**George Herbert Walker School of Business & Technology**

Department of Computer and Information Sciences

Master of Science in Data Analytics

CSDA 6010 Data Analytics Practicum



DEPARTMENT CHAIR: DR. JAMES CURTIS

DATA ANALYTICS PRACTICUM INSTRUCTORS:

1. DR. JIANGPING WANG

2. DR. ALI OVLIA

PRAVEEN KUMAR RATIKINDI

WEBSTER UNIVERSITY

CSDA 6010 DATA ANALYTICS PRACTICUM

FALL 2024

Dr. JIANGPING WANG

Table of Contents

[EXECUTIVE SUMMARY: 6](#_Toc177569351)

[INTRODUCTION 6](#_Toc177569352)

[Business Considerations: 7](#_Toc177569353)

[Outcome Variables: 7](#_Toc177569354)

[BUSINESS PROBLEMS: 7](#_Toc177569355)

[BUSINESS GOALS: 7](#_Toc177569356)

[ANALYTICS GOALS: 8](#_Toc177569357)

[ANALYTICAL APPROACH: 8](#_Toc177569358)

[DATA PREPROCESSING: 9](#_Toc177569359)

[Missing Values: 10](#_Toc177569360)

[Null values: 10](#_Toc177569361)

[Zeros in dataset: 10](#_Toc177569362)

[DATA EXPLORATION: 10](#_Toc177569363)

[FEATURE EXPLORATION: 13](#_Toc177569364)

[PREDICTOR RELEVANCY: 23](#_Toc177569365)

[Satisfaction Level: 23](#_Toc177569366)

[Last Evaluation: 24](#_Toc177569367)

[Number of Projects: 25](#_Toc177569368)

[Average Monthly Hours: 26](#_Toc177569369)

[Time Spent in Company: 27](#_Toc177569370)

[Work Accident: 28](#_Toc177569371)

[Promotion in Last 5 Years: 29](#_Toc177569372)

[Department (Sales): 30](#_Toc177569373)

[Salary: 31](#_Toc177569374)

[DIMENSION REDUCTION: 32](#_Toc177569375)

[DATA TRANSFORMATION: 32](#_Toc177569376)

[DATA PARTITION: 32](#_Toc177569377)

[TRANSLATING GOALS: 33](#_Toc177569378)

[MODEL SELECTION: 34](#_Toc177569379)

[MODEL FITTING: 34](#_Toc177569380)

[MODEL EVALUATION: 40](#_Toc177569381)

[HYPOTHESIS ANALYSIS: 43](#_Toc177569382)

[CONCLUSION AND FUTURE WORK: 45](#_Toc177569383)

[BUSINESS RECOMMENDATIONS: 46](#_Toc177569384)

Table of Figures

[Figure 1Structure of the Employee data 11](#_Toc177569385)

[Figure 2 Head of the Employee data 11](#_Toc177569386)

[Figure 3 Summary of the Employee data 11](#_Toc177569387)

[Figure 4 Satisfaction level of employees 13](#_Toc177569388)

[Figure 5 Last evaluation of employees 14](#_Toc177569389)

[Figure 6 Number of projects an employee involved in 15](#_Toc177569390)

[Figure 7 Average monthly hours of employees 16](#_Toc177569391)

[Figure 8 Time spent at the company by the employees in years 17](#_Toc177569392)

[Figure 9 Work accidents 18](#_Toc177569393)

[Figure 10 Did the employee leave the company? 19](#_Toc177569394)

[Figure 11 Promotion in the last 5 years 20](#_Toc177569395)

[Figure 12 Number of employees working the respective departments 21](#_Toc177569396)

[Figure 13 Salary level of employees 22](#_Toc177569397)

[Figure 14 Satisfaction level of employees 23](#_Toc177569398)

[Figure 15 Last Evaluation 24](#_Toc177569399)

[Figure 16 Number of Projects 25](#_Toc177569400)

[Figure 17 Average Monthly Hours 26](#_Toc177569401)

[Figure 18 Time Spent in Company 27](#_Toc177569402)

[Figure 19 Work Accident 28](#_Toc177569403)

[Figure 20 Promotion in Last 5 Years 29](#_Toc177569404)

[Figure 21 Department (Sales) 30](#_Toc177569405)

[Figure 22 Salary 31](#_Toc177569406)

[Figure 23 Dimension Reduction 32](#_Toc177569407)

[Figure 24 Random Forest for determining key factors for turnover 35](#_Toc177569408)

[Figure 25 Decision Tree for determining key factors for turnover 36](#_Toc177569409)

[Figure 26 Logistic Regression for Predicting Turnover 37](#_Toc177569410)

[Figure 27 Logistic Reg Confusion matrix 38](#_Toc177569411)

[Figure 28 KNN Confusion matrix 39](#_Toc177569412)

[Figure 29 Naive bayes confusion matrix 40](#_Toc177569413)

[Figure 30 KNN confusion matrix on holdout set 41](#_Toc177569414)

# EXECUTIVE SUMMARY:

This project focuses on addressing two primary business goals: predicting employee turnover and identifying key factors driving turnover. Using a dataset of employee records, various machine learning techniques were employed to achieve these objectives. The Random Forest and Decision Tree models were utilized to uncover the most influential factors behind employee turnover. Both models consistently identified satisfaction level, number of projects, and last evaluation scores as the top predictors, with lower satisfaction being strongly linked to higher turnover rates.

