Random Forest Considerations

- **Complexity:** More complex and less interpretable compared to single decision trees.
- **Computational Cost:** Requires more computational resources than individual decision trees, especially with large forests.

13.4 Neural Networks

Neural Networks are advanced models inspired by the human brain's structure. They consist of multiple layers of interconnected nodes (neurons) that process inputs and learn complex patterns through training.

13.4.1 Neural Networks Advantages

- **Flexibility:** This algorithm is capable of handling complex and non-linear relationships between variables.

- **Performance:** Models that use neural networks can obtain relatively high performances, especially in large datasets.

13.4.2 Neural Networks Considerations

- **Complexity:** Requires more computational resources and time for training.

- **Interpretability:** Less transparent compared to simpler models, making it harder to understand the influence of individual predictors.

13.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm that classifies new instances based on the majority class among its k-nearest neighbors in the training set.

13.5.1 KNN Advantages

- **Simplicity:** Easy to understand and implement, with no explicit training phase.

- **Adaptability:** Naturally handles multi-class problems and can model complex boundaries.

13.5.2 KNN Considerations

- **Computational Cost:** Requires significant computation and storage, especially with large datasets, as it needs to compare each test instance to all training instances.
- **Choice of k:** Performance depends on the choice of k and the distance metric used.
- **Euclidean Distance:** Because the model relies on measuring Euclidean distance between variables, it works better with numerical variables and, although it can be used with categorical variables, it is more challenging.

13.6 Model Selection Conclusion

The selection of these models is driven by the need to balance interpretability, computational efficiency, and predictive accuracy. Logistic Regression provides a strong baseline with its interpretability and efficiency. Neural Networks offer the capability to model complex relationships but at the cost of increased computational demands. Decision Trees provide a clear visualization of decisions, though they may overfit, Random Forest balances high accuracy and robustness but can be complex and computationally demanding, while KNN offers simplicity and adaptability but can also be computationally expensive.

Section 14: Model Fitting

In this section, the different algorithms discussed earlier will be trained on the dataset with the goal of capturing patterns that can help classify employees as either having left the company (left=1) or having stayed (left=0).

14.1 Logistic Regression

The first model built was based on logistic regression, which assumes a linear relationship between the predictors and the natural logarithm of odds (log-odds) of the target variable. By estimating the log-odds of the outcome as a linear function of the predictor variables, logistic regression allows these log-odds to be converted into probabilities using the logistic function, which ensures predicted probabilities are between 0 and 1.

```
> summary(stepwise_model)

Call:
glm(formula = left ~ satisfaction_level + last_evaluation + number_project +
    average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years + department + salary, family = binomial,
    data = train_set)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.3388681  0.2297670  -5.827 5.64e-09 ***
satisfaction_level     -4.1234486  0.1169649 -35.254  < 2e-16 ***
last_evaluation         0.7031485  0.1779951   3.950 7.80e-05 ***
number_project         -0.3178972  0.0253265 -12.552  < 2e-16 ***
average_montly_hours    0.0044971  0.0006129   7.338 2.17e-13 ***
time_spend_company      0.2761742  0.0187723  14.712  < 2e-16 ***
Work_accident1         -1.5365054  0.1070103 -14.358  < 2e-16 ***
promotion_last_5years1 -1.5610938  0.3227940  -4.836 1.32e-06 ***
departmenthr            0.1770270  0.1559702   1.135 0.256373
departmentIT           -0.2143942  0.1439415  -1.489 0.136368
departmentmanagement   -0.4727506  0.1860713  -2.541 0.011063 *
departmentmarketing    -0.1557508  0.1580524  -0.985 0.324409
departmentproduct_mng  -0.2788059  0.1553330  -1.795 0.072671 .
departmentRandD        -0.5882774  0.1739140  -3.383 0.000718 ***
departmentsales        -0.0934090  0.1211768  -0.771 0.440797
departmentsupport      -0.0245670  0.1297454  -0.189 0.849820
departmenttechnical     0.0384488  0.1264317   0.304 0.761046
salarylow               1.8319235  0.1494253  12.260  < 2e-16 ***
salarymedium            1.3381037  0.1502077   8.908  < 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: 11495.5  on 10498  degrees of freedom
Residual deviance:  8990.2  on 10480  degrees of freedom
AIC: 9028.2

Number of Fisher Scoring iterations: 5
```

Figure 14.1: Summary of Stepwise Logistic Regression Model

14.1.1 Formula

The summary of the model provides an overview of the variables included in the model as well as the target variable “left”. The model function is composed by an intercept and multiple coefficients. It is possible to notice that for categorical variables, the categories representing accounting department, high salary, no promotion in the last 5 years, and no work accident are missing. The reason for that is that they were used as reference categories and their effect is being captured in the intercept, which reflects the log-odds of the outcome variable when all predictors are at their reference levels.

With exception of those set as reference categories, each predictor has a coefficient that represents their effect on the log-odds of the outcome variable (left). A positive coefficient means that an increase in that predictor leads to a higher probability of the employee leaving, while a negative coefficient suggests that an increase in the predictor’s value leads to a lower probability of leaving. The magnitude (absolute value) of each coefficient describes the strength of this relationship, since it shows how a one-unit change in the predictor influences the log-odds of the outcome. In summary, the sign of the coefficient indicates the direction of the effect (negative or positive), and the absolute value represents the magnitude of the impact.

14.1.2 Standard Error

The standard error measures the variability or uncertainty associated with the estimated coefficient. It indicates how much the estimate of the coefficient might vary from sample to sample. A smaller standard error means that the coefficient is estimated more precisely. Large standard errors relative to the coefficient estimates might indicate less confidence in the estimates.

14.1.3 P-Value

The Pr(>|z|) column displays the p-value, which indicates whether a feature is statistically significant in predicting the outcome variable. The smaller the p-value, the more significant the feature is. Each p-value is also associated with a significance code, which helps identify the level of significance of a variable for the model. P-values less than 0.001 are considered extremely significant, those less than 0.01 are highly significant, those less than 0.05 are significant, and p-values greater than 0.05 are not considered statistically significant.

14.1.4 Significant Predictors

Based on the p-values, the significant predictors are “satisfaction_level”, “last_evaluation”, “number_project”, “average_montly_hours “, “time_spend_company”, “Work_accident”, “promotion_last_5years”, “salary_low” and “salary_medium”. Therefore, it is possible to interpret each of them as follows:

- **satisfaction_level:** This variable has a strong negative coefficient (-4.123), indicating that as employee satisfaction increases, the likelihood of leaving the company decreases significantly.

- **last_evaluation:** A positive coefficient (0.703) indicates that higher performance evaluations are associated with a higher likelihood of leaving.

- **number_project:** A negative coefficient (-0.318) suggests that employees who work on more projects are less likely to leave.
- **average_montly_hours:** A small positive coefficient (0.004) indicates that more monthly work hours slightly increase the chance of an employee leaving.

- **time_spend_company:** A positive coefficient (0.276) shows that employees who have spent more years at the company are more likely to leave.

- **Work_accident:** A negative coefficient (-1.537) suggests that employees who had a work accident are less likely to leave.

- **promotion_last_5years:** A negative coefficient (-1.561) indicates that employees who were promoted in the last 5 years are less likely to leave.

- **salarylow and salarymedium:** Employees with low (1.832) or medium (1.338) salaries are more likely to leave compared to those with high salaries (which is the baseline).

14.1.5 Deviance

The summary of the model also displays information about the deviance, which is a measure of how much the fitted model deviates from a perfect model that can explain all he variability in the data. The lower the deviance, the better the model fits the data.

The null deviance shows the deviance of the model considering only the intercept, therefore informing how well the model predicts the outcome with no features. The residual deviance measures how well the model fits the data when predictors are included. In this case, it is possible to observe that the null deviance is 11495.5, while the residual deviance is 8990.2. These results indicate that the predictors are effective in improving the model's fit to the data.

14.1.6 Akaike Information Criterion (AIC)

Another measure available is the AIC (Akaike Information Criterion), which is a measure of the model's quality. It balances goodness of fit with model complexity by penalizing according to the number of parameters. Lower AIC values indicate better models; however, the measure is more useful when comparing between different models.

14.1.7 Summary

In summary, the model effectively identifies key predictors of employee attrition and shows a good fit to the data, with significant predictors that offer actionable insights into factors influencing employee decisions to leave.

14.2 Decision Tree

The second model was built using the decision tree algorithm. This algorithm provides a hierarchical structure that makes it easy to understand how different predictors and their values lead to a certain target variable outcome (left = 0 or left = 1). As mentioned previously, a decision tree makes classifications by recursively splitting the data based on variable values that will lead to the highest reduction in impurity. This process helps in effectively separating employees who left the company from those who stayed, by creating subsets of data that are as homogenous as possible with respect to the target variable.

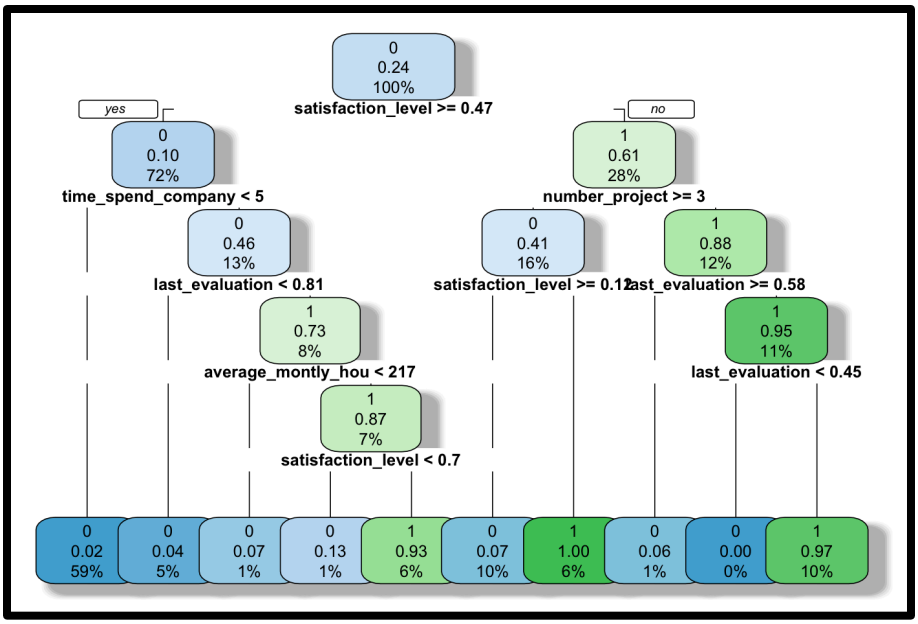

Figure 14.2: Decision Tree Model

14.2.1 Classification Tree Rules

The figure provides a clear representation of the model. The classification tree can be explained as follows:

Root Node (First Split)

It is possible to see that the root node contains 100% of the data, where 24% of the employees who left the company.

