PREDICTING EMPLOYEE TURNOVER: AN HR ANALYTICS APPROACH USING LOGISTIC REGRESSION IN R

Project submitted to ST. Teresa's College (Autonomous), Ernakulam, affiliated to Mahatma Gandhi University in partial completion of

PGDM-BUSINESS ANALYTICS

Submitted by

NAYANA B JOSEPH

REGISTRATION NO: SM23PGDM010

Under the supervision and guidance of

DR. SUNITHA T.R

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

COLLEGE WITH POTENTIAL FOR EXCELLENCE

Nationally Re-Accredited at A++ Level (Fourth Cycle)

Affiliated to Mahatma Gandhi University Kottayam

Valued.
Dr. Nimitha Aboobaker
SMS, CUSAT

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the project entitled **" Predicting Employee Turnover: An HR Analytics Approach Using Logistic Regression in R"** is a bonafide record of the project work carried out by **Ms. NAYANA B JOSEPH (Reg No : SM23PGDM010)** final year student of **PGDM-BUSINESS ANALYTICS** under my supervision and guidance during 2023-2025. The project report represents the work of the candidate and is hereby approved for submission.

5/12/24

Dr Sunitha T.R

Asst. Professor

Dept. of Management Studies

Countersigned

Principal

DECLARATION

I hereby declare that the project work entitled **" Predicting Employee Turnover: An HR Analytics Approach Using Logistic Regression in R"** submitted to **ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM** is a record of an original work done by me under the guidance of **Dr. SUNITHA T.R, Asst. Professor, ST. Teresa's College, Ernakulam**, and this project work is submitted in the partial fulfilment of the requirements for the aware of the degree of **PGDM – Business Analytics.** The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree in diploma.

Place : Ernakulam

Date : 04-12-2024

Nayana

NAYANA B JOSEPH

ACKNOWLEDGEMENT

An undertaking of work life – this is never an outcome of a single person, rather it bears the imprints of some people who directly or indirectly helped me in completing the present study. I would be falling in my duties if I don't say a word of thanks to all those who made my training period educative and pleasurable.

First of all, I thank almighty God for his mercy and love which kept me in good health and sound mind and helped me to complete the project work successfully and gave me strength and inspiration for making this project work a great success.

I express my sincere gratitude to our Director Rev. Sr. Emeline CSST. I thank Dr. Alphonsa Vijaya Joseph, Principal, ST. Teresa's College (Autonomous), Ernakulam for her valuable support and encouragement.

I am grateful to Mrs Namitha Peter, Head of the Department of Management Studies, and all other members of the faculty of the Department for all the support and help given to me in the preparation of this project. I must also thank my faculty guide Dr. Sunitha T.R, ST. Teresa's College, Ernakulam, for her continuous support, mellow criticism and able directional guidance during the project.

Finally, I would like to thank all lecturers, friends, and my family for their kind support and all who have directly or indirectly helped me in preparing this project report. And at last, I am thankful to all divine light and my parents, who kept my motivation and zest for knowledge always high through the tides of time.

NAYANA B JOSEPH

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

HR analytics has emerged as a powerful tool for understanding and addressing workforce challenges. By leveraging employee data, organizations can gain insights into the factors that contribute to turnover and develop strategies to improve retention. Predictive modeling—particularly the use of logistic regression—is one of the most effective methods for forecasting employee turnover based on key indicators such as job satisfaction, tenure, salary, workload, and performance.

HR analytics helps organizations optimise human resource management. It helps answers questions like what factors contribute to employee turnover and what specific employees are at a high risk of leaving an organization soon, amongst others. The unplanned departure of employees often results in substantial direct and indirect costs, meaning even a modest reduction in turnover can lead to significant savings, which would be highly valued by company management

Predictive analytics in HR utilizes data and AI to improve decision-making by identifying employees at risk of leaving and refining retention strategies. By examining trends in historical data, organizations can create proactive, customized solutions that effectively reduce turnover and cultivate a more stable, satisfied workforce. Predictive analytics can be used to calculate Employee Satisfaction scores, offering a comprehensive overview of areas that are positively received and highlighting those that require improvement. It can also assist in building a predictive model by utilizing factors such as job tenure, monthly engagement scores, and recent promotion history to generate a turnover risk score for each employee. Apply anomaly detection algorithms to performance metrics to pinpoint significant changes, such as a decline in productivity for a top performer by a specific percentage. Early identification enables managers to have constructive conversations with these employees, potentially addressing their concerns and preventing possible resignations.

This project aims to develop a predictive model using logistic regression to identify employees who are at risk of leaving the company. Logistic regression is well-suited for this task, as it models the probability of a binary outcome (in this case, whether an employee will stay or leave). By analyzing various employee attributes and behaviors, the model can generate insights that will allow HR teams to implement targeted retention strategies and minimize turnover.

To assess which factors and the extent of their impact on employee departures. this analysis employs logistic regression. The logistic regression model will assist in identifying the most significant variables in predicting employee turnover.

1.2 STATEMENT OF THE PROBLEM

Employee turnover is a significant challenge for organizations. as it leads to disruptions in workflow. increased recruitment and training costs. and potential declines in morale and productivity. HR analytics enables organizations to make better decisions that impact employees and the organization using the evidence data reveals. Uncover and remedy inefficiencies to improve employee and organizational productivity and reduce costs. With HR analytics. organizations can move beyond simply reporting past data to predict future outcomes. understand what drives employee engagement and retention is crucial for a healthy work environment. Despite advancements in HR analytics. many organizations still struggle to accurately predict which employees are at risk of turnover. leading to missed opportunities for intervention.

Understanding the factors that contribute to employee departures is crucial for Human Resources (HR) departments to develop strategies that can improve retention. The goal of this project is to identify the key factors influencing employees' decisions to leave the company. based on available data including job satisfaction, performance evaluations, workload. tenure. accidents. promotions. department, and salary. By analysing these factors, the HR team can devise data-driven interventions to reduce employee turnover and improve overall employee satisfaction and retention.

By leveraging logistic regression, this study aims to analyse various employee attributes and behaviours that contribute to employee turnover and to provide actionable insights using predictive capabilities to help HR departments minimize employee departures and their associated costs.

1.3 SIGNIFICANCE OF THE STUDY

The significance of this study lies in its ability to help HR Departments understand and address the high employee turnover it is currently facing. By analyzing key factors such as employee satisfaction. work hours. promotion history. and department-specific dynamics. the predictive model can pinpoint the main drivers of employee departures. This enables HR to develop

targeted strategies to retain talent, enhance job satisfaction, and improve the overall work environment. Furthermore, reducing turnover can lower recruitment and training costs while maintaining productivity, ultimately leading to a more stable and engaged workforce, which is critical for the company's long-term success.

1.4 SCOPE OF THE STUDY

This scope of this study focuses on understanding the key factors influencing employee turnover in a company by analyzing employee survey data. The objective is to predict which employees are likely to leave the company using a machine learning model. The dataset includes various factors that potentially affect employee retention, such as satisfaction levels, performance evaluations, number of projects, average monthly hours worked, years spend at the company, work accidents, promotions in the last five years, department, and salary level.

The primary methodology employed will be logistic regression, a statistical technique well-suited for predicting binary outcomes (i.e., whether an employee will stay or leave). The study will involve identifying and selecting relevant predictor variables that are statistically significant in forecasting employee turnover.

The scope includes examining various factors that may influence employee turnover, such as job satisfaction, workload, work accidents, previous evaluations, career development opportunities, and salary package.

The scope will include suggestions for future research avenues based on the findings, such as exploring other predictive modeling techniques, investigating additional factors influencing turnover, or conducting longitudinal studies to assess the effectiveness of implemented strategies over time.

This study can serve as a foundation for designing retention strategies, improving work-life balance, and fostering a positive work environment that encourages long-term employee engagement.

1.5 OBJECTIVES OF THE STUDY

1. To develop a logistic regression model that accurately predicts whether an employee will leave the company based on various factors.

2. To identify key predictors of employee turnover- factors like satisfaction levels, number of projects, average monthly hours, performance evaluations, and salary to understand their impact on employee retention.
3. To assess the model's effectiveness by calculating the predictive accuracy of the logistic regression model using appropriate metrics (e.g., accuracy, precision, recall) to determine its reliability in predicting employee turnover.

1.6 METHODOLOGY

To investigate the factors contributing to employee turnover, a comprehensive research methodology was employed. The study began with the collection of a dataset containing various employee attributes, including satisfactory levels, last evaluation scores, number of projects, average monthly hours, tenure at the company, work accident history, promotion status over the past five years, departmental affiliation, and salary levels.

The logistic regression model will be fitted to the training data using R, statistical software to estimate the coefficients for each independent variable.

Research Design

The study will employ a quantitative research design, utilizing statistical methods to analyze employee data and develop a predictive model for turnover.

Data Collection

Historical employee data was collected from an organization's HR information system (HRIS) using the public data repository for analytics projects.

Data Preprocessing

- Data Cleaning:
  The collected data will be cleaned to remove duplicates, address missing values, and correct inaccuracies.
- Feature Engineering:
  New variables may be created from existing data (e.g., categorizing job tenure into ranges) to enhance the predictive power of the model.

Dependent Variable:

The primary dependent variable will be employee turnover, defined as a binary outcome (1 = employee left, 0 = employee retained).

Independent Variables:

- Satisfactory Levels- The employee's self-reported satisfaction level [0-1]
- Last Evaluation - Score of employee's last performance review [0–1]
- Number Project- Number of projects employee contributes to
- Average Monthly Hour- Average number of hours employee worked per month
- Time Spent- How long the employee has been with the company (years)
- Work Accident- Whether or not the employee experienced an accident while at work
- Promotion last 5 years- Whether or not the employee was promoted in the last 5 years
- Department- The employee's department
- Salary- The employee's salary (low, medium, or high)
- Left - Whether or not the employee left the company

Model Development Logistic

Regression:

Logistic regression will be used to model the relationship between the independent variables and the probability of employee turnover.

The model will be trained using a portion of the dataset (training set) while reserving a separate portion (testing set) for model validation.

1.7 SOFTWARE USED FOR ANALYSIS

R PROGRAMMING

R is a powerful and flexible programming language primarily used for statistical analysis, data visualization, and data science. Originating in the early 1990s as an open-source implementation of the S programming language, R has grown to become a cornerstone in the fields of statistics and data analytics due to its rich ecosystem and extensive libraries. The language excels at data manipulation and transformation, with packages like dplyr facilitating streamlined workflows for filtering, summarizing, and aggregating data. R's visualization capabilities are particularly noteworthy, with the ggplot2 package allowing users to create sophisticated and customizable graphics based on a layered approach to data representation. This enables users to convey complex information clearly and effectively. In addition to its

manipulation and visualization prowess. R offers a wide range of statistical tools, from basic descriptive statistics to advanced techniques like regression analysis, time series forecasting, and machine learning. The Comprehensive R Archive Network (CRAN) boasts thousands of packages that cater to various domains, including biostatistics, social sciences, finance, and machine learning, further enhancing R's versatility. Furthermore, R's integration with other programming languages and platforms, such as Python and SQL, allows for seamless interoperability in diverse data workflows. Its active community contributes to a wealth of resources, including tutorials, forums, and user-contributed packages, fostering continuous learning and development. With its user-friendly syntax and powerful capabilities, R remains an indispensable tool for statisticians, data analysts, and researchers seeking to extract insights from complex datasets and communicate their findings effectively.

- **Statistical Analysis**: R is designed for statistical analysis and data visualization, making it a go-to for statisticians and data scientists.
- **Open Source**: R is free to use and has a large community contributing to its development and packages.
- **Rich Ecosystem**: R has a vast array of packages available through CRAN, Bioconductor, and GitHub for various applications, including machine learning, data manipulation, and more.
- **Data Manipulation**: Libraries like dplyr and tidyr simplify data cleaning and transformation tasks.
- **Visualization**: Powerful libraries like ggplot2 provide advanced tools for creating informative and attractive visualizations.
- **Reproducibility**: R supports reproducible research through RMarkdown, allowing you to combine code, output, and narrative.
- **Interactivity**: Packages like shiny enable the development of interactive web applications directly from R.

CHAPTER 2

INDUSTRY PROFILE

2.1 INDUSTRY PROFILE

HUMAN RESOURCE FUNCTION IN ORGANIZATIONS

Human Resource (HR) in the industry plays a crucial role in managing the workforce and ensuring that an organization operates efficiently and effectively. Here are some key aspects of HR in the industry:

Key Functions of HR

1. **Recruitment and Staffing**:
   - Attracting, selecting, and onboarding new employees.
   - Developing job descriptions and specifications.
   - Utilizing various recruitment channels (job boards, social media, recruitment agencies).
2. **Training and Development**:
   - Providing employee training programs to enhance skills and knowledge.
   - Offering career development opportunities and succession planning.
   - Ensuring compliance with industry regulations and standards through ongoing education.
3. **Performance Management**:
   - Setting performance standards and conducting evaluations.
   - Providing feedback and coaching to employees.

- Implementing reward and recognition programs to motivate staff.

4. **Employee Relations**:
- Managing relationships between employees and management.
- Addressing grievances and resolving conflicts.
- Promoting a positive workplace culture and employee engagement.

5. **Compensation and Benefits**:
- Designing competitive salary structures and benefit packages.
- Conducting market research to ensure fair compensation.
- Managing payroll and employee benefits programs.

6. **Compliance and Legal Issues**:
- Ensuring adherence to labor laws and regulations.
- Handling workplace safety and health standards.
- Managing policies related to diversity, equity, and inclusion.

7. **Strategic Planning**:
- Aligning HR strategy with organizational goals.
- Analyzing workforce data to inform decision-making.
- Forecasting future staffing needs based on business trends.

Current Trends in HR

1. **Technology Integration**:
- Utilizing HR software and systems for recruitment, payroll, and employee management.
- Implementing AI and data analytics to improve decision-making.

2. **Remote Work and Flexibility**:
- Adapting to the rise of remote and hybrid work models.
- Offering flexible working hours to improve work-life balance.

3. **Employee Wellness and Mental Health**:
- Focusing on employee well-being programs, including mental health support.
- Creating a supportive environment that prioritizes work-life balance.

4. **Diversity and Inclusion**:
- Implementing initiatives to promote diversity in hiring and company culture.
- Ensuring that all employees feel valued and included.

5. **Continuous Learning**:
- Encouraging a culture of lifelong learning and development.

- Providing resources for skills training in response to changing industry needs.

Challenges Facing HR in the Industry

1. **High Turnover Rates**: Retaining talent can be difficult, especially in competitive industries.
2. **Skill Shortages**: Finding qualified candidates for specialized roles is often a challenge.
3. **Changing Workforce Demographics**: Adapting to the needs of a diverse workforce, including millennials and Gen Z.
4. **Regulatory Changes**: Keeping up with evolving labor laws and compliance requirements.
5. **Technology Adoption**: Navigating the complexities of integrating new technologies into HR practices.

HR plays a pivotal role in the success of any organization. By focusing on strategic management of human capital, HR professionals can enhance employee satisfaction, improve retention, and drive overall organizational performance. In a rapidly changing business environment, staying adaptable and innovative is essential for effective HR management.

TECHNOLOGIES IN HUMAN RESOURCE

Technologies in human resources (HR) have transformed the way organizations manage their workforce, enhancing efficiency and decision-making. Here are some key technologies used in HR:

1. **Applicant Tracking Systems (ATS)**
   - Streamlines the recruitment process by managing job postings, applications, and candidate communication.
   - Helps HR teams track applicants through various stages of the hiring process.
2. **Human Resource Information Systems (HRIS)**
   - Centralizes employee data, including personal information, payroll, benefits, and performance records.
   - Facilitates reporting and compliance by providing easy access to HR metrics.
3. **Performance Management Software**

- Enables goal setting, performance tracking, and feedback processes.
- Supports continuous performance management through regular check-ins and reviews.

4. **Learning Management Systems (LMS)**
   - Provides training and development resources for employees.
   - Tracks training completion and employee progress, allowing for personalized learning experiences.
5. **Employee Engagement Tools**
   - Surveys and feedback platforms help measure employee satisfaction and engagement levels.
   - Features like pulse surveys and feedback tools allow for real-time insights.
6. **Payroll Software**
   - Automates payroll processing, tax calculations, and compliance with labor laws.
   - Ensures timely and accurate payments to employees, reducing administrative burdens.
7. **Time and Attendance Systems**
   - Tracks employee attendance, work hours, and leave requests.
   - Integrates with payroll systems to streamline compensation processes.
8. **Cloud-based HR Solutions**
   - Offers flexibility and accessibility, allowing HR teams to manage operations from anywhere.
   - Facilitates collaboration among remote teams and employees.
9. **Artificial Intelligence (AI) and Machine Learning**
   - Enhances recruitment processes by screening resumes and predicting candidate success.
   - Analyzes employee data to identify trends and inform decision-making.
10. **Chatbots and Virtual Assistants**

- Provides instant answers to employee queries regarding HR policies, benefits, and procedures.
- Improves communication and reduces the workload on HR teams.

11. Data Analytics and Business Intelligence Tools

- Analyzes HR data to provide insights into workforce trends, turnover rates, and employee performance.
- Supports strategic decision-making by visualizing data in dashboards and reports.

12. Video Interview Platforms

- Facilitates remote interviewing through video conferencing tools.
- Enhances the candidate experience and expands the talent pool beyond geographical limitations.

13. Onboarding Software

- Streamlines the onboarding process for new hires with digital checklists and resources.
- Ensures a smooth transition into the organization, improving retention rates.

14. Wellness and Employee Assistance Programs (EAPs)

- Provides resources and support for employee mental health and well-being.
- Incorporates tools for wellness tracking and initiatives.

The integration of technology in HR has led to more efficient processes, better employee experiences, and enhanced decision-making capabilities. By leveraging these tools, HR professionals can focus more on strategic initiatives and less on administrative tasks, ultimately contributing to organizational success.

ANALYTICS IN HR: EMPLOYEE TURNOVER PREDICTION

Employee turnover prediction in HR analytics involves using data-driven methods to identify which employees are at risk of leaving the organization. By collecting and analyzing various data points—such as employee demographics, performance metrics, job satisfaction scores, workload, and promotion history—HR can uncover patterns and factors that contribute to turnover. Techniques like regression analysis, decision trees, and machine learning algorithms

help predict future turnover, allowing organizations to implement proactive retention strategies. These insights enable HR to address issues before they lead to departures, ultimately reducing costs associated with recruitment and training, improving employee engagement, and fostering a more stable and satisfied workforce. By leveraging analytics, HR professionals can make informed, strategic decisions that enhance overall organizational effectiveness.

Application of analytics in Human Resources in a company

Analytics in HR has numerous applications that enhance decision-making, improve workforce management, and drive organizational success. Here are some key applications:

1. **Employee Turnover Prediction**
   - Analyzing historical data to identify factors that lead to employee attrition. Predictive models help HR proactively address retention issues and develop targeted interventions.
2. **Recruitment Optimization**
   - Utilizing data analytics to refine recruitment processes, identify successful sourcing channels, and assess candidate fit. Metrics such as time-to-fill and quality-of-hire can be analyzed to improve hiring strategies.
3. **Performance Management**
   - Leveraging analytics to evaluate employee performance, set benchmarks, and provide actionable feedback. Continuous performance data can inform training needs and development plans.
4. **Employee Engagement Analysis**
   - Analyzing engagement survey results to understand employee sentiment, identify areas for improvement, and track the effectiveness of engagement initiatives over time.
5. **Learning and Development**
   - Assessing training effectiveness through analytics to determine which programs yield the best outcomes. Data-driven insights can guide the development of tailored learning paths for employees.
6. **Workforce Planning**

- Using analytics to forecast future workforce needs based on business trends and organizational goals. This ensures that the right talent is available at the right time.

7. **Compensation and Benefits Analysis**

- Evaluating compensation structures and employee benefits to ensure competitiveness and fairness. Analytics can help identify discrepancies and inform adjustments.

8. **Diversity and Inclusion Metrics**

- Tracking diversity metrics and analyzing recruitment and retention data to promote equity in the workplace. This helps organizations assess the effectiveness of their diversity initiatives.

9. **Succession Planning**

- Analyzing employee performance and potential to identify future leaders within the organization. Data-driven succession planning supports career development and ensures leadership continuity.

10. **Sentiment Analysis**

- Utilizing text analysis on employee feedback and survey comments to gauge overall sentiment. This helps HR teams understand employee concerns and areas needing attention.

EMPLOYEE TURNOVER PREDICTION IN HUMAN RESOURCES

Employee turnover prediction in HR involves using data analytics to identify employees who are at risk of leaving the organization. This predictive approach allows HR to take proactive measures to improve retention and enhance employee engagement. Here's a closer look at the process and its significance:

Process of Employee Turnover Prediction

1. **Data Collection**:
    - Gather data from various sources, including employee demographics, performance metrics, satisfaction surveys, tenure, and compensation details.

2. **Data Analysis**:

- Use statistical techniques and machine learning algorithms to analyze the data. Common methods include regression analysis, decision trees, and clustering to identify patterns and correlations.

3. **Identifying Predictors**:

   - Determine key indicators of turnover, such as low job satisfaction, lack of promotions, high workload, and short tenure. These factors often signal potential attrition.