- **Satisfaction Level ≥ 0.47**:
    - **Yes (Left Branch)**: 90% of the employee's who have a satisfaction level ≥ 0.47 stayed in the company while 10% left. This node contains 72% of the total observations
    - **No (Right Branch)**: 61% of the employees with a satisfaction level < 0.47 left, while 39% stayed. This node contains 28% of the total observations

Left Subtree (For Employees With Satisfaction Level ≥ 0.47)

- **Time Spent at Company < 5**:
    - **Yes**: 98% of employees who have satisfaction level ≥ 0.47 and have spent less than 5 years at the company stayed, while only 2% left. This node contains 59% of the total observations.

- **No**: 54% of the employees who have satisfaction level ≥ 0.47 and have spent 5 or more years at the company stayed, while 46% left. This node contains 13% of the total observations.

- **Last Evaluation < 0.81**:

  - **Yes**: 96% of employees who have satisfaction level ≥ 0.47 , have spent 5 or more years at the company and have a last evaluation < 0.81 stayed, while 4% left.

  - **No**: 73% of employees who have satisfaction level ≥ 0.47 , have spent 5 or more years at the company and have a last evaluation ≥ 0.81 left, while 27% stayed. This node contains 8% of the total observations.

- **Average Monthly Hours < 217**

  - **Yes**: 93% of the employees who have satisfaction level ≥ 0.47 , have spent 5 or more years at the company , have a last evaluation ≥ 0.81 and have worked < 217 average monthly hours stayed, while 7% left. This node contains 1% of the total observatons.

  - **No**: 87% of the employees who have satisfaction level ≥ 0.47 , have spent 5 or more years at the company , have a last evaluation ≥ 0.81 and have worked ≥ 217 average monthly hours left, while 13% stayed. This node contains 7 % of the total number of observations.

- **Satisfaction Level < 0.7**

  - **Yes**: 87% of the employees who have a satisfaction level ≥ 0.47 but < 0.7, have spent 5 or more years at the company , have a last evaluation ≥ 0.81 and have worked ≥ 217 average monthly hours stayed. This node contains 1% of the total observations.

  - **No**: 93% of employees who have a satisfaction level > 0.7, have spent 5 or more years at the company, have a last evaluation ≥ 0.81 and have worked ≥ 217 average monthly hours left. This node contains 6% of the total observations.

Right Subtree (For Employees With Satisfaction Level < 0.47)

- **Number of Projects ≥ 3**

  - **Yes:** 59% of the employees with a satisfaction level < 47 and number of projects ≥ 3 stayed, while 41% left. This node contains 16% of the total observations.

  - **No:** 88% of the employees with satisfaction level < 47 and number of projects < 3 left, while 12% stayed. This node contains 12% of the total observations.

- **Satisfaction Level ≥ 0.12**
    - **Yes:** 93% of the employees with a satisfaction level ≥ 0.12 but < 47, and number of projects ≥ 3 stayed, while 7% left. This node contains 10% of the total observations.
    - **No:** 100% of the employees with a satisfacion level < 0.12 and number of projects ≥ 3 left. This node contains 6% of the total observations.
- **Last Evaluation ≥ 0.58**
    - **Yes:** 94% of the employees with satisfaction level < 47, number of projects <3, and last evaluation ≥ 0.58 stayed, while 6% left. This node contains 1% of the total observations.
    - **No:** 95% of the employees with a satisfaction level <0.47, number of project < 3, last evaluation < 0.58 left, while 5% stayed. This node contains 11% of the total observations.
- **Last Evaluation < 0.45**
    - **Yes:** There are no records of employees who have a satisfaction level <47, number of projects <3, and last evaluation <0.45.
    - **No:** 97% of employees who have a satisfaction level <47, number of projects < 3, last evaluation ≥ 0.45 but < 0.58 left. This node contains 10 % of the total observations.

It can also be observed that the model judges the features “satisfaction_level”, “time_spend_company”, “number_project” and “last_evaluation” to be the most important predictors of employee turnover.

14.2.2 Classification of Employees at Risk of Leaving

The model classifies that employees will leave the company (left=1) if they meet any of the following conditions:

1. Employees who have a satisfaction level > 0.7, have spent 5 or more years at the company, have a last evaluation ≥ 0.81 and have worked ≥ 217 average monthly hours.
2. Employees with a satisfacion level < 0.12 and number of projects ≥ 3.
3. Employees who have a satisfaction level <47, number of projects < 3, last evaluation ≥ 0.45 but < 0.58.

14.3 Random Forest

The third model was built using a Random Forest algorithm. This technique involves generating multiple decision trees using randomly selected subsets of features and samples. This diversity among the trees ensures that the model captures a wide range of patterns and relationships in the data. Consequently, it reduces the risk of overfitting and enhances the model's ability to generalize, leading to a more robust and reliable performance compared to individual decision trees.

```
> rf_model

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

        OOB estimate of  error rate: 0.94%
Confusion matrix:
     0    1 class.error
0 7999   13 0.001622566
1   86 2401 0.034579815
```

Figure 14.3: Random Forest Model

The model summary ***(Figure 14.3)*** indicates that the target variable is “left,” with 500 classification trees generated, considering 3 randomly selected variables at each split. The Out-Of-Bag (OOB) error rate, a cross-validation measure in Random Forests where each tree is tested on data not used for its training, is 0.94%, reflecting strong overall model performance.

The confusion matrix reveals a misclassification rate of 0.16% for records incorrectly predicted as “left = 1” and 3.46% for records incorrectly predicted as “left = 0.”

These results suggest the model performs well on the training data, demonstrating a very low OOB error rate and high accuracy, particularly for class 0 (employees who stayed). However, the slightly higher error rate for class 1 (employees who left) indicates some difficulty in accurately identifying individuals at risk of leaving.

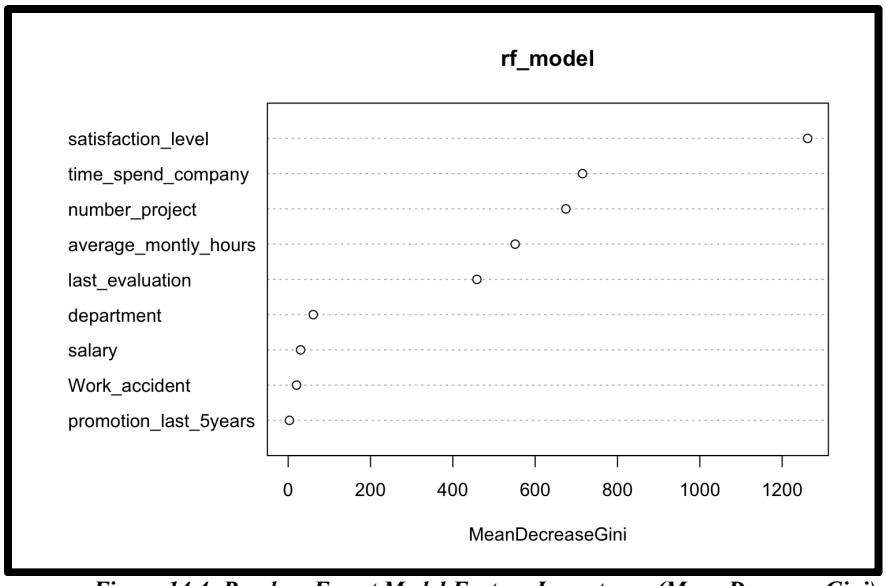

Figure 14.4: Random Forest Model Feature Importance (Mean Decrease Gini)

The plot ***(Figure 14.4)*** illustrates the importance of the predictors used in the model accoding to the Mean Decrease Gini, which provides the importance of each variable in terms of how much it improves the purity of the nodes in the trees. Higher values indicate more importance.

It is possible to observe that “satisfaction level” is by far the feature that improves purity the most, followed by “time spent in the company” and “number of projects”. On the other hand, the predictors “work accident” and “promotion within the last 5 years” are the ones with the lowest effect in impurity reduction.

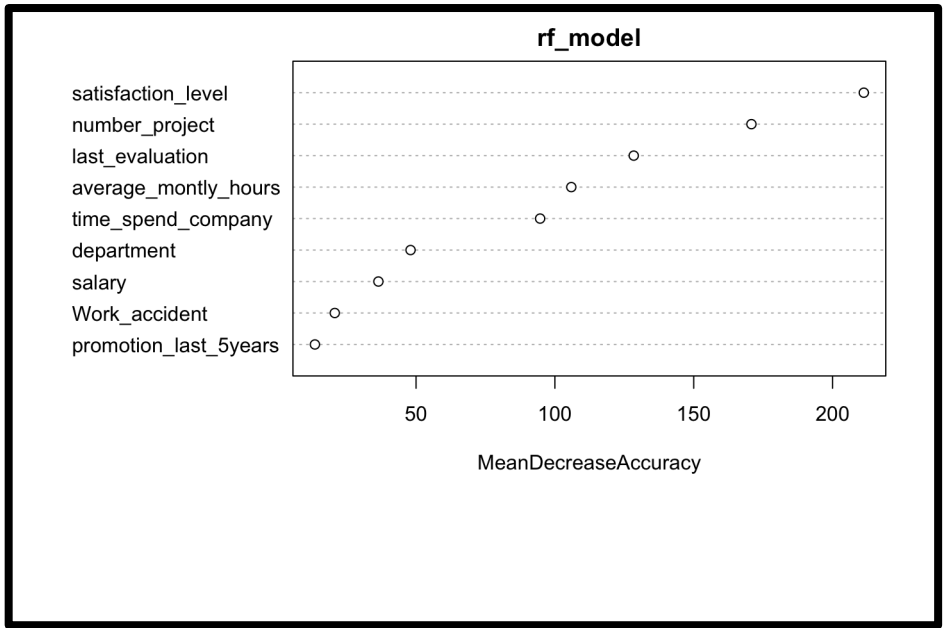

Figure 14.5: Random Forest Model Feature Importance (Mean Decrease Accuracy)

This second plot ***(Figure 14.5)*** displays the importance of the predictors used in the model according to the Mean Decrease Accuracy, which assesses how much the accuracy of the model is decreased when a certain predictor value is shuffled, thereby disrupting the original relationship between the predictor and the target variable. In simpler terms, it measures the impact of each variable on the model's accuracy. Higher values indicate that the variable improves prediction.