4. **Model Development**:

   - Build predictive models based on historical data to forecast which employees are most likely to leave. Validate the model using techniques like cross-validation to ensure accuracy.

5. **Intervention Strategies**:

   - Use insights from the model to develop targeted retention strategies, such as tailored engagement programs, career development opportunities, and recognition initiatives.

6. **Monitoring and Adjustment**:

   - Continuously monitor turnover rates and the effectiveness of retention strategies, adjusting the model and interventions as needed based on new data and feedback.

Significance of Turnover Prediction

- **Proactive Retention**: By identifying at-risk employees early, HR can intervene before turnover occurs, potentially saving costs associated with recruitment and training.
- **Improved Employee Engagement**: Understanding the factors contributing to turnover allows organizations to create a more satisfying work environment.
- **Data-Driven Decision Making**: Leveraging analytics provides a scientific basis for HR strategies, helping organizations make informed decisions about workforce management.

- **Cost Savings**: Reducing turnover translates to significant savings in hiring and onboarding expenses, as well as minimizing disruptions to team performance.

Employee turnover prediction is a vital application of HR analytics that empowers organizations to understand and mitigate the factors leading to employee attrition. By leveraging data effectively, HR can foster a more engaged and stable workforce, ultimately enhancing organizational performance and culture.

CHAPTER 3

THEORETICAL FRAMEWORK

3.1 THEORETICAL FRAMEWORK FOR EMPLOYEE TURNOVER PREDICTION USING LOGISTIC REGRESSION

Regression theory provides a framework for understanding relationships between variables, making it particularly useful in predicting outcomes based on various predictors. In the context of this study, the primary goal is to analyze factors contributing to employee turnover and predict the likelihood that an employee will leave the company.

1. Concept of Regression Analysis

At its core, regression analysis seeks to establish a relationship between a dependent variable and one or more independent variables. In this study, the dependent variable is binary—indicating whether an employee has left the company (Left), while the independent variables encompass various factors such as employee satisfaction, performance evaluations, workload, tenure, and promotions.

2. Logistic Regression

Given the binary nature of the outcome, logistic regression is the most appropriate model. Unlike linear regression, which predicts continuous outcomes, logistic regression estimates the probability of a categorical outcome. It employs the logistic function to map predicted values to a range between 0 and 1, allowing for a straightforward interpretation of the likelihood of turnover.

3. Model Interpretation

The logistic regression model generates coefficients for each independent variable, reflecting their respective influence on the log-odds of the dependent variable. A positive coefficient indicates an increased likelihood of leaving the company with an increase in the predictor, while a negative coefficient suggests a protective effect against turnover. This interpretative power enables HR to identify key factors affecting employee retention.

4. Statistical Significance and Predictive Power

Regression theory also encompasses the assessment of statistical significance, allowing researchers to determine which predictors have a meaningful impact on the outcome. By evaluating the model's overall fit and predictive accuracy, such as through metrics like the area under the ROC curve, the effectiveness of the model can be gauged.

5. Practical Application

The insights derived from regression analysis enable organizations to devise targeted interventions to enhance employee satisfaction and retention. By identifying critical factors influencing turnover, HR can implement policies and practices tailored to address the specific needs and concerns of employees, ultimately fostering a more engaged and stable workforce.

In summary, regression theory serves as the backbone for this study, facilitating a nuanced understanding of employee turnover dynamics and empowering HR to make data-driven decisions aimed at improving retention and job satisfaction.

Logistic Regression

Logistic regression is a powerful statistical method used to model and analyze binary outcomes, making it particularly suitable for studying employee turnover in this dataset. The primary objective is to predict whether an employee will leave the company (Left: 1) or not (Left: 0) based on various independent factors.

1. Binary Outcome Variable

In this study, the outcome variable, Left, is binary, representing two possible states: an employee has left (1) or has not left (0). This binary nature makes logistic regression the ideal choice, as it effectively handles categorical dependent variables.

2. Predictor Variables

The model incorporates multiple independent variables believed to influence employee turnover, including:

- **Satisfactory Levels**: Reflects how content an employee is with their job, on a scale from 0 to 1. Higher satisfaction is expected to correlate with lower turnover.

- **Last Evaluation**: The score from the last performance review, also on a scale from 0 to 1. Higher evaluation scores may suggest better job security and satisfaction.
- **Number of Projects**: The count of projects an employee is involved in, which can indicate their engagement and role within the company.
- **Average Monthly Hours**: Represents the average number of hours worked each month. Excessive hours can lead to burnout, potentially increasing turnover.
- **Time Spent at Company**: The number of years an employee has been with the company, where longer tenure typically suggests greater loyalty.
- **Work Accident**: A binary variable indicating whether the employee experienced a work-related accident. Such incidents may negatively impact job satisfaction.
- **Promotion Last 5 Years**: Indicates whether an employee has received a promotion in the past five years, which may enhance their commitment to the organization.
- **Department**: The specific department the employee works in, which can affect turnover dynamics.
- **Salary**: Categorical data indicating the employee's salary level (low, medium, high), influencing their overall job satisfaction.

3. Logistic Regression Model Specification

The logistic regression model can be mathematically represented as:

P(Left=1)=11+e−(β0+β1X1+β2X2+...+βnXn)P(Left = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}P(Left=1)=1+e−(β0+β1X1+β2X2+...+βnXn)1

In this formula:

- P(Left=1)P(Left = 1)P(Left=1) is the predicted probability that an employee will leave the company.
- β0\beta_0β0 is the intercept, while β1,β2,....,βn\beta_1, \beta_2, ..., \beta_nβ1,β2,....,βn are the coefficients associated with each predictor variable X1,X2,....,XnX_1, X_2, ..., X_nX1,X2,....,Xn.

4. Coefficient Interpretation

The coefficients obtained from the model provide insight into the relationship between each predictor and the likelihood of turnover. A positive coefficient suggests that an increase in the predictor variable is associated with an increased likelihood of leaving, while a negative coefficient indicates a protective effect against turnover.

5. Model Evaluation and Applications

The effectiveness of the logistic regression model can be assessed using metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC). These evaluations help determine how well the model predicts employee departures.

By utilizing logistic regression, HR can identify key factors contributing to employee turnover, allowing for targeted interventions aimed at improving job satisfaction and retention. This data-driven approach enables the development of effective policies that can enhance employee engagement and foster a more stable workforce.

Logistics Regression Analysis in R?

Logistic regression is employed to examine and predict employee turnover within the company. The dataset includes critical variables such as employee satisfaction levels, performance evaluations, number of projects, average monthly hours worked, tenure, work accident history, promotion status, department, and salary levels. After preparing the data by handling any missing values and converting categorical variables into factors, the logistic regression model is fitted to assess the likelihood of an employee leaving the company. Each predictor's coefficient indicates its impact on turnover, with positive coefficients suggesting an increased risk of departure and negative coefficients indicating a protective effect. The model's performance is evaluated using metrics like accuracy and the area under the ROC curve (AUC), providing HR with valuable insights into the key factors driving employee turnover. These insights can guide targeted interventions aimed at improving job satisfaction and retention strategies within the organization.

Conceptual Framework for predictive modelling

Dependent Variable (Target Variable)

- Left: This binary variable indicates whether an employee has left the company (1) or not (0). It serves as the primary outcome of interest in the study, reflecting employee turnover.

2. Independent Variables (Predictors)

These variables are hypothesized to influence the likelihood of an employee leaving the company:

- Satisfactory Levels: A self-reported measure of employee satisfaction on a scale from 0 to 1. Higher satisfaction levels may correlate with lower turnover rates.
- Last Evaluation: The score from the employee's most recent performance review, also on a scale from 0 to 1. Higher evaluation scores might suggest better job performance and greater likelihood of retention.
- Number of Projects: The total number of projects the employee is involved in. A higher number of projects may indicate engagement or, conversely, potential overload, influencing turnover.
- Average Monthly Hours: The average number of hours worked each month. This variable can reflect work-life balance; excessive hours might lead to burnout and increased turnover.
- Time Spent in Company: The number of years the employee has been with the organization. Typically, longer tenure is associated with lower turnover rates, as employees may develop a stronger attachment to the company.
- Work Accident: A binary variable indicating whether the employee has experienced a workplace accident (1 for yes, 0 for no). The experience of accidents could impact employee satisfaction and turnover decisions.
- Promotion Last 5 Years: A binary variable indicating if the employee has received a promotion in the last five years. Promotions may correlate with job satisfaction and commitment to the organization.
- Department: This categorical variable represents the employee's department within the company. Different departments may have varying cultures and turnover rates, impacting overall employee retention.

- Salary: A categorical variable indicating salary level (low, medium, or high). Salary can play a crucial role in job satisfaction and employee loyalty, influencing turnover rates.

3. **Relationships Among Variables**

The conceptual framework suggests that the independent variables collectively influence the dependent variable (Left) through various mechanisms. For instance, higher satisfaction levels and better performance evaluations are likely to decrease turnover likelihood, while factors such as longer hours or lack of promotion could increase it.

3.2 PREDICTIVE MODELLING IN HR FIELD USING LOGISTIC REGRESSION

1. Create a more effective payroll system

Payroll systems can sometimes get unnecessarily complicated. Ineffective bonus strategies or issues with employees can create problems.

A simple, yet effective predictive analysis can help prevent these issues. It can make the payroll system more effective, thus saving resources and revenue.

2. Better risk management

Any business faces certain risks. How you manage them can make the difference between success and failure. Whether we're talking about security risks, external, or internal threats, you need to be one step ahead.

Predictive HR analytics may not address all the risks you face. It might not tell you when and where a hacker will strike. But it can help address many other risks.

For instance, you can forecast when and which employees will need training. By predicting a critical situation so that you can avoid it altogether.

3. Improved retention strategies

We already mentioned how important it is to reduce voluntary turnover rates. Retention strategies are how you achieve that. But they aren't always the easiest to craft.

With the correct data at hand, the HR team can revise its retention strategies. They can choose those methods that will keep all employees happy and will minimize turnover.

4. Reduce post-turnover problems

No matter how good your retention strategies are, you will sometimes have employees who resign.

Maybe it's for reasons that have nothing to do with your company, like moving to a different city or country. Or maybe they simply want something different.

Whatever the reason, turnover will not make you happy. Plus, replacing the lost employees will cost you money.

Since you can't always prevent turnover, the best you can do is to mitigate the problems it comes with. A data-driven approach through predictive analytics can help. With a predictive model, you can filter data like employee engagement, commute time, performance, and more.

This can help human resources better understand which employees are more likely to quit and when, and what variables (preventable or unpreventable) cause that attrition. By knowing that, you can avoid situations when you need to hire fast and in a panic.