It is noticeable that "satisfaction level", "number of projects", and "last evaluation" play a significant role in the model's predictive performance. In contrast, "work accident" and "promotion within the last 5 years" are not as impactful.

In summary, understanding which variables are important in reducing impurity or increasing accuracy, is crucial for improving the model's performance when necessary as well as for obtaining business insights. In this case, the features "satisfaction level" and "number of projects" were consistently at the top of both measures, suggesting they are crucial factors to consider for reducing employee turnover.

14.4 Neural Networks

The fourth and fifth model were built using Neural Networks. Neural Networks models are composed of nodes, also called neurons and they are structured with three types of layers: the input layer, one or more hidden layers, and the output layer. The input layer receives the values of each predictor, the hidden layers apply activation functions to transform these values, and the output layer provides the final prediction or classification result.

The Neural Networks algorithm is very flexible and allows for the choice of different numbers of hidden layers and nodes within each layer. It is also possible to choose among different activation functions depending on the specific goal. In this case, a sigmoid activation function was selected for both models because the goal is binary classification, and this function is well-suited for this purpose.

Before training the model, each data partition (training, validation, and test) was normalized, ensuring that the values of each variable fall within the range of 0 to 1. This normalization is a best practice that improves the model's ability to learn patterns more effectively and efficiently. Normalizing all partitions using the training set's statistics is crucial to prevent data leakage and ensure that the validation and test phases involve data scaled consistently with the training process, providing a fair and realistic assessment of the model's performance.

In addition, the binary target variable was converted into numeric format to be processed correctly by the network and its associated loss functions. This numeric format is essential for the network to perform binary classification effectively.

14.4.1 Single Hidden Node Model

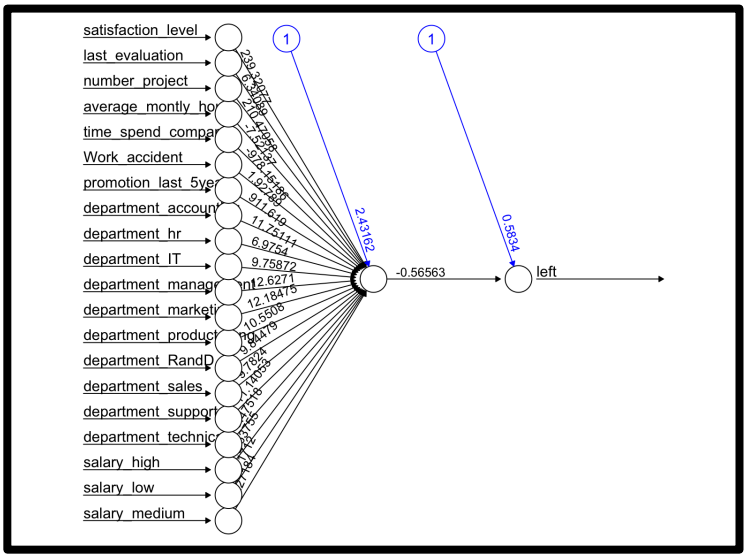

Figure 14.6: Single Hidden Node Model

The first model ***(Figure 14.6)*** was created using the simplest possible structure, where there are 20 nodes in the input layer, each representing a predictor, only one node in the hidden layer, and one node in the output layer representing the target variable.

It is possible to see that the each connection between neurons have a weight associated with it. These weights act as coefficient, determining how much influence a particular input feature has on the neuron's output. By learning and adjusting these weights, the neural network can represent complex relationships between inputs and outputs. Besides the weights, each neuron (except those in the input layer) also has a bias term, which is an additional parameter that shifts the activation function and adds flexibility to the model. It helps the network learn and fit the data more effectively.

The weights are obtained through a method called backpropagation. It begins with randomly initialized small weights for the network's connections. The input data is passed through the network to generate predictions. These predictions are then compared to the actual target values using a loss function, which measures the discrepancy between the predictions and the targets.

The backpropagation algorithm updates the weights in the network to minimize this error. It calculates the gradient of the loss function with respect to each weight by propagating the error backward through the network. Weights are adjusted in the direction that reduces the loss. This process is repeated iteratively until the errors are minimized to an acceptable level or no further significant improvement can be made.

The calculations made in the hidden node and output node of the final model can be written as follows:

$$N_{Hidden\ Layer}\ \frac{1}{1+e^{-(\,2.43162\,+(239.32077\,*satisfaction\_level)+(6.34089*\text{last\_evaluation})+\cdots+(10.27184*\text{salary\_medium}))}}$$

$$N_{Output\ Layer}\ \frac{1}{1+e^{-(0.5834+(-0.56563*\,\mathrm{N_{Hidden\ Layer}}\,))}}$$

14.4.2 Two Hidden Node Model

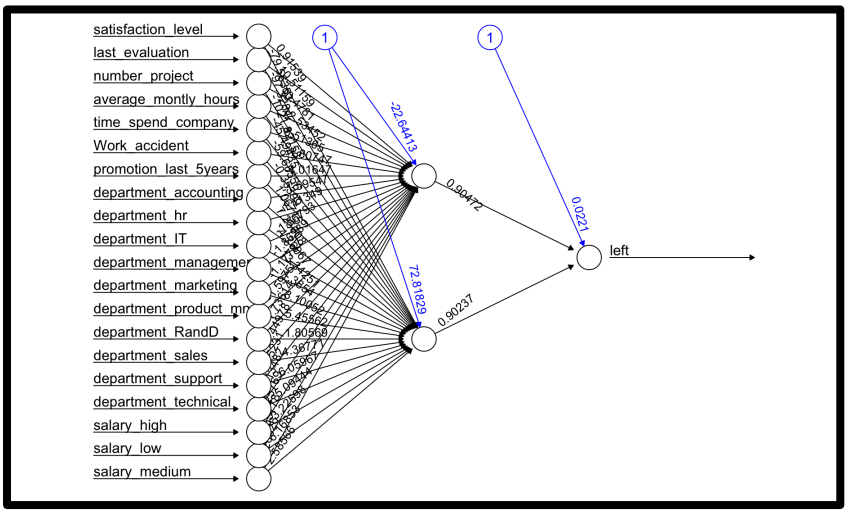

Figure 14.7: Two Hidden Node Model

```
> NN_model$result.matrix
                                  [,1]
error                     5.529748e+02
reached.threshold         9.154094e-03
steps                     4.639600e+04
```

Figure 14.8: Training Results for Model 1

```
> NN_model2$result.matrix
                                  [,1]
error                     2.792015e+02
reached.threshold         9.694263e-03
steps                     1.385400e+04
```

Figure 14.9: Training Results for Model 2

The second neural network ***(Figure 14.7)*** is more complex than the first, since it has 2 neurons in its hidden layer. Although the fundamental training process is similar, the path to the output layer is more elaborate due to the additional neuron, which can potentially enhance the model's ability to capture complex patterns.

Despite being more complex, the second model ***(Figure 14.9)*** achieved approximately half the error of the first model ***(Figure 14.8)*** and required significantly fewer iterations to converge. The second model reached convergence in 13,854 iterations, while the first model needed 46,396 iterations. This indicates that the second model, although more complex, learned more efficiently and achieved better performance in the training data with fewer steps.

14.5 K-Nearest Neighbors (KNN)

The last model created was based on the KNN algorithm, in which the goal is to identify k records in the training dataset that are most similar to the new records being classified. Similarity is measured using Euclidean distance, which calculates the straight-line distance between data points. Once the nearest neighbors are identified, the new records are classified based on the majority class among these neighbors.

To accurately measure the distances between records, the data had to be normalized so that all features are on the same scale. This ensures that features with larger ranges do not disproportionately affect the distance calculations, leading to more balanced and reliable results.

In order to choose the optimal number of neighbours (k), a 10-fold cross validation was performed on the training data. This process involves splitting the dataset into 10 folds. For each k value, the model is trained on 9 folds and validated on the remaining fold, repeating this process 10 times with each fold serving as the test set once. Performance metrics from each fold are averaged to assess the model's overall performance and identify the best k. This method ensures that the model's performance is robust and evaluates how each k value performs on different subsets of the training data.