Statistical assumptions and considerations

1. Dependent Variable: Binary logistic regression requires the dependent variable to be binary
2. **Independence**: The observations to be independent of each other
3. **Linearity**: logistic regression assumes linearity of independent variables and log odds of the dependent variable. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds of the dependent variable.
4. **Multicollinearity**: Check for high correlations between independent variables (e.g., Satisfactory Level and last_evaluation). High multicollinearity can inflate standard errors and make it hard to identify the individual effect of each predictor.

CHAPTER 4

DATA ANALYSIS AND INTERPRETATION

4.1 DATA PREPROCESSING

Data cleaning

Data cleaning, also known as data cleansing or data scrubbing, is the process of detecting and correcting errors and inconsistencies in data sets to improve data quality. It's a vital step in any data analysis project, as dirty data can lead to inaccurate and misleading results.

Common Data Cleaning Tasks:

1. **Handling Missing Values:**

   Deletion: Remove rows or columns with missing values.

   - **Imputation:** Fill missing values with estimated values.
     - Mean/Median Imputation: Replace missing values with the mean or median of the column.
     - Mode Imputation: Replace missing values with the most frequent value.
     - Regression Imputation: Predict missing values using regression models.
     - K-Nearest Neighbors Imputation: Fill missing values based on the values of similar data points.

2. **Detecting and Correcting Outliers:**
   - **Statistical Methods:** Use techniques like Z-scores or IQR to identify outliers.
   - **Visualization:** Use box plots or scatter plots to visualize outliers.
   - **Handling Outliers:**
     - Remove outliers if they are clearly errors.
     - Cap outliers to a certain value.
     - Winsorize outliers (replace with a specified percentile value).

3. **Identifying and Removing Duplicates:**
   - Use techniques like grouping and aggregation to identify and remove duplicate records.

4. **Data Formatting and Standardization:**
   - Ensure consistency in data formats (e.g., date, currency, text).
   - Standardize units of measurement.
   - Correct typos and inconsistencies in text data.

5. **Data Integration:**
   - Combine data from multiple sources into a single dataset.
   - Handle inconsistencies in data structures and formats.

Tools for Data Cleaning:

- **Python Libraries:** Pandas, NumPy, Scikit-learn
- **R:** dplyr, tidyr
- **Excel:** Built-in functions and Power Query
- **Specialized Data Cleaning Tools:** OpenRefine, Trifacta

Best Practices for Data Cleaning:

- **Understand the Data:** Gain a deep understanding of the data sources, structure, and potential issues.
- **Document the Cleaning Process:** Keep track of the cleaning steps and decisions made.
- **Validate the Cleaned Data:** Ensure the data is accurate and consistent after cleaning.
- **Iterative Process:** Data cleaning is often an iterative process, so be prepared to revisit and refine your cleaning steps.

Feature Engineering

- **Satisfactory Levels**: Reflects how content an employee is with their job, on a scale from 0 to 1. Higher satisfaction is expected to correlate with lower turnover.
- **Last Evaluation**: The score from the last performance review, also on a scale from 0 to 1. Higher evaluation scores may suggest better job security and satisfaction.
- **Number of Projects**: The count of projects an employee is involved in, which can indicate their engagement and role within the company.
- **Average Monthly Hours**: Represents the average number of hours worked each month. Excessive hours can lead to burnout, potentially increasing turnover.
- **Time Spent at Company**: The number of years an employee has been with the company, where longer tenure typically suggests greater loyalty.

- **Work Accident**: A binary variable indicating whether the employee experienced a work-related accident. Such incidents may negatively impact job satisfaction.
- **Promotion Last 5 Years**: Indicates whether an employee has received a promotion in the past five years, which may enhance their commitment to the organization.
- **Department**: The specific department the employee works in, which can affect turnover dynamics.

Salary: Categorical data indicating the employee's salary level (low, medium, high), influencing their overall job satisfaction.

4.2 DESCRIPTIVE ANALYSIS

4.2.1 Satisfaction level distribution

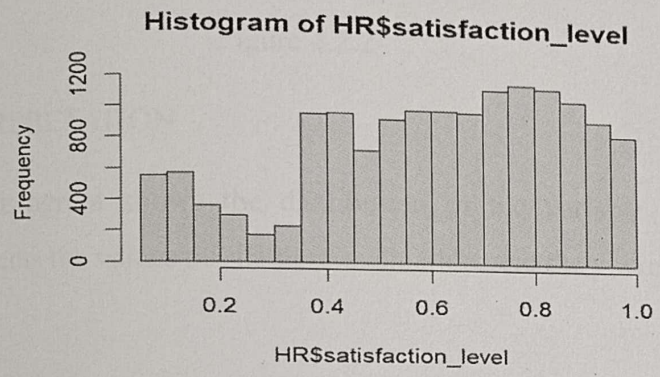


Figure 4.2.1

INTERPRETATION

The histogram shows the distribution of the variable HR$satisfaction_level. The x-axis represents the values of HR$satisfaction_level, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The distribution of HR$satisfaction_level is unimodal, meaning that there is a single peak in the distribution.
- The peak is around 0.8, suggesting that most employees have a high satisfaction level.
- The distribution is slightly skewed to the left, with a longer tail extending to the left.

This suggests that there are a few employees who have a low satisfaction level.

Overall, the histogram suggests that most employees are satisfied with their jobs, with a few employees being less satisfied. The distribution is relatively symmetric, indicating that there is no strong bias in the employee satisfaction ratings.

4.2.2 Evaluation level Distribution

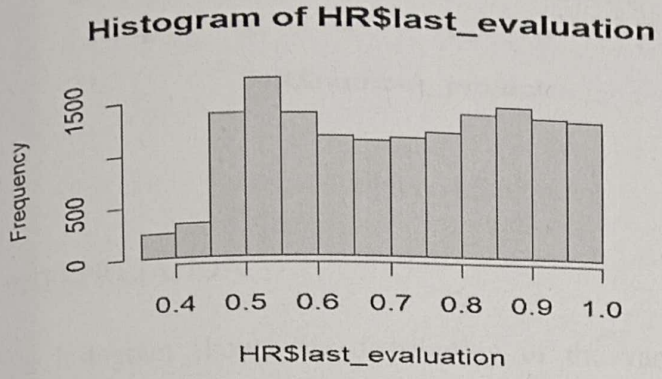


Figure 4.2.2

INTERPRETATION

The histogram shows the distribution of the variable HR$last_evaluation. The x-axis represents the values of HR$last_evaluation, and the y-axis represents the frequency of each value.

- The distribution of HR$last_evaluation is unimodal, meaning that there is a single peak in the distribution.
- The peak is around 0.6, suggesting that most employees received a performance evaluation score of around 0.6.
- The distribution is relatively symmetric, with a slight skew to the right. This suggests that there are a few employees who received very high performance evaluation scores.

Overall, the histogram suggests that most employees received a performance evaluation score of around 0.6, with a few employees receiving higher scores. The distribution is relatively symmetric, indicating that there is no strong bias in the performance evaluations.

4.2.3 Number project Distribution

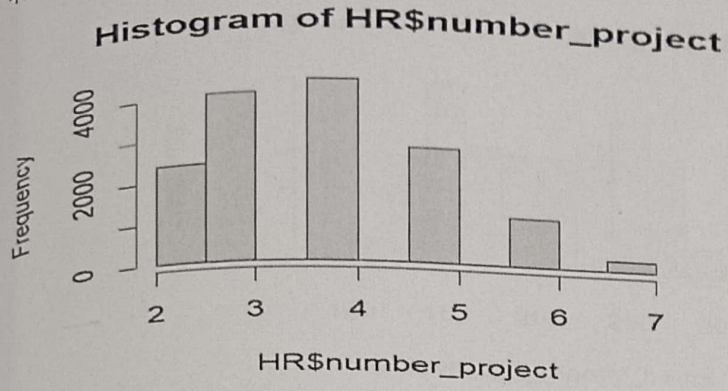


Figure 4.2.3

INTERPRETATION

The histogram shows the distribution of the variable HR$number_project. The x-axis represents the values of HR$number_project, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The most common number of projects assigned to employees is 3 and 4, with approximately 4000 employees in each category.
- The number of employees decreases as the number of projects increases. For example, there are fewer employees assigned to 2 or 7 projects compared to 3 or 4.
- The distribution is skewed to the right, with a longer tail extending to the right. This suggests that there are a few employees who are assigned to a large number of projects.

Overall, the histogram suggests that most employees are assigned to between 2 and 4 projects, with a smaller number of employees assigned to more or fewer projects. There are a few employees who are assigned to a large number of projects, which may be worth investigating further.

4.2.4 Average monthly hour distribution

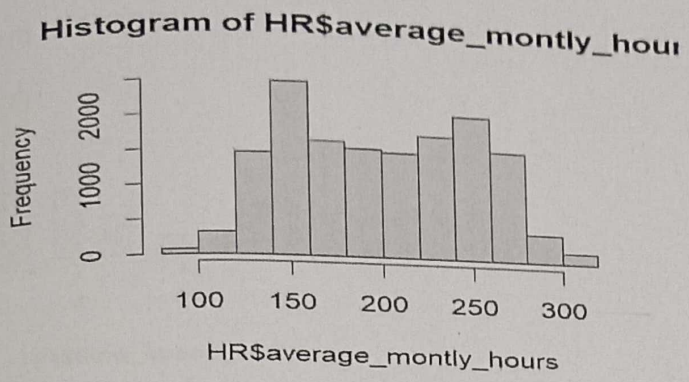


Figure 4.2.4

INTERPRETATION

The histogram shows the distribution of the variable HR$average_montly_hours. The x-axis represents the values of HR$average_montly_hours, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The distribution of HR$average_montly_hours is unimodal, meaning that there is a single peak in the distribution.
- The peak is around 150, suggesting that most employees work an average of 150 hours per month.
- The distribution is slightly skewed to the right, with a longer tail extending to the right. This suggests that there are a few employees who work significantly more than 150 hours per month.

Overall, the histogram suggests that most employees work an average of 150 hours per month, with a few employees working longer hours. The distribution is relatively symmetric, indicating that there is no strong bias in the average monthly hours worked.