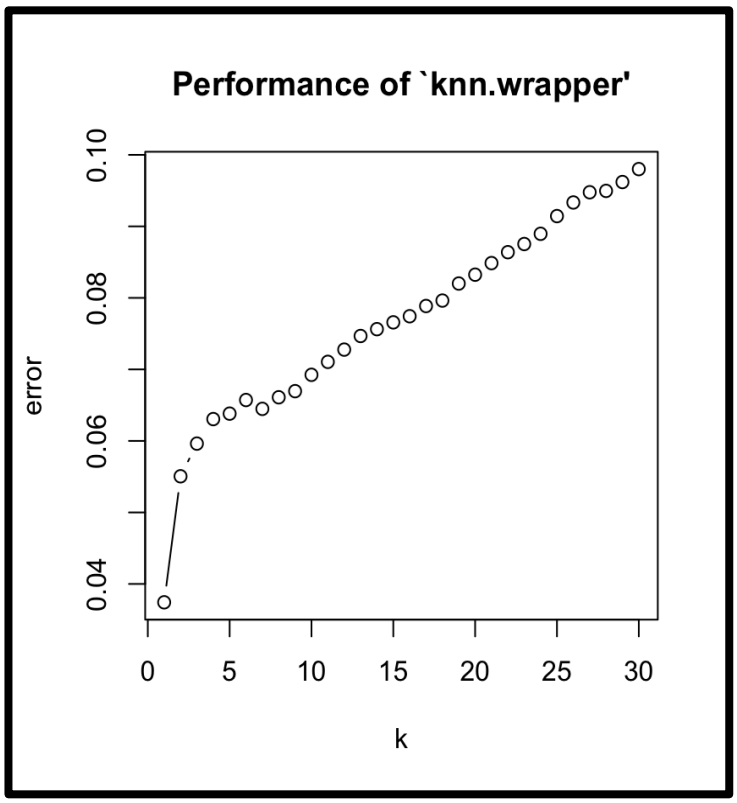

Figure 14.10: Impact of k on Error Rate

```
Parameter tuning of 'knn.wrapper':

- sampling method: 10-fold cross validation

- best parameters:
 k
 1

- best performance: 0.03743257

- Detailed performance results:
    k       error  dispersion
1   1 0.03743257 0.006509840
2   2 0.05505307 0.006316221
3   3 0.05962440 0.005158099
4   4 0.06305307 0.007638162
5   5 0.06381524 0.005745267
6   6 0.06571973 0.007019918
7   7 0.06448191 0.005288728
8   8 0.06610096 0.008372979
9   9 0.06695765 0.009161041
10 10 0.06924381 0.008111954
11 11 0.07105370 0.007207348
12 12 0.07276799 0.008816045
13 13 0.07467302 0.008816502
14 14 0.07562549 0.008827835
15 15 0.07657824 0.008784147
16 16 0.07743465 0.009827317
17 17 0.07886368 0.009080262
18 18 0.07962540 0.009388397
19 19 0.08200672 0.007925054
20 20 0.08324509 0.009432782
21 21 0.08486395 0.009922290
22 22 0.08638822 0.009807794
23 23 0.08753135 0.009986119
24 24 0.08895992 0.010192206
25 25 0.09143647 0.009515940
26 26 0.09334132 0.010841763
27 27 0.09477017 0.011255444
28 28 0.09496119 0.011113870
29 29 0.09619937 0.010865528
30 30 0.09800908 0.010473089
```

Figure 14.11: KNN Parameter Tuning Results

As shown in the figures ***14.10*** and ***14.11***, the optimal number of neighbors (k) is 1, with an error rate of 0.0374. This indicates that the model achieves the best performance by classifying a new record based on the class of its nearest neighbor. Although cross-validation was used to minimize overfitting, relying on a single k value can still introduce model instability. To improve stability while maintaining strong performance, k = 7 was selected as it offers a favorable balance between model accuracy and robustness.

Section 15: Performance Evaluation

With all the models trained, this phase of the project involves evaluating their performance and selecting the model that has the best performance. This will be done by applying them to the validation set, where unseen data with already known outcomes will be classified. This will approach allows for the comparison between predicted and actual classes. The validation set contains 3000 observations.

The key metrics for evaluation will be accuracy, sensitivity, and specificity. These metrics will be derived from a confusion matrix, which summarizes the classifications made by the model in comparison to the actual outcomes.

In a confusion matrix for binary classification, it is possible to see four different classes, as follows:

- **True Positives (TP)**: The number of instances correctly predicted as the positive class.
- **True Negatives (TN)**: The number of instances correctly predicted as the negative class.
- **False Positives (FP)**: The number of instances incorrectly predicted as the positive class when they are actually negative.
- **False Negatives (FN)**: The number of instances incorrectly predicted as the negative class when they are actually positive.

These classes are crucial for calculating the performance metrics. Their formulas and definitions are as follows:

Accuracy: Proportion of correct predictions out of all predictions.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity: Proportion of actual positives that are correctly identified.

$$\frac{TP}{TP + FN}$$

Specificity: Proportion of actual negatives that are correctly identified.

$$\frac{TN}{TN + FP}$$

Precision (Positive Predicted Value): Precision is the proportion of predicted positive cases that are true positives. It measures the accuracy of the model's positive predictions.

$$\frac{TP}{TP + FP}$$

In this analysis, the positive class are employees who left (left=1) and the negative class are employees who stayed (left=0). Since the analytics goal is to identify employees who are likely to leave, considering sensitivity is of extreme importance, because failing to identify at-risk individuals could mean the loss of valuable and talented workers to competitors, which directly conflicts with the business goal and can weaken the company's position in the market. Therefore, a model that maximizes sensitivity is essential for effective employee retention.

Additionally, precision is also a critical metric, as it indicates how accurately the model identifies workers who may leave. A model may be effective at flagging employees likely to

leave but if it lacks precision, the HR team might expend unnecessary effort on employees who are not truly at risk. This not only waste resources but can also cause confusion and disrupt the workplace environment.

15.1 Naïve Model

The Naïve model is a benchmark model that serves as a baseline for evaluating more sophisticated models. It sets the minimum performance standard expected from any predictive model. By comparing the naïve model's performance with that of other models, we can determine whether those models are valid and offer improvements by outperforming naïve predictions.

The naïve model works by predicting the majority class for all instances. In this case, based on the training set, where most employees stayed (8,012) compared to those who left (2,487), the naïve model predicts that all employees will stay (left = 0).

When applied to the validation set, the confusion matrix in ***Table 15.1*** is obtained:

	REFERENCE	
PREDICTION	0	1
0	2,256	744
1	0	0

Table 15.1: Confusion Matrix for Naïve Model

In this case, True Positives are 0, True Negatives are 2,256, False Positives are 0, and False Negatives are 744. Therefore, this classification provides an accuracy of 75.2%, a specificity of 100%, a sensitivity of 0%, and precision is undefined.

While the naïve model achieves perfect specificity (100%) by correctly identifying all negative instances (employees who stayed, class 0), it completely fails to identify any positive instances (employees who left, class 1), resulting in a recall (sensitivity) of 0%. Its accuracy of 75.2% reflects the imbalance in the dataset, as it benefits from the large proportion of class 0 (employees who stayed). Precision cannot be calculated because the model never predicts class 1.

This demonstrates that while the naïve model is a useful benchmark, it is not suitable for tasks where identifying both classes (stayed and left) is important, especially the minority class (employees who left).

15.2 Logistic Regression

```
> confusion_matrix_lreg
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2093  470
         1  163  274

               Accuracy : 0.789
                 95% CI : (0.774, 0.8035)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : 1.007e-06

                  Kappa : 0.3435

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

            Sensitivity : 0.36828
            Specificity : 0.92775
         Pos Pred Value : 0.62700
         Neg Pred Value : 0.81662
             Prevalence : 0.24800
         Detection Rate : 0.09133
   Detection Prevalence : 0.14567
      Balanced Accuracy : 0.64801

       'Positive' Class : 1
```

Figure 15.1: Confusion Matrix for Stepwise Logistic Regression

The confusion matrix ***(Figure 15.1)*** shows an accuracy of 78.9%, which is higher than Naïve, indicating the model provides a more effective prediction compared to the baseline. The specificity is 92.77%, demonstrating the model is capable to correctly identify most of employees who are likely to stay. On the other hand, it has a low sensitivity (36.83%), showing that it is not great at identifying employees who are likely to leave.

Despite the model's current low sensitivity, there is potential to enhance its performance on this metric by adjusting the decision threshold. Specifically, by lowering the threshold from the default value of 0.5 (which classifies records with probabilities above 0.5 as positive), sensitivity can be improved. This adjustment may result in a trade-off where accuracy and specificity are sacrificed in favor of increased sensitivity.

```
> confusion_matrix_lreg_improved
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1797  208
         1  459  536

               Accuracy : 0.7777
                 95% CI : (0.7624, 0.7924)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : 0.0005382

                  Kappa : 0.4645

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

            Sensitivity : 0.7204
            Specificity : 0.7965
         Pos Pred Value : 0.5387
         Neg Pred Value : 0.8963
             Prevalence : 0.2480
         Detection Rate : 0.1787
   Detection Prevalence : 0.3317
      Balanced Accuracy : 0.7585

       'Positive' Class : 1
```

Figure 15.2: Confusion Matrix for Stepwise Logistic Regression with Threshold 0.3

By lowering the threshold to 0.3 **(Figure 15.2)**, the model's ability to identify employees likely to leave improved significantly (from 36.83% to 72.04%). This adjustment led to a slight decrease in accuracy, from 78.9% to 77.77%, and a more impactful decrease in specificity (from 92.77% to 79.65%) and precision (from 62.7% to 53.87%). Despite these changes, the trade-off appears reasonable, given the notable improvement in sensitivity and the model's enhanced performance in detecting potential employee turnover.

15.3 Decision Tree

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

          Reference
Prediction    0    1
         0 2224   60
         1   32  684

               Accuracy : 0.9693
                 95% CI : (0.9625, 0.9752)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9167

 Mcnemar's Test P-Value : 0.004879

            Sensitivity : 0.9194
            Specificity : 0.9858
         Pos Pred Value : 0.9553
         Neg Pred Value : 0.9737
             Prevalence : 0.2480
         Detection Rate : 0.2280
   Detection Prevalence : 0.2387
      Balanced Accuracy : 0.9526

       'Positive' Class : 1
```

Figure 15.3: Confusion Matrix for Decision Tree

The confusion matrix ***(Figure 15.3)*** reveals that the decision tree model excels with an accuracy of 96.93%, sensitivity of 91.94%, specificity of 98.58%, and precision of 95.53%. This model outperforms the logistic regression model in all these aspects. The high performance aligns well with the business goal of identifying employees who may leave. Additionally, decision trees are straightforward to interpret, providing clear insights into the decision-making process.

In terms of practicality, the decision tree model is both efficient and scalable, making it a strong candidate for implementation. Considering all factors, the decision tree emerges as a strong candidate as the final model.

15.4 Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2251   20
         1    5  724

               Accuracy : 0.9917
                 95% CI : (0.9877, 0.9946)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9775

 Mcnemar's Test P-Value : 0.00511

            Sensitivity : 0.9731
            Specificity : 0.9978
         Pos Pred Value : 0.9931
         Neg Pred Value : 0.9912
             Prevalence : 0.2480
         Detection Rate : 0.2413
   Detection Prevalence : 0.2430
      Balanced Accuracy : 0.9855

       'Positive' Class : 1
```

Figure 15.4: Confusion Matrix for Random Forest

By analyzing the confusion matrix ***(Figure 15.4)***, it is possible to see that the Random Forest delivers outstanding performance, with an accuracy of 99.17%, sensitivity of 97.31%, specificity of 99.78%, and precision of 99.31%. These percentages represent exceptionally high levels of accuracy and reliability, indicating that the model performs extremely well in classifying both positive and negative cases.

The Random Forest model's superior sensitivity means it effectively identifies almost all true positives, while its high specificity ensures that most true negatives are correctly classified. Compared to previous models, the Random Forest not only excels in overall accuracy but also in minimizing both false positives and false negatives.

These metrics suggest that the Random Forest model offers a significant improvement over the other models considered, which also makes it a great potential choice for the final model, especially given its higher robustness compared to the decision tree.

15.5 Neural Networks

15.5.1 Single Hidden Node Model

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

          Reference
Prediction    0    1
         0 1760   35
         1  496  709

               Accuracy : 0.823
                 95% CI : (0.8089, 0.8365)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6071

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

            Sensitivity : 0.9530
            Specificity : 0.7801
         Pos Pred Value : 0.5884
         Neg Pred Value : 0.9805
             Prevalence : 0.2480
         Detection Rate : 0.2363
   Detection Prevalence : 0.4017
      Balanced Accuracy : 0.8665

       'Positive' Class : 1
```

Figure 15.5: Confusion Matrix for Neural Networks with Single Hidden Node

The first Neural Networks model ***(Figure 15.5)***, with one node in the hidden layer, provides an accuracy of 82.3%, sensitivity of 95.30%, a specificity of 78.01%, and a precision of 58.84%.

Although its accuracy is not as high, the model excels in sensitivity. Therefore, it effectively identifies employees who are likely to leave the organization, which aligns very well with the analytics goal.

15.5.2 Two Hidden Node Model

```
> print(confusion_matrix_NN2)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2137   89
         1  119  655

               Accuracy : 0.9307
                 95% CI : (0.921, 0.9395)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8166

 Mcnemar's Test P-Value : 0.04435

            Sensitivity : 0.8804
            Specificity : 0.9473
         Pos Pred Value : 0.8463
         Neg Pred Value : 0.9600
             Prevalence : 0.2480
         Detection Rate : 0.2183
   Detection Prevalence : 0.2580
      Balanced Accuracy : 0.9138

       'Positive' Class : 1
```

Figure 15.6: Confusion Matrix for Neural Networks with Two Hidden Nodes

The second neural networks model ***(Figure 15.6)***, with 2 nodes in the hidden layer, obtained a significantly higher accuracy, specificity, and precision than the simpler model. However, it underperformed in sensitivity.

Based on these results, the decision was made to lower the threshold to 0.3 in an attempt to enhance its performance in identifying employees at risk of leaving.

```
> print(confusion_matrix_NN2_improved)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2099   65
         1  157  679

               Accuracy : 0.926
                 95% CI : (0.916, 0.9351)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8095

 Mcnemar's Test P-Value : 1.012e-09

            Sensitivity : 0.9126
            Specificity : 0.9304
         Pos Pred Value : 0.8122
         Neg Pred Value : 0.9700
             Prevalence : 0.2480
         Detection Rate : 0.2263
   Detection Prevalence : 0.2787
      Balanced Accuracy : 0.9215

       'Positive' Class : 1
```

Figure 15.7: Confusion Matrix for Neural Networks with Two Hidden Nodes and Threshold 0.3

In ***Figure 15.7*** it is evident that by lowering the threshold, there has been a significant improvement in the model's overall performance, since the sensitivity went from 88.04% to 91.26% while maintaining a high accuracy, specificity, and a fair precision.

15.6 K-Nearest Neighbors (KNN)

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

          Reference
Prediction    0    1
         0 2156   21
         1  100  723

               Accuracy : 0.9597
                 95% CI : (0.952, 0.9664)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8956

 Mcnemar's Test P-Value : 1.332e-12

            Sensitivity : 0.9718
            Specificity : 0.9557
         Pos Pred Value : 0.8785
         Neg Pred Value : 0.9904
             Prevalence : 0.2480
         Detection Rate : 0.2410
   Detection Prevalence : 0.2743
      Balanced Accuracy : 0.9637

       'Positive' Class : 1
```

Figure 15.8: Confusion Matrix for KNN (K=1)

In this model ***(Figure 15.8)***, the number of neighbors (k) used is 1. The model provided an accuracy of 95.97%, a very high sensitivity of 97.18%, a specificity of 95.57%, and a precision of 87.85%. These results show the model has successfully achieved the task at hand, with one of the highest sensitivities among all models. However, as discussed previously, using a k=1 may not be practical, even though the model also performed well in the validation phase.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2145   74
         1  111  670

               Accuracy : 0.9383
                 95% CI : (0.9291, 0.9467)
    No Information Rate : 0.752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8374

 Mcnemar's Test P-Value : 0.008126

            Sensitivity : 0.9005
            Specificity : 0.9508
         Pos Pred Value : 0.8579
         Neg Pred Value : 0.9667
             Prevalence : 0.2480
         Detection Rate : 0.2233
   Detection Prevalence : 0.2603
      Balanced Accuracy : 0.9257

       'Positive' Class : 1
```

Figure 15.9: Confusion Matrix for KNN (K=7)

By analyzing the confusion matrix of the KNN model where k=7 was used ***(Figure 15.9)***, it is possible to observe the model's performance metrics decreased, however an accuracy of 93.83%, sensitivity of 90.05%, specificity of 95.08%, and precision of 85.79% are still considered great.

15.7 Model Performance via ROC and AUC

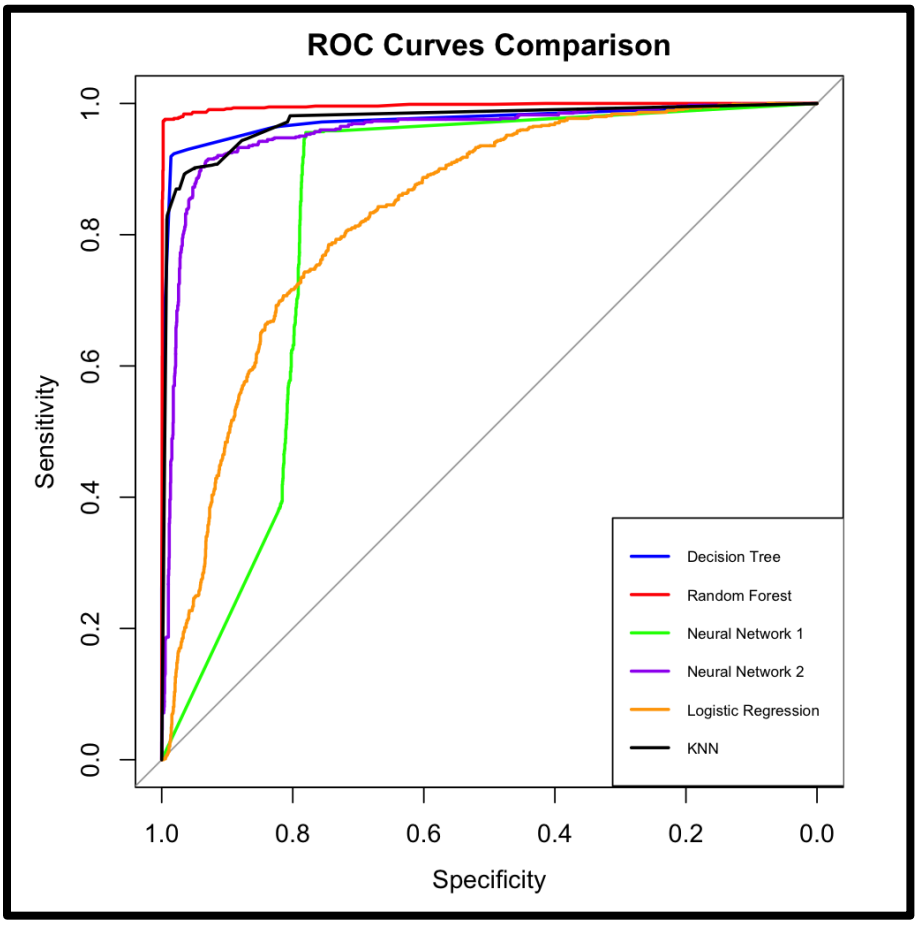

Figure 15.10: Comparison of ROC Curves

The ROC (Receiver Operating Characteristic) curve ***(Figure 15.10)*** illustrates the trade-off between sensitivity and specificity across different thresholds, allowing for the analysis of how changes in sensitivity affect specificity. It is possible to observe that at a given sensitivity level, the ROC curve shows the corresponding specificity, revealing how much of one metric is sacrificed for the other.

The curve also facilitates the comparison of different models. Models that achieve high sensitivity and specificity with minimal trade-offs are considered superior. This comparison is often summarized using the AUC (Area Under the Curve) metric. AUC values range from 0 to 1, with a perfect model scoring an AUC of 1. The closer the AUC is to 1, the better the model's overall performance.

The diagonal line in the ROC plot represents a random classifier with an AUC of 0.5, indicating no discrimination ability. Models with AUC values above this line show better-than-random performance.

By observing the plot, models like Random Forest, Decision Tree, and Neural Network with 2 nodes in the hidden layer are better performers compared to the Neural Network with 1 node in the hidden layer and logistic regression. For instance, the Random Forest model, with an AUC of 0.9961, nearly reaches perfection, indicating its exceptional ability to distinguish between classes.

15.8 Performance Comparison

Method	Accuracy	Sensitivity	Specificity	Precision	AUC
Naïve	75.2%	0	100%	-	0.5
Logistic Regression	77.77%	72.04%	79.65%	53.87%	0.8322
Random Forest	99.17%	97.31%	99.78%	99.31%	0.9961
NN with 1 Node in the Hidden Layer	82.30%	95.30%	78.01%	58.84%	0.823
NN with 2 Nodes in the Hidden Layer	92.60%	91.26%	93.04%	81.22%	0.9556
Decision Tree	96.93%	91.94%	98.58%	95.53%	0.9737
KNN	93.83%	90.05%	95.08%	85.79%	0.9733

Table 15.2: Performance Comparison Across Models

Given the results obtained for evaluating model performance ***(Figure 15.11)***, it is possible to make a final comparison.

Logistic Regression and Neural Network with 1 hidden node perform significantly worse than the other models and can therefore be discarded.

Random Forest, Neural Network with 2 hidden nodes, Decision Tree and KNN demonstrated strong and relatively similar performances. However, Neural Network and KNN still lagged behind Decision Tree and Random Forest in terms of accuracy and other metrics. In addition to that, KNN has the disadvantage of being a “lazy learner,” meaning that it needs to store the entire training dataset and calculate the distances for each prediction. This process can be slow, especially as the dataset grows. KNN is also sensitive to outliers, which can impact its predictions. Neural Networks can capture complex patterns, but they are computationally expensive, act as a “black box model,” and did not offer the same accuracy as other models, making the cost-benefit ratio less attractive.

Random Forest and Decision Tree emerged as the top contenders for achieving the analytics goal. Decision Tree offers the advantage of interpretability and transparency in how it makes classifications. However, its instability (sensitivity to small changes in the data) can be a concern. Random Forest, on the other hand, had the best performance across all metrics and addresses the issue of instability of individual decision trees by being a robust model trained on 500 trees. Although Random Forest is less transparent and more resource intensive, the benefits outweigh these drawbacks. Therefore, it was selected as the final model for implementation.

Section 16: Test phase

After building multiple models based on different algorithms and evaluating how each of them perform in the validation phase, the project has reached the test phase.

As concluded earlier, the Random Forest model has been identified as the most effective and suitable for deployment. However, before proceeding, it will undergo a final evaluation to confirm its effectiveness and real-world applicability. This test phase involves applying the model to an unseen sample of 1,500 records to validate its reliability and performance.