4.2.5 Time Spend Company Distribution

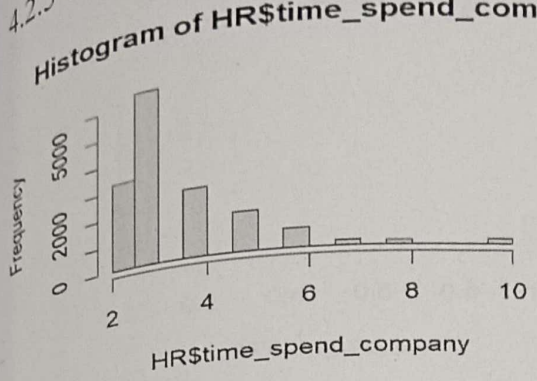


Figure 4.2.5

INTERPRETATION

The histogram shows the distribution of the variable HR$time_spend_company. The x-axis represents the values of HR$time_spend_company, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The most common number of years that employees spend with the company is 3, with approximately 6000 employees in this category.
- The number of employees decreases as the number of years spent with the company increases. For example, there are fewer employees who have spent 2 or 4 years with the company compared to 3 years.
- The distribution is skewed to the right, with a longer tail extending to the right. This suggests that there are a few employees who have spent a very long time with the company.

Overall, the histogram suggests that most employees stay with the company for a relatively short period of time, with a significant majority staying for 3 years or less. However, there are a few employees who have spent a much longer time with the company, which may be worth investigating further.

4.2.6 Work accident Distribution

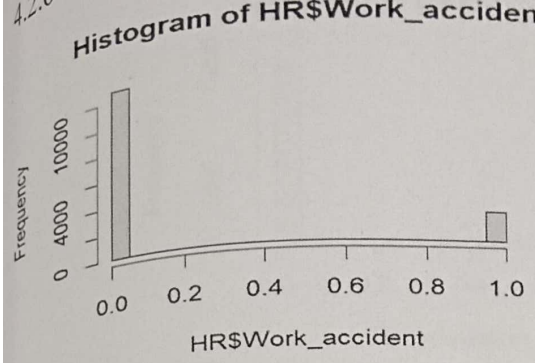


Figure 4.2.6

INTERPRETATION

The histogram shows the distribution of the variable HR$Work_accident. The x-axis represents the values of HR$Work_accident, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The majority of the observations have a value of 0, indicating that most employees did not experience a work accident.
- There is a smaller number of observations with a value of 1, indicating that a minority of employees did experience a work accident.
- The distribution is highly skewed to the right, with a long tail extending to the right. This suggests that there are a few employees who experienced multiple work accidents, which is unusual.

Overall, the histogram suggests that work accidents are relatively rare in the dataset, with most employees not experiencing any accidents. However, there are a few employees who have experienced multiple accidents, which may be worth investigating further.

4.2.7 promotion last year distribution

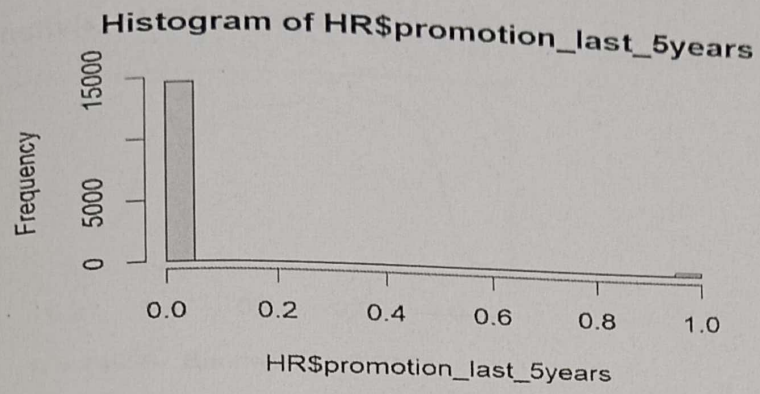


Figure 4.2.7

INTERPRETATION

histogram shows the distribution of the variable HR$promotion_last_5years. The x-axis represents the values of HR$promotion_last_5years, and the y-axis represents the frequency of each value.

Based on the histogram, we can see that:

- The majority of the observations have a value of 0, indicating that most employees were not promoted in the last 5 years.
- There is a smaller number of observations with a value of 1, indicating that a minority of employees were promoted in the last 5 years.
- The distribution is highly skewed to the right, with a long tail extending to the right. This suggests that there are a few employees who were promoted multiple times in the last 5 years, which is unusual.

Overall, the histogram suggests that promotions are relatively rare in the dataset, with most employees not being promoted in the last 5 years. However, there are a few employees who have been promoted multiple times, which may be worth investigating further.

4.2.8 Satisfaction level distribution

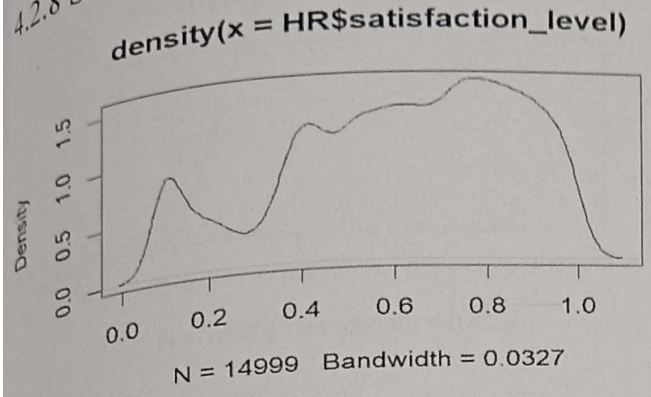


Figure 4.2.8

INTERPRETATION

The plot shows a density plot of the variable HR$satisfaction_level. A density plot is a smooth curve that shows the distribution of a continuous variable.

- The distribution of HR$satisfaction_level is bimodal, meaning that there are two peaks in the distribution.
- The first peak is around 0.2, suggesting that a group of employees have a low satisfaction level.
- The second peak is around 0.8, suggesting that another group of employees have a high satisfaction level.
- There is a valley between the two peaks, suggesting that there are few employees with a moderate satisfaction level.

Overall, the density plot suggests that employee satisfaction is either very low or very high, with few employees having a moderate satisfaction level. This may indicate that there are two distinct groups of employees in the dataset with different levels of satisfaction.

4.2.9 Last evaluation distribution

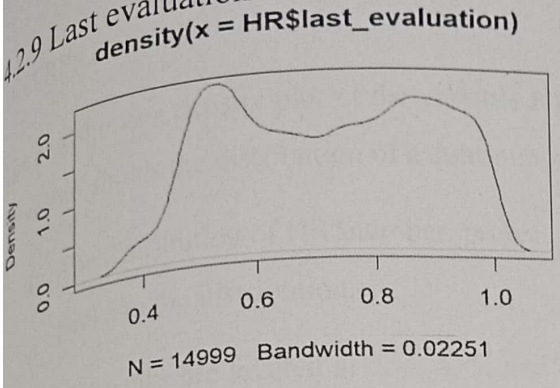


Figure 4.2.9

INTERPRETATION

The plot shows a density plot of the variable HR$last_evaluation. A density plot is a smooth curve that shows the distribution of a continuous variable.

- The distribution of HR$last_evaluation is unimodal, meaning that there is a single peak in the distribution.
- The peak is around 0.6, suggesting that most employees received a performance evaluation score of around 0.6.
- The distribution is relatively symmetric, with a slight skew to the right. This suggests that there are a few employees who received very high performance evaluation scores.

Overall, the density plot suggests that most employees received a performance evaluation score of around 0.6, with a few employees receiving higher scores. The distribution is relatively symmetric, indicating that there is no strong bias in the performance evaluations.

4.2.10 Number project distribution

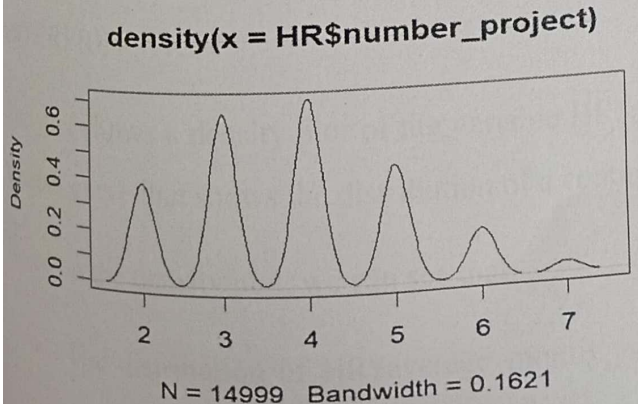


Figure 4.2.10

INTERPRETATION

The plot shows a density plot of the variable HR$number_project. A density plot is a smooth curve that shows the distribution of a continuous variable.

- The distribution of HR$number_project is multimodal, meaning that there are multiple peaks in the distribution.
- The peaks are located at 2, 3, 4, 5, and 6, suggesting that these are the most common numbers of projects that employees are assigned to.
- The distribution is relatively symmetric, with a slight skew to the right. This suggests that there are a few employees who are assigned to a large number of projects.

Overall, the density plot suggests that employees are assigned to a variety of projects, with most employees being assigned to between 2 and 6 projects. There are a few employees who are assigned to more than 6 projects, which may be worth investigating further.

4.2.11 Average montly hours

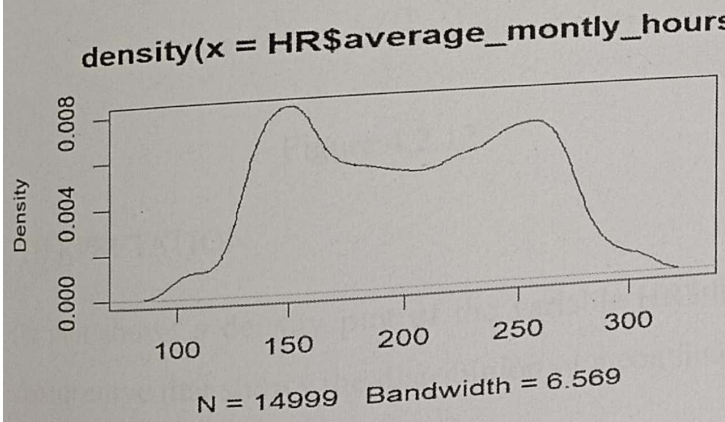


Figure 4.2.11

INTERPRETATION

The plot shows a density plot of the variable HR$average_montly_hours. A density plot is a smooth curve that shows the distribution of a continuous variable.

Based on the density plot, we can see that:

- The distribution of HR$average_montly_hours is unimodal, meaning that there is a single peak in the distribution.

- The peak is around 150, suggesting that most employees work an average of 150 hours per month.
- The distribution is slightly skewed to the right, with a longer tail extending to the right. This suggests that there are a few employees who work significantly more than 150 hours per month.

Overall, the density plot suggests that most employees work an average of 150 hours per month, with a few employees working longer hours. The distribution is relatively symmetric, indicating that there is no strong bias in the average monthly hours worked.

4.2.12 Time spend company

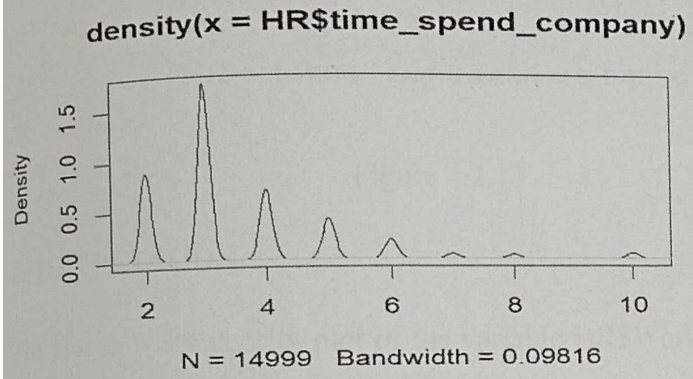


Figure 4.2.12

INTERPRETATION

The plot shows a density plot of the variable HR$time_spend_company. A density plot is a smooth curve that shows the distribution of a continuous variable.

Based on the density plot, we can see that:

- The distribution of HR$time_spend_company is multimodal, meaning that there are multiple peaks in the distribution.
- The peaks are located at 2, 3, 4, 5, 6, and 8, suggesting that these are the most common number of years that employees spend with the company.
- The distribution is skewed to the right, with a longer tail extending to the right. This suggests that there are a few employees who have spent a very long time with the company.

Overall, the density plot suggests that employees tend to stay with the company for a relatively short period of time, with most employees staying for between 2 and 6 years. However, there are a few employees who have spent a much longer time with the company, which may be worth investigating further

4.2.13 Work accident distribution

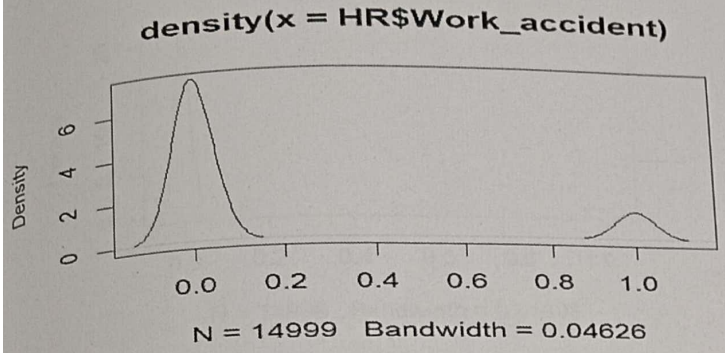


Figure 4.2.13

INTERPRETATION

The plot shows a density plot of the variable HR$Work_accident. A density plot is a smooth curve that shows the distribution of a continuous variable.

Based on the density plot, we can see that:

- The distribution of HR$Work_accident is bimodal, meaning that there are two peaks in the distribution.
- The first peak is around 0, suggesting that most employees did not experience a work accident.
- The second peak is around 1, suggesting that a smaller number of employees did experience a work accident.
- The distribution is highly skewed to the right, with a long tail extending to the right. This suggests that there are a few employees who experienced multiple work accidents, which is unusual.

Overall, the density plot suggests that work accidents are relatively rare in the dataset, with most employees not experiencing any accidents. However, there are a few employees who have experienced multiple accidents, which may be worth investigating further.

4.2.14 Promotion last 5 years

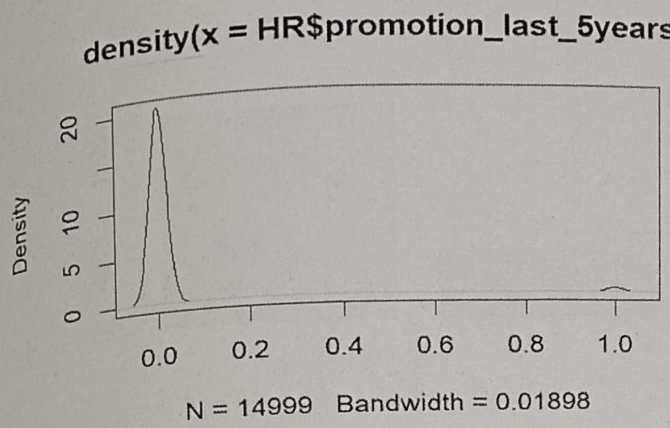


Figure 4.2.14

INTERPRETATION

The plot shows a density plot of the variable HR$promotion_last_5years. A density plot is a smooth curve that shows the distribution of a continuous variable.

Based on the density plot, we can see that:

- The distribution of HR$promotion_last_5years is bimodal, meaning that there are two peaks in the distribution.
- The first peak is around 0, suggesting that most employees were not promoted in the last 5 years.
- The second peak is around 1, suggesting that a smaller number of employees were promoted in the last 5 years.
- The distribution is highly skewed to the right, with a long tail extending to the right. This suggests that there are a few employees who were promoted multiple times in the last 5 years, which is unusual.

Overall, the density plot suggests that promotions are relatively rare in the dataset, with most employees not being promoted in the last 5 years. However, there are a few employees who have been promoted multiple times, which may be worth investigating further.

4.2.15 Barplot for left Distribution

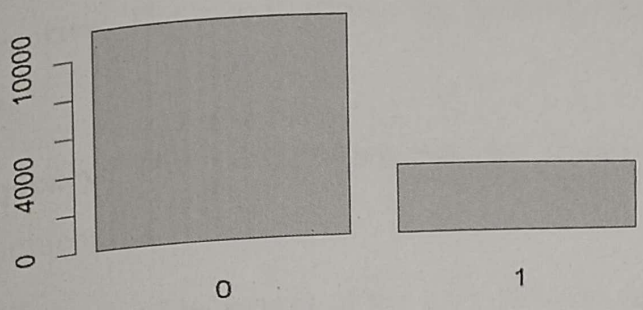


Figure 4.2.15

INTERPRETATION

- **categories:** The x-axis shows the two categories: 0 and 1. These likely represent "No Work Accident" and "Work Accident," respectively.
- **Frequency:** The y-axis represents the frequency or count of observations in each category.
- **Comparison:** The height of the bars indicates the relative frequency of each category. In this case, the bar corresponding to 0 (No Work Accident) is significantly taller than the bar corresponding to 1 (Work Accident). This suggests that a much larger portion of the dataset consists of employees who did not experience a work accident compared to those who did.

In summary: The bar chart indicates that work accidents are relatively rare in the dataset, with a significant majority of employees not experiencing any accidents.

4.2.16 Barplot for Department distribution

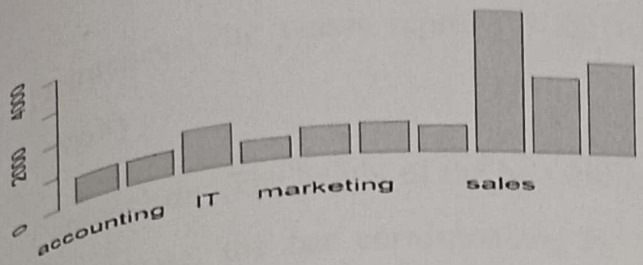


Figure 4.2.16

INTERPRETATION

- **Categories:** The x-axis shows the four departments: accounting, IT, marketing, and sales.
- **Frequency:** The y-axis represents the frequency or count of employees in each department.
- **Comparison:** The height of the bars indicates the relative size of each department. In this case, the bar corresponding to sales is significantly taller than the others, suggesting that it is the largest department.

In summary: The bar chart indicates that the sales department is the largest department in the dataset, followed by marketing, IT, and accounting.

4.2.17 Barplot for Salary distribution

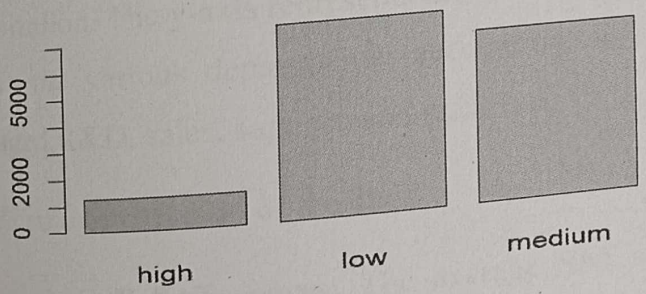


Figure 4.2.17

INTERPRETATION

- **Categories:** The x-axis shows the three categories: high, low, and medium.

- **Frequency:** The y-axis represents the frequency or count of observations in each category.
- **Comparison:** The height of the bars indicates the relative frequency of each category. In this case, the bar corresponding to low is significantly taller than the others, suggesting that a larger portion of the dataset falls into this category.

In summary: The bar chart indicates that the "low" category is the most frequent among the three categories, followed by "medium" and then "high.

4.2.18 BAR GRAPH FOR SAT

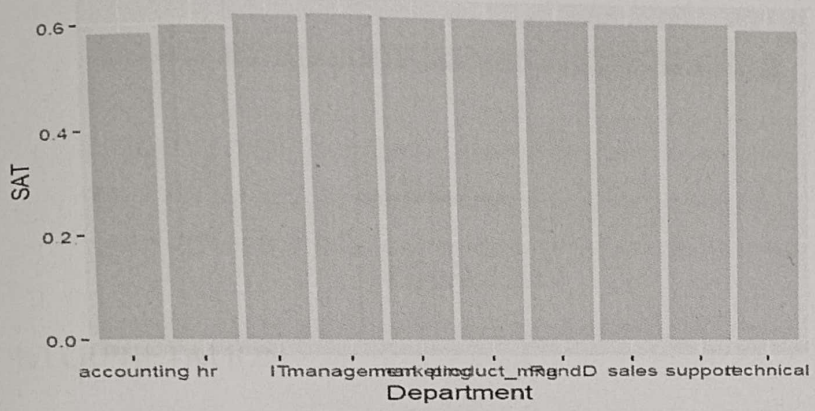


Figure 4.2.18

INTERPRETATION

The graph is a bar chart that displays the SAT scores for different departments within an organization. The y-axis represents the SAT scores, ranging from 0 to 0.6, while the x-axis shows the various departments: accounting, hr, IT, management, marketing, product managed,R&D, sales, support, and technical.

Here's an interpretation of the chart:

- **Overall SAT scores:** The average SAT scores across all departments seem to be relatively low, with most scores falling below 0.6.
- **Departmental variations:** There are some variations in SAT scores among departments. For example, the accounting department appears to have the lowest average SAT score, while the technical department seems to have the highest.

- **Comparison of departments:** It's possible to compare the SAT scores of specific departments to identify which departments have higher or lower levels of performance. For instance, the IT department's SAT score is higher than the hr department's.