```
> final_confusion_matrix
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1158   12
         1    2  328

               Accuracy : 0.9907
                 95% CI : (0.9844, 0.9949)
    No Information Rate : 0.7733
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9731

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.9647
            Specificity : 0.9983
         Pos Pred Value : 0.9939
         Neg Pred Value : 0.9897
             Prevalence : 0.2267
         Detection Rate : 0.2187
   Detection Prevalence : 0.2200
      Balanced Accuracy : 0.9815

       'Positive' Class : 1
```

Figure 16.1: Confusion Matrix for Random Forest Model on Test Set

The final confusion matrix confirms the model's success and readiness for implementation since it provided a 99.07% accuracy, a 96.47% sensitivity, a 99.83% specificity, and a 99.39% precision. These results are consistent with the ones obtained in the previous phases.

Section 17: Model Implementation

To effectively implement the model and identify employees at risk of leaving the company, the original dataset, which included both current employees and those who had left, was filtered to retain only current employees ***(Figure 17.1).***

```
> employee_final.df<- subset(employee.df, left == 0)
> head(employee_final.df)
     satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accident left
2001               0.58            0.74              4                  215                 3             0    0
2002               0.82            0.67              2                  202                 3             0    0
2003               0.45            0.69              5                  193                 3             0    0
2004               0.78            0.82              5                  247                 3             0    0
2005               0.49            0.60              3                  214                 2             0    0
2006               0.36            0.95              3                  206                 4             0    0
     promotion_last_5years department salary
2001                     0      sales    low
2002                     0      sales    low
2003                     0      sales    low
2004                     0      sales    low
2005                     0      sales    low
2006                     0      sales    low
```

Figure 17.1: Preview of Current Employees Dataset After Filtering

After filtering, a random forest model was applied to estimate each current employee's likelihood of leaving. The predictions from the model were added to the dataset as a new column, 'probability_leaving' ***(Figure 17.2).***

```
> head(employee_final.df)
     satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accident left promotion_last_5years
2001               0.58            0.74              4                  215                 3             1    0                     0
2002               0.82            0.67              2                  202                 3             1    0                     0
2003               0.45            0.69              5                  193                 3             1    0                     0
2004               0.78            0.82              5                  247                 3             1    0                     0
2005               0.49            0.60              3                  214                 2             1    0                     0
2006               0.36            0.95              3                  206                 4             1    0                     0
     department salary probability_leaving
2001      sales    low               0.020
2002      sales    low               0.006
2003      sales    low               0.016
2004      sales    low               0.002
2005      sales    low               0.004
2006      sales    low               0.054
```

Figure 17.2: Preview of Current Employees Dataset with Probability of Leaving

Next, the probabilities generated by the model were categorized into turnover risk categories to provide HR with a clearer understanding of each employee's risk level ***(Figure 17.3)***. The thresholds used for categorization are detailed in ***Table 17.1****:*

Category	Threshold	Description
Low Risk	Probability ≤ 0.1	Employees with a very low likelihood of leaving.
Low-Moderate Risk	0.1 < Probability < 0.3	Employees with a slightly elevated likelihood of leaving.
Moderate Risk	0.3 ≤ Probability < 0.5	Employees with a moderate likelihood of leaving.
High Risk	Probability ≥ 0.5	Employees with a high likelihood of leaving.

Table 17.1: Turnover Risk Categories and Thresholds

```
> head(employee_final.df)
     satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accident left promotion_last_5years
2001               0.58            0.74              4                  215                 3             1    0                     0
2002               0.82            0.67              2                  202                 3             1    0                     0
2003               0.45            0.69              5                  193                 3             1    0                     0
2004               0.78            0.82              5                  247                 3             1    0                     0
2005               0.49            0.60              3                  214                 2             1    0                     0
2006               0.36            0.95              3                  206                 4             1    0                     0
     department salary probability_leaving turnover_risk
2001      sales    low               0.020      Low Risk
2002      sales    low               0.006      Low Risk
2003      sales    low               0.016      Low Risk
2004      sales    low               0.002      Low Risk
2005      sales    low               0.004      Low Risk
2006      sales    low               0.054      Low Risk
```

Figure 17.3: Preview of Current Employees Dataset with Turnover Risk Categories

Subsequently, the impact score was reintroduced into the dataset of current employees. This score enables HR to assess not only an employee's turnover risk but also the significance of their departure's impact on the organization, allowing for prioritization of cases that require urgent attention ***(Figure 17.4)***.

```
> head(employee_final.df)
     satisfaction_level last_evaluation number_project average_montly_hours time_spend_company Work_accident left promotion_last_5years
2001               0.58            0.74              4                  215                 3             1    0                     0
2002               0.82            0.67              2                  202                 3             1    0                     0
2003               0.45            0.69              5                  193                 3             1    0                     0
2004               0.78            0.82              5                  247                 3             1    0                     0
2005               0.49            0.60              3                  214                 2             1    0                     0
2006               0.36            0.95              3                  206                 4             1    0                     0
     department salary probability_leaving turnover_risk impact_score
2001      sales    low               0.020      Low Risk          0.2
2002      sales    low               0.006      Low Risk          0.2
2003      sales    low               0.016      Low Risk          0.2
2004      sales    low               0.002      Low Risk          0.4
2005      sales    low               0.004      Low Risk          0.2
2006      sales    low               0.054      Low Risk          0.6
```

Figure 17.4: Preview of Current Employees Dataset with Turnover Risk Categories and Impact Scores

The dataset was segmented by risk categories and prioritized within each category based on a hierarchical structure, first emphasizing employees with the highest impact score, followed by those with the greatest likelihood of leaving. This organization ensures that employees requiring the most urgent attention are clearly identified within each risk category, enabling HR to take targeted and efficient action.

Figures 17.5 to 17.8 display the top 10 employees for each risk category, prioritized by impact score and then by probability of leaving.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary	probability_leaving	turnover_risk	impact_score
1	0.95	0.90	4	221	10	2	0	0	management	high	0.066	Low Risk	1
2	0.95	0.90	4	221	10	2	0	0	management	high	0.066	Low Risk	1
3	0.24	0.81	6	263	7	1	0	0	management	high	0.062	Low Risk	1
4	0.24	0.81	6	263	7	1	0	0	management	high	0.062	Low Risk	1
5	0.50	1.00	5	264	8	1	0	1	accounting	high	0.032	Low Risk	1
6	0.50	1.00	5	264	8	1	0	1	accounting	high	0.032	Low Risk	1
7	0.89	0.96	3	179	8	1	0	0	management	high	0.012	Low Risk	1
8	0.89	0.96	3	179	8	1	0	0	management	high	0.012	Low Risk	1
9	0.55	0.81	3	239	8	1	0	0	accounting	high	0.004	Low Risk	1
10	0.55	0.81	3	239	8	1	0	0	accounting	high	0.004	Low Risk	1

Figure 17.5: Top 10 Current Low-Risk Employees by Impact Score and Turnover Risk

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary	probability_leaving	turnover_risk	impact_score
1	0.74	0.95	5	266	10	1	0	1	management	high	0.214	Low-Moderate Risk	1.0
2	0.74	0.95	5	266	10	1	0	1	management	high	0.214	Low-Moderate Risk	1.0
3	0.77	0.89	4	269	10	1	0	0	management	high	0.156	Low-Moderate Risk	1.0
4	0.77	0.89	4	269	10	1	0	0	management	high	0.156	Low-Moderate Risk	1.0
5	0.81	0.90	4	270	10	1	0	0	accounting	medium	0.230	Low-Moderate Risk	0.9
6	0.81	0.90	4	270	10	1	0	0	accounting	medium	0.230	Low-Moderate Risk	0.9
7	0.88	0.83	4	273	10	1	0	0	sales	medium	0.186	Low-Moderate Risk	0.9
8	0.88	0.83	4	273	10	1	0	0	sales	medium	0.186	Low-Moderate Risk	0.9
9	0.88	0.99	3	190	5	1	0	0	technical	high	0.218	Low-Moderate Risk	0.8
10	0.67	0.85	3	160	4	1	0	0	hr	medium	0.296	Low-Moderate Risk	0.7

Figure 17.6: Top 10 Current Low-Moderate Risk Employees by Impact Score and Turnover Risk

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary	probability_leaving	turnover_risk	impact_score
1	0.90	1.00	2	114	5	1	0	0	support	high	0.362	Moderate Risk	0.8
2	0.90	0.81	6	273	5	1	0	0	technical	medium	0.476	Moderate Risk	0.7
3	0.82	0.87	5	273	6	1	0	0	support	medium	0.474	Moderate Risk	0.7
4	0.81	0.98	5	243	6	1	0	0	sales	medium	0.440	Moderate Risk	0.7
5	0.92	0.97	4	238	5	2	0	0	support	medium	0.402	Moderate Risk	0.7
6	0.97	0.96	4	250	6	1	0	0	RandD	medium	0.346	Moderate Risk	0.7
7	0.84	0.84	6	261	5	1	0	0	product_mng	low	0.496	Moderate Risk	0.6
8	0.73	0.83	5	266	5	1	0	0	sales	low	0.462	Moderate Risk	0.6
9	0.77	0.85	5	221	5	1	0	0	technical	low	0.444	Moderate Risk	0.6
10	0.70	0.84	6	225	6	1	0	0	accounting	low	0.432	Moderate Risk	0.6

Figure 17.7: Top 10 Current Moderate Risk Employees by Impact Score and Turnover Risk

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary	probability_leaving	turnover_risk	impact_score
1	0.79	0.87	4	223	5	1	0	0	sales	medium	0.980	High Risk	0.7
2	0.79	0.87	4	223	5	1	0	0	sales	medium	0.980	High Risk	0.7
3	0.85	0.96	4	240	6	1	0	0	technical	medium	0.974	High Risk	0.7
4	0.90	0.87	4	231	5	1	0	0	management	low	0.970	High Risk	0.6
5	0.80	0.99	4	255	5	2	0	0	technical	low	0.934	High Risk	0.6
6	0.86	0.96	5	238	5	1	0	0	technical	low	0.502	High Risk	0.6
7	0.39	0.57	2	132	3	1	0	0	support	low	1.000	High Risk	0.0
8	0.42	0.50	2	151	3	1	0	0	sales	low	1.000	High Risk	0.0

Figure 17.8: Top 10 Current High Risk Employees by Impact Score and Turnover Risk

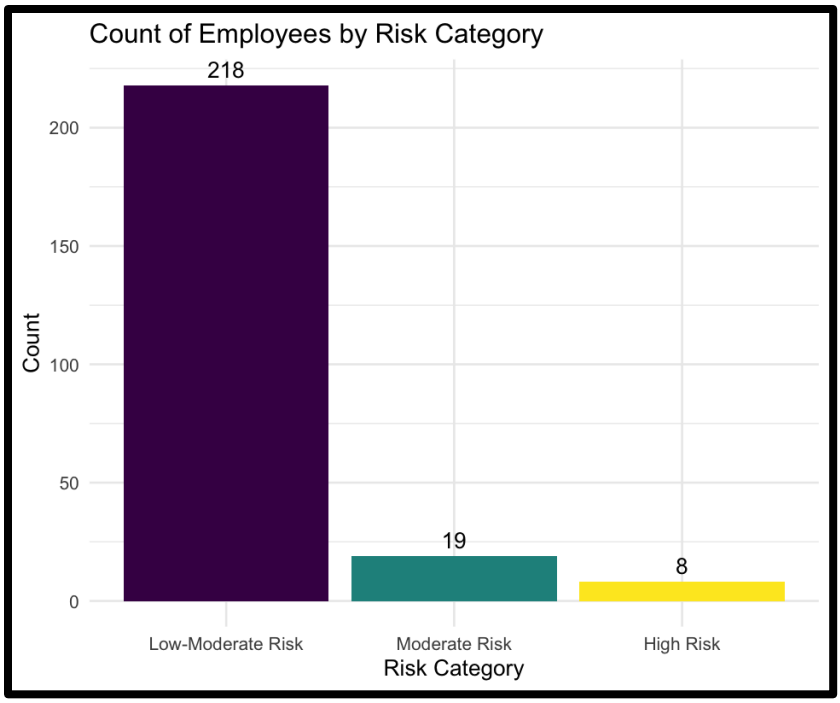

Figure 17.9: Count of Employees by Risk Category (Low Risk Excluded)

The analysis revealed the following distribution of employees across the risk categories ***(Figure 17.9)***:

- **High Risk**: 8 employees.
- **Moderate Risk**: 19 employees.
- **Low-Moderate Risk**: 218 employees.
- **Low Risk**: 11,183 employees.

This categorization and prioritization empower HR to focus their efforts on employees who are both at risk of leaving and whose departure would have the most significant impact on the organization.

Section 18: Post Model Implementation Data Exploration

The goal of this section is to explore the patterns present in employees with 0% likelihood of leaving the company across different variables to identify the range of values that lead to the highest retention.

18.1 Summary Statistics and Histograms

```
> summary(safe_employees)
 satisfaction_level last_evaluation  number_project  average_montly_hours time_spend_company Work_accident left     promotion_last_5years
 Min.   :0.1300     Min.   :0.3600   Min.   :2.000   Min.   : 96.0        Min.   : 2.000     0:3469        0:4244   0:4156
 1st Qu.:0.6200     1st Qu.:0.5900   1st Qu.:3.000   1st Qu.:166.0        1st Qu.: 2.000     1: 775        1:   0   1:  88
 Median :0.7300     Median :0.7000   Median :4.000   Median :193.0        Median : 3.000
 Mean   :0.7403     Mean   :0.7101   Mean   :3.689   Mean   :196.5        Mean   : 2.905
 3rd Qu.:0.8800     3rd Qu.:0.8400   3rd Qu.:4.000   3rd Qu.:229.0        3rd Qu.: 3.000
 Max.   :1.0000     Max.   :1.0000   Max.   :6.000   Max.   :285.0        Max.   :10.000