4.2.19 RELATIONSHIP BETWEEN LEFT AND SATISFACTION LEVEL

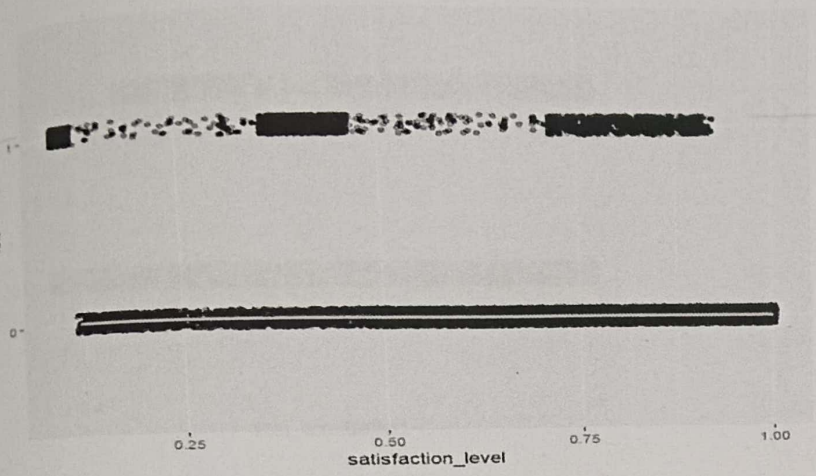


Figure 4.2.19

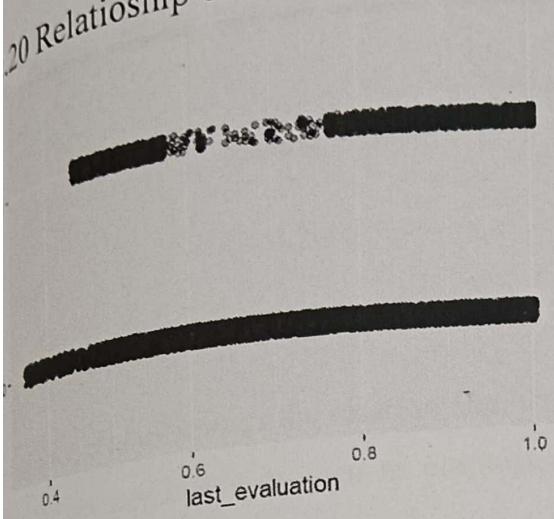
INTERPRETATION

The graph is a scatter plot that displays the relationship between satisfaction level and whether or not an employee left the company (left). The x-axis represents the satisfaction level, ranging from 0 to 1, while the y-axis indicates whether an employee left (1) or stayed (0).

- **Distribution of satisfaction levels:** The majority of employees have a satisfaction level between 0.75 and 1.00, suggesting that most employees are generally satisfied with their jobs.
- **Relationship with leaving:** The plot shows a clear separation between employees who stayed and those who left based on their satisfaction level. Employees who left tend to have lower satisfaction levels, while employees who stayed have higher satisfaction levels.
- **Threshold for leaving:** There seems to be a threshold around a satisfaction level of 0.50. Employees with satisfaction levels below this threshold are more likely to leave the company.

rall, the plot suggests that satisfaction level is a strong predictor of employee turnover. ployees with higher satisfaction levels are more likely to stay with the company, while those lower satisfaction levels are more likely to leave.

20 Relatioship between left and last evaluation



4.2.20

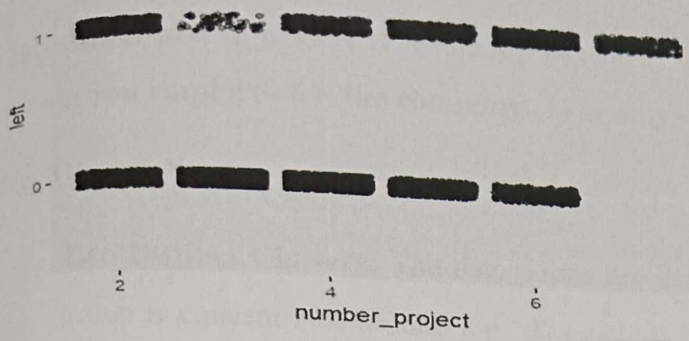
TERPRETATION

he plot shows the relationship between an employee's last evaluation score and whether they ft the company. The x-axis represents the last evaluation score, ranging from 0.4 to 1, with igher values indicating better performance. The y-axis shows whether an employee left (1) or ayed (0).

- **No clear trend:** There doesn't seem to be a strong relationship between the last evaluation score and leaving the company. Employees with both high and low evaluation scores left the company.
- **Possible clustering:** There might be a slight clustering of points around certain evaluation scores, but it's not very pronounced.

Overall, the plot suggests that the last evaluation score alone is not a strong predictor of employee turnover in this dataset. Other factors might be influencing employees' decisions to leave.

4.2.21 Relationship between left and number project



4.2.21

INTERPRETATION

This scatter plot examines the relationship between "number_project" and "left," likely indicating the number of projects an employee was involved in and whether they left (1) or stayed (0).

- **No clear trend:** There doesn't seem to be a strong relationship between the number of projects and leaving the company. Employees with both low and high project numbers left the company.
- **Possible clustering:** There might be a slight clustering of points around certain project numbers, but it's not very pronounced.
- Overall, the plot suggests that the number of projects alone is not a strong predictor of employee turnover in this dataset. Other factors might be influencing employees' decisions to leave

4.2.22 Relationship between left and average monthly hours

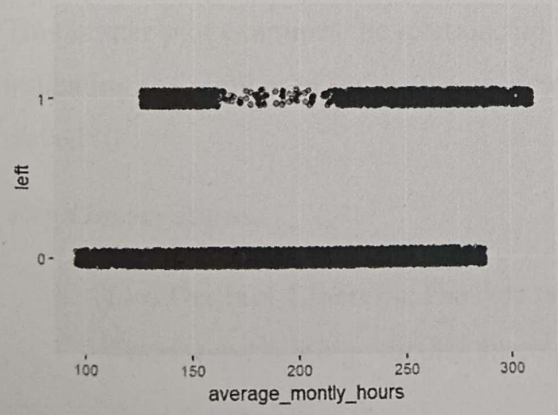


Fig 4.2.22

INTERPRETATION

This scatter plot shows the relationship between the average monthly hours worked and whether an employee left the company (1) or stayed (0).

Key Observations:

1. **Two Distinct Clusters:** The data points are clearly clustered into two groups. One group is concentrated around the "0" (stayed) level of the "left" variable. while the other group is around the "1" (left) level.
2. **Overlap in Average Monthly Hours:** There is a significant overlap in the average monthly hours worked between the two groups. This suggests that **average monthly hours alone may not be a strong predictor of whether an employee will leave.**

4.2.23 Relationship between left and time spend company

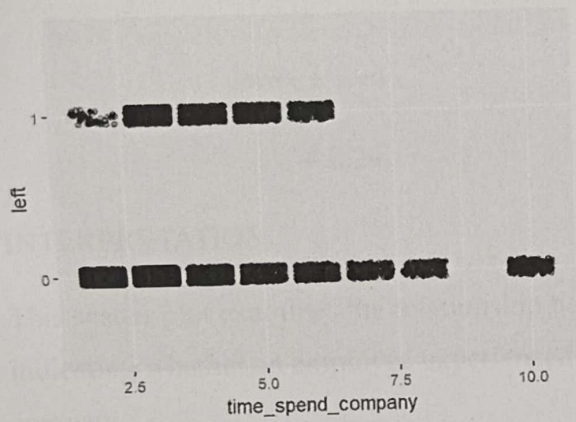


Fig 4.2.23

INTERPRETATION

This scatter plot examines the relationship between "time_spend_company" and "left," likely indicating the number of years an employee spent at the company and whether they left (1) or stayed (0).

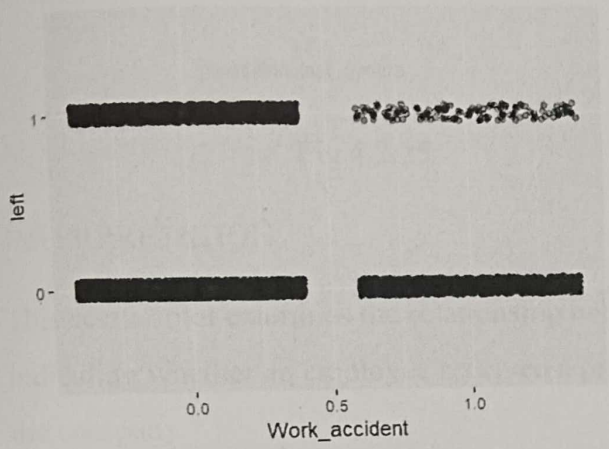
Key Observations:

1. **Two Distinct Clusters:** The data points are clearly clustered into two groups along the y-axis, representing whether an employee left (1) or stayed (0).

2. **Time Spent Distribution:** The distribution of "time_spend_company" shows multiple clusters. suggesting that employees tend to leave after specific durations with the company.

3. **Separation and Overlap:** There is some separation between the clusters. but there is also significant overlap. especially in the lower time spans.

4.2.24 Relationship between left and work accident



4.2.24

INTERPRETATION

This scatter plot examines the relationship between "Work_accident" and "left," likely indicating whether an employee experienced a work accident and subsequently left the company.

1. **Two Distinct Clusters:** The data points are clearly clustered into two groups along the y-axis, representing whether an employee left (1) or stayed (0).

2. **Work Accident Distribution:** The distribution of "Work_accident" seems to be somewhat similar across both clusters, with a concentration around 0 (no work accident) and some spread towards 1 (work accident).

3. **Overlap and Separation:** There is a significant overlap in the "Work_accident" values between the two clusters. However, a small portion of data points from the "left" cluster seem to have higher "Work_accident" values compared to the "stayed" cluster.

4.2.25 Relationship between left and promotion last 5 years

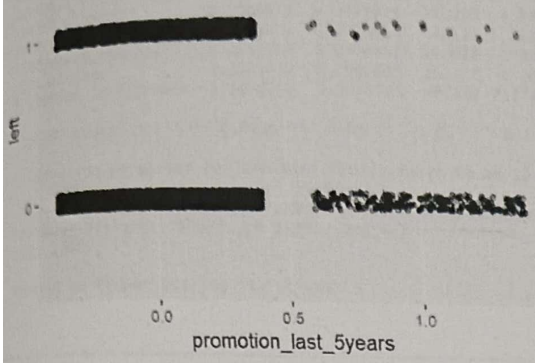


Fig 4.2.25

INTERPRETATION

This scatter plot examines the relationship between "promotion_last_5years" and "left," likely indicating whether an employee received a promotion in the last 5 years and subsequently left the company.