       department      salary     probability_leaving turnover_risk  impact_score
 sales      :1137   high  : 456   Min.   :0           Min.   :0      Min.   :0.0000
 technical  : 749   low   :1734   1st Qu.:0           1st Qu.:0      1st Qu.:0.2000
 support    : 597   medium:2054   Median :0           Median :0      Median :0.3000
 IT         : 360                 Mean   :0           Mean   :0      Mean   :0.2985
 marketing  : 310                 3rd Qu.:0           3rd Qu.:0      3rd Qu.:0.4000
 product_mng: 307                 Max.   :0           Max.   :0      Max.   :1.0000
 (Other)    : 784
```

Figure 18.1: Summary Statistics for No-Risk Employees

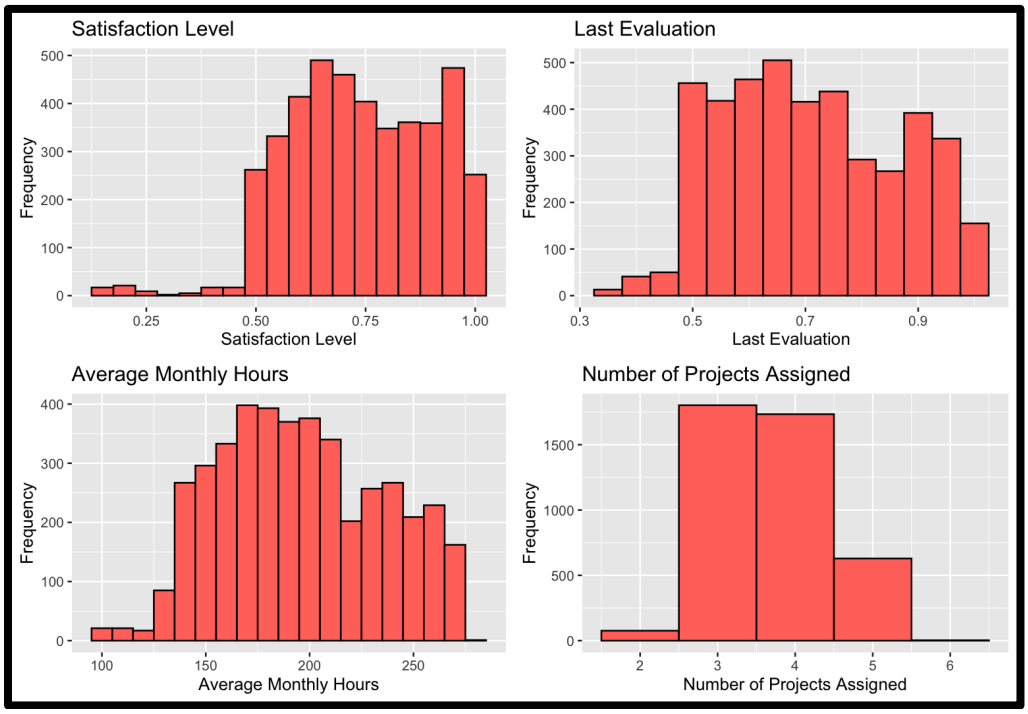

Figure 18.2: Histograms of Key Employee Metrics for No-Risk Employees

As it is provided by the summary statistics table ***(Figure 18.1)*** and histograms ***(Figure 18.2)***, it is possible to observe that the safest ranges for employees to be in terms of satisfaction level is between 0.62 and 1, for last evaluation it is between 0.59 and 1, for average monthly hours it is between 166 and 229, and for number of projects, between 3 and 4.

18.2 Salary and Promotions Comparison

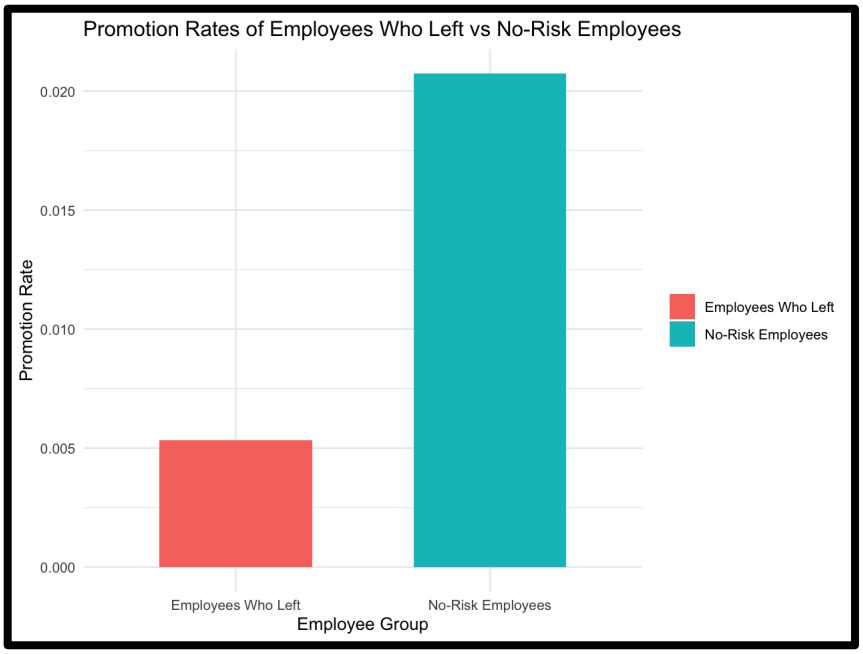

Figure 18.3: Histograms of Promotion Rates for Employees Who left and No-Risk Employees

In ***Figure 18.3***, it is possible to see that though promotion rates are extremely low for both groups, no-risk employees have 4 times higher promotion rates, suggesting, again, that increasing the company's promotion rates can significantly strengthen retention.

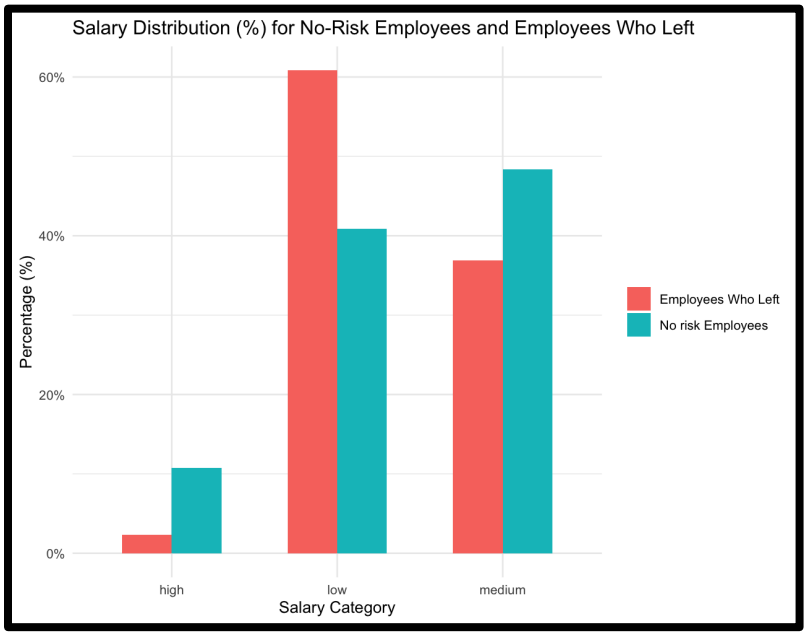

Figure 18.4: Salary Distribution for No-Risk Employees and Employees Who Left

By analyzing the percentages of salary distribution between the group of employees who left and the group of employees who offer no risk of leaving ***(Figure 18.4)***, it is possible to see that the no-risk employees, have a higher proportion of higher salaries then those who left.

18.3 Satisfaction Improvement Analysis

Based on insights from earlier analyses, it was concluded that improving workforce satisfaction is critical to maximizing employee retention. To achieve this, identifying the factors that influence employee satisfaction is essential. For this purpose, the Boruta algorithm was applied to the entire dataset, with satisfaction level as the target variable, to determine its most significant predictors. The variables 'left' and time_spend_company were excluded due to their redundancy and limited analytical value. Knowing that an employee's decision to stay or leave the company affects satisfaction, or that longer tenure correlates with satisfaction, does not provide meaningful insights, as these relationships are inherently self-evident. Also, the decision to use the entire dataset, rather than focusing only on current employees, was made to capture the broader drivers of satisfaction, encompassing both current and former employees, ensuring a more comprehensive understanding of the factors influencing satisfaction.

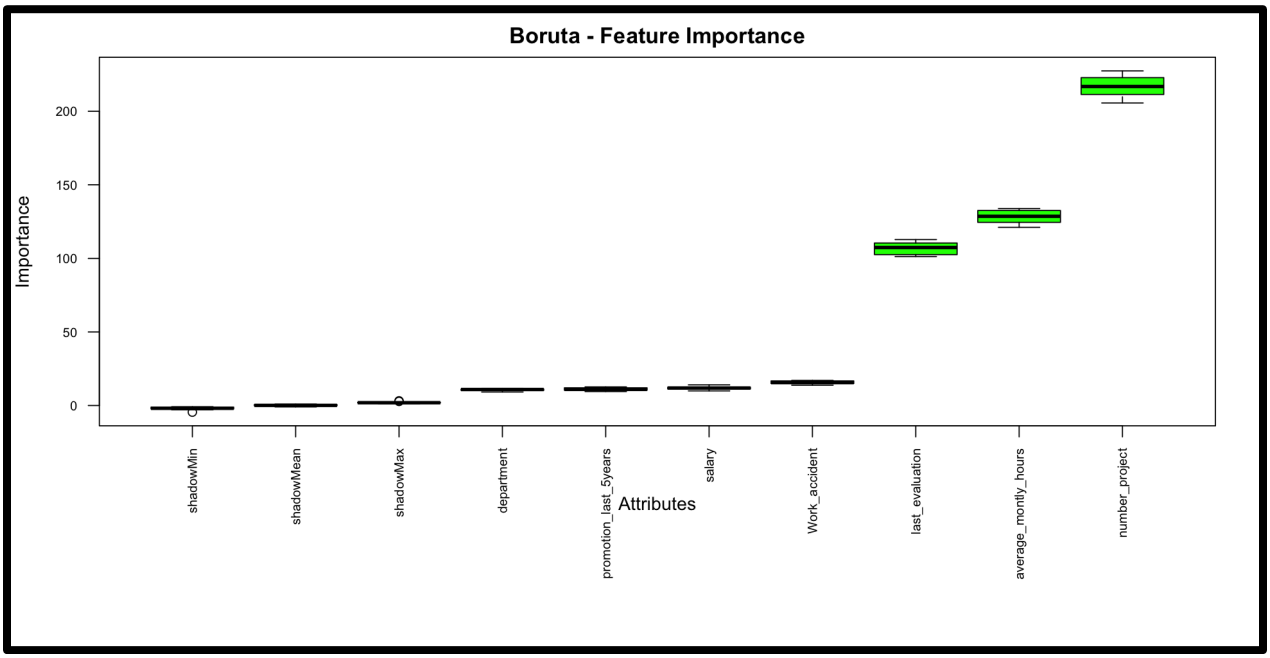

Figure 18.5: Boruta Feature Importance for Predicting Employee Satisfaction

The plot ***(Figure 18.5)*** indicates that the number of projects and average monthly hours worked are the two most significant factors affecting satisfaction. Therefore, ensuring that the values for these variables remain within the recommended ranges is crucial for achieving the company's retention goals.

Conclusion

The analysis of employee turnover has uncovered critical factors influencing satisfaction and retention. Employee satisfaction, number of projects, performance evaluation scores, and monthly hours worked emerged as the primary drivers of retention, while salary and promotion opportunities also play an important role. Keeping these variables within ranges associated with low turnover can create a balanced work environment, fostering engagement, satisfaction, and reduced turnover. This will help the company retain talent, strengthen its market position, and reduce costs linked to hiring and training.

The analysis identified distinct profiles of employees more likely to leave. Disengaged employees leave early due to a lack of involvement and underutilization. Overworked achievers, despite handling heavy workloads and numerous projects, often feel unrecognized, leading to burnout. Stagnated employees, while in good standing and managing reasonable workloads, leave seeking better opportunities. These patterns are influenced by broader systemic issues, such as limited promotions and low salaries, which affect most employees and further contribute to their final decision to leave.

The predictive models developed in this project effectively identified employees at risk of leaving. Categorizing employees by risk level and organizing them by priority has equipped the HR team with a valuable tool to address turnover proactively and efficiently.

Since satisfaction level is a key driver of retention, the analysis explored ways to enhance it. Findings revealed that managing workload and project assignments is critical for improving satisfaction, helping to foster a more supportive and engaging workplace. This approach provides actionable steps to reduce turnover while building a more committed and productive workforce.

Recommendations

In this section, the primary objective is to provide actionable recommendations to the HR department on how to maximize employee retention. With the most influential predictors of turnover identified and their optimal ranges determined, the insights from this analysis will inform HR strategies, allowing for targeted interventions that enhance employee satisfaction, productivity, and overall organizational stability. Through data-driven recommendations, the goal is to minimize turnover rates and foster a more committed and engaged workforce.

1. **Employee Satisfaction**: Satisfaction is the most significant factor affecting turnover. Ensure employee satisfaction remains above 0.6 to reduce the risk of departures.

2. **Employee Performance**: Higher-performing employees tend to stay longer. Employees with performance scores above 0.6 have shown the lowest risk of leaving. HR should intervene with employees performing below this threshold to support their development and success.

3. **Number of Projects**: The optimal number of projects assigned per employee should be between 3 and 4. Maintaining this range can prevent burnout and enhance satisfaction.

4. **Working Hours**: Keep average monthly hours between 160 and 220 to ensure employees are not overworked while maintaining productivity.

5. **Promotion Opportunities**: Provide more promotion opportunities. Employees who left had four times fewer promotions than those with no risk of leaving. Increasing promotion rates can significantly enhance retention.

6. **Salary Improvements**: Higher salaries are more common among employees at no risk of leaving. Improving salary structures will help retain talent, particularly among high-performing employees.

7. **Engage with At-Risk Employees**: Conduct one-on-one meetings with employees identified as at risk of leaving to understand their concerns and explore potential solutions that may encourage them to stay.