Key Observations:

1. **Two Distinct Clusters:** The data points are clearly clustered into two groups along the y-axis, representing whether an employee left (1) or stayed (0).
2. **Promotion Distribution:** The distribution of "promotion_last_5years" shows two distinct patterns:
    - A large cluster around 0, indicating a majority of employees did not receive a promotion in the last 5 years.
    - A smaller cluster around 1, indicating a smaller group of employees who did receive a promotion.
3. **Separation:** The two clusters are relatively well-separated along the x-axis

```
average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years, family = "binomial", data = train_HR)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.1585696  0.1310104   1.210    0.226
satisfaction_level   -4.1320825  0.1080741 -38.234   < 2e-16 ***
last_evaluation       0.7284889  0.1628630   4.473 7.71e-06 ***
number_project       -0.3044955  0.0233694 -13.030   < 2e-16 ***
average_montly_hours  0.0045909  0.0005657   8.116 4.82e-16 ***
time_spend_company    0.2342700  0.0166497  14.070   < 2e-16 ***
Work_accident        -1.5131837  0.0997615 -15.168   < 2e-16 ***
promotion_last_5years -1.6458291  0.2670765  -6.162 7.17e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13172  on 11998  degrees of freedom
Residual deviance: 10643  on 11991  degrees of freedom
AIC: 10659

Number of Fisher Scoring iterations: 5

>
```

INTERPRETATION

The model is a logistic regression with the dependent variable being "left" (presumably whether an employee left the company or not). The independent variables are:

- satisfaction_level
- last_evaluation
- number_project
- average_montly_hours
- time_spend_company
- work_accident
- promotion_last_5years [1]

Coefficients:

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	0.158570	0.131010	1.210	0.226	
satisfaction_level	-4.132082	0.108074	-38.234	< 2e-16	***
last_evaluation	0.728489	0.162863	4.473	7.71e-06	***
number_project	-0.304496	0.023369	-13.030	< 2e-16	***
average_montly_hours	0.004591	0.000566	8.116	4.82e-16	***
time_spend_company	0.234270	0.016650	14.070	< 2e-16	***
work_accident	-1.513184	0.099762	-15.168	< 2e-16	***
promotion_last_5years	-1.645829	0.267076	-6.162	7.17e-10	***

Interpretation of Coefficients:

- **Intercept:** The intercept represents the log-odds of an employee leaving when all independent variables are zero. In this case, it's not very interpretable since most variables can't be zero in reality.
- **satisfaction_level:** A one-unit increase in satisfaction level is associated with a decrease in the log-odds of leaving by 4.132. In other words, higher satisfaction decreases the likelihood of leaving.
- **last_evaluation:** A one-unit increase in the last evaluation score is associated with an increase in the log-odds of leaving by 0.728. This suggests that a higher evaluation score might increase the likelihood of leaving, possibly due to dissatisfaction with current compensation or lack of promotion opportunities.
- **number_project:** A one-unit increase in the number of projects is associated with a decrease in the log-odds of leaving by 0.304. This indicates that more projects might lead to a lower likelihood of leaving, potentially due to increased engagement or job satisfaction.
- **average_montly_hours:** A one-unit increase in average monthly hours is associated with an increase in the log-odds of leaving by 0.0046. While the effect is small, it suggests that longer working hours might slightly increase the likelihood of leaving.
- **time_spend_company:** A one-unit increase in time spent at the company is associated with an increase in the log-odds of leaving by 0.234. This could indicate that employees might be more likely to leave after a certain tenure, possibly due to reaching a plateau or seeking new challenges.
- **work_accident:** Having a work accident is associated with a decrease in the log-odds of leaving by 1.513. This suggests that employees who have had work accidents might be less likely to leave, potentially due to increased loyalty or fear of job loss.
- **promotion_last_5years:** Not being promoted in the last 5 years is associated with an increase in the log-odds of leaving by 1.646. This indicates that lack of promotion is a strong predictor of leaving.

Model Fit:

intercept. ...ance of a model with only the

- **Residual deviance:** The residual deviance represents the deviance of the fitted model.
- **AIC:** The Akaike Information Criterion (AIC) is a measure of model fit, with lower values indicating better fit.

Interpretation of Significance:

The significance levels (Pr(>|z|)) indicate the probability of observing the estimated coefficient by chance if the true coefficient were zero. The asterisks (***) indicate that the coefficient is statistically significant at the 0.001 level, meaning that the relationship between the variable and the outcome is very likely not due to chance.

Overall Interpretation:

The model suggests that employee satisfaction, last evaluation, number of projects, time spent at the company, work accidents, and promotions are all important factors influencing whether an employee leaves. The model can be used to identify employees at risk of leaving and to implement strategies to improve retention.

Caveats:

- The model assumes a linear relationship between the log-odds of leaving and the independent variables. This assumption should be checked using diagnostic plots.
- The model assumes that the independent variables are not perfectly correlated with each other. Multicollinearity can affect the estimation of coefficients.
- The model is based on the data provided and may not generalize to other populations or contexts.

CHAPTER 5 INFERENCE

5.1 SUMMARY OF FINDINGS OF THE STUDY

1. **Influential Factors on Employee Turnover**:
    - **Satisfactory Level**: Higher self-reported satisfaction levels are strongly associated with lower turnover rates. Employees who feel satisfied are less likely to leave the company.
    - **Last Evaluation Score**: Employees who receive higher performance evaluation scores tend to stay longer. This suggests that recognition and feedback play a crucial role in retention.
    - **Number of Projects**: Employees involved in more projects may feel more engaged and valued, which positively correlates with retention.
    - **Average Monthly Hours**: Excessive working hours can lead to burnout, increasing the likelihood of turnover. Finding a balance is essential for employee well-being.
    - **Time Spent at Company**: Employees with longer tenure are generally more loyal, indicating that investment in employee development can lead to higher retention.
    - **Work Accidents**: Employees who have experienced accidents at work are more likely to leave, highlighting the importance of workplace safety and employee health.

- **Promotions**: Employees who have received promotions in the past five years are less likely to leave, emphasizing the significance of career advancement opportunities.
- **Department and Salary**: Variability in turnover rates across departments and salary levels suggests that some departments may need more focused retention strategies, particularly for lower-salaried roles.

2. **Recommendations for HR**:
   - **Enhance Job Satisfaction**: Implement initiatives to increase employee satisfaction through regular feedback, recognition programs, and work-life balance policies.
   - **Career Development**: Offer clear pathways for promotion and professional growth to retain top talent.
   - **Safety and Well-being Programs**: Focus on improving workplace safety to reduce work-related accidents and associated turnover.
   - **Tailored Strategies**: Analyze department-specific turnover rates and develop targeted interventions to address unique challenges within those teams.

The logistic regression model provided valuable insights into the predictors of employee turnover. By understanding these factors, the company can take proactive measures to improve employee retention, ultimately leading to a more stable workforce and a positive organizational culture. Implementing the recommendations based on these findings will enhance employee satisfaction and reduce turnover rates, fostering long-term success for the company.

5.2 SUGGESTIONS OF THE STUDY

The analysis of employee turnover using logistic regression has provided valuable insights into the factors influencing employee retention. Based on these findings, the following suggestions are proposed for HR to improve retention rates and enhance job satisfaction:

1. **Enhance Employee Satisfaction**:

- **Regular Feedback and Recognition**: Implement structured performance reviews that provide constructive feedback. Recognize employees' contributions to increase morale and satisfaction.
- **Employee Engagement Surveys**: Conduct regular surveys to gauge employee satisfaction and address concerns proactively.

2. **Career Development Opportunities**:
   - **Training Programs**: Offer professional development opportunities and skills training to help employees advance in their careers.
   - **Clear Promotion Pathways**: Establish transparent criteria for promotions and career progression, ensuring employees understand the steps they need to take.
3. **Work-Life Balance Initiatives**:
   - **Flexible Work Arrangements**: Consider implementing flexible work hours or remote work options to help employees balance personal and professional responsibilities.
   - **Manage Workloads**: Regularly assess project assignments and workloads to prevent employee burnout from excessive hours.
4. **Safety and Well-being Focus**:
   - **Workplace Safety Programs**: Enhance safety training and resources to minimize work-related accidents. Create a culture that prioritizes employee well-being.
   - **Employee Support Services**: Provide access to mental health resources and employee assistance programs to support overall well-being.
5. **Department-Specific Strategies**:
   - **Analyze Turnover Trends**: Review turnover rates by department to identify specific issues that may need targeted interventions.
   - **Tailored Retention Programs**: Develop department-specific initiatives based on unique challenges or characteristics that may affect employee satisfaction and retention.

6. **Competitive Compensation and Benefits**:
    - **Salary Reviews**: Regularly review and adjust salaries to ensure competitiveness within the industry. Consider offering benefits that enhance overall compensation.
    - **Incentive Programs**: Implement bonus or incentive programs tied to performance to motivate and retain employees.
7. **Promote a Positive Company Culture**:
    - **Inclusive Environment**: Foster an inclusive workplace culture that values diversity and encourages collaboration.
    - **Team-Building Activities**: Organize team-building events to strengthen relationships among employees and promote a sense of belonging.

By implementing these suggestions, HR can create a supportive and engaging workplace that not only addresses the factors contributing to employee turnover but also enhances overall job satisfaction. These proactive measures will ultimately lead to improved employee retention and a more positive organizational culture.

5.3 CONCLUSION

In conclusion, the analysis of employee turnover within the company has highlighted critical factors influencing whether employees decide to leave their positions. By leveraging logistic regression on the dataset, we have identified that variables such as satisfactory levels, performance evaluation scores, number of projects, average monthly hours worked, tenure at the company, and experiences with workplace accidents significantly contribute to turnover rates. Higher employee satisfaction and favorable performance reviews are strongly correlated with retention, while excessive working hours and lack of promotion opportunities tend to increase the likelihood of departure. Furthermore, the data indicates that departments and salary levels also play a vital role in shaping employee decisions to remain with the company. Armed with these insights, HR can formulate targeted strategies aimed at enhancing employee engagement and satisfaction, such as implementing more robust feedback mechanisms, providing clearer pathways for career advancement, and promoting a healthier work-life balance. By addressing these areas, the company can work towards reducing

turnover rates, fostering a more committed workforce, and ultimately cultivating a positive organizational culture that supports both current and future employees.

BIBLIOGRAPHY

Alfred J. Walker: Human Resources Information Systems Development, 1982

Ashok K. Gupta: Developing Human Resource Information System, 2008

Dr. P.K. Gupta and Susheel Chhabra: Human Resource Information System, 2015

Jayant Mukherjee: Designing Human Resource Management Systems - A Leader's Guide, 2012

Kelvin Molly: Human Resource Information System, 2014

Michael J. Kavanagh, Mohan Thite, Richard D. Johnson: Human Resource Information Systems - Basics, Applications, and Future Directions, Third Edition, 2015

Satish M Badgi: Practical Guide to Human Resource Information Systems, 2012

Teresa Torres-Coronas and Mario Arias-Oliva: Encyclopedia of Human Resources Information Systems - Challenges in e-HRM, 2008

Yorrick Bakker: Back to the Future of Human Resource Information Systems, 2012