On the predictive side, logistic regression, KNN, and Naive Bayes models were implemented to forecast employee turnover. These models, after appropriate tuning, delivered reasonable accuracy. For instance, the KNN model achieved a validation accuracy of over 94%, while Logistic Regression and Naive Bayes models also performed well, with KNN slightly outperforming others in accuracy. However, Naive Bayes showed slightly weaker performance in precision and predictive power, especially when dealing with employees expected to stay versus those likely to leave.

The analysis revealed a critical insight: employee satisfaction is the most significant factor influencing turnover, which aligns with common retention challenges. Other factors, such as time spent at the company and workload (measured by average monthly hours), also play essential roles in employee decisions to leave.

From a business perspective, focusing on improving employee satisfaction, reducing excessive workloads, and regularly evaluating employees’ contributions may help in reducing turnover. Implementing these changes, along with predictive models to proactively identify employees at risk of leaving, can help the company retain valuable talent and reduce turnover costs. Overall, the combination of predictive analytics and business strategy offers a comprehensive solution to tackling employee turnover effectively.

# INTRODUCTION

In today’s competitive business environment, retaining top talent is crucial for the growth and success of any organization. For Future Tech Inc., a mid-sized software-as-a-service (SaaS) provider specializing in financial technology solutions, employee retention is more than just a goal — it's a necessity. With a significant portion of its workforce in sales, support, and technical roles, turnover can disrupt operations, inflate costs, and impact overall company performance.

As Zig Ziglar wisely noted, “You don't build a business. You build people, and people build the business.” This philosophy underscores the importance of understanding and addressing the factors driving employee turnover. The aim of this project is twofold: firstly, to develop a predictive model to identify employees likely to leave the company; and secondly, to uncover the key factors responsible for turnover.

By leveraging data-driven insights, Future Tech Inc. seeks to enhance its human capital strategy, minimize turnover-related disruptions, and cultivate a more engaged and motivated workforce, ultimately driving sustained growth and competitive advantage.

## Business Considerations:

1. Retention Strategies: Use insights to develop targeted strategies addressing key turnover factors like job satisfaction and compensation.

2. Cost Reduction: Lower costs associated with hiring and training by reducing turnover through focused interventions.

3. Workplace Improvements: Enhance work-life balance and career opportunities to boost employee satisfaction and reduce turnover.

4. Predictive Planning: Integrate predictive models into workforce planning to manage staffing proactively.

5. Data-Driven Decisions: Utilize analytics to refine HR policies and foster a data-driven talent management approach.

# Outcome Variables:

The primary outcome variable in this project is employee turnover, represented by whether an employee leaves the company ("left") or stays. This binary variable serves as the key target for predictive modeling, allowing the identification of patterns and factors influencing turnover. By analyzing this outcome, the project aims to determine the secondary outcome in which variables, such as job satisfaction, salary, number of projects, or time spent at the company, significantly correlate with the likelihood of an employee leaving. Understanding the influence of these predictors on the outcome variable will help in developing effective retention strategies and optimizing human resource management practices.

# BUSINESS PROBLEMS:

**Primary Problem:** Future Tech Inc. is experiencing a high employee turnover rate, leading to increased costs associated with recruiting, onboarding, and training new employees. This turnover disrupts the company's operations, affects project timelines, and reduces overall productivity.

**Secondary Problem**: The company lacks a clear understanding of the main factors contributing to employee turnover, resulting in inefficient HR policies and employee engagement strategies. Without knowing the key drivers of turnover, the company struggles to create an effective retention plan.

# BUSINESS GOALS:

**Primary Goal:** Develop a predictive model to accurately identify employees who are at high risk of leaving the company. By predicting potential turnover, Future Tech Inc. can proactively engage with at-risk employees, provide support, and implement retention strategies to reduce the turnover rate and associated costs.

**Secondary Goal**: Identify and analyze the main factors influencing employee turnover at Future Tech Inc. Using data-driven insights, propose targeted HR interventions and policy changes that address the underlying reasons for turnover, improve employee satisfaction, and foster a more stable and motivated workforce.

# ANALYTICS GOALS:

**Primary Goal:** Create a machine learning model to predict the likelihood of employees leaving Future Tech Inc. This model will use various employee attributes to accurately predict turnover, enabling the company to take proactive measures.

**Secondary Goal**: Perform exploratory data analysis (EDA) and feature importance analysis to uncover the primary factors influencing employee turnover. This analysis will include statistical techniques and data visualization methods to identify patterns and trends that contribute to employee attrition.

# ANALYTICAL APPROACH:

1. DATA PREPROCESSING:

This step involves defining and understanding each attribute in the dataset, such as satisfaction levels, evaluation scores, and salary. It also includes data exploration, checking for missing values, zeros, and ensuring overall data quality before proceeding to further analysis.

2. PREDICTOR ANALYSIS AND RELEVANCY:

Predictor’s analysis is performed and analyzed the relevance of each predictor variable, such as satisfaction level, number of projects, and average monthly hours, in relation to employee turnover. This step identifies which factors are most strongly associated with the likelihood of an employee leaving the company.

3. DIMENSION REDUCTION:

All the predictors are important, so no predictor is removed during analysis.

4. DATA TRANSFORMATION:

Data transformation includes changing the salary

5. DATA PARTITIONING METHODS:

The dataset is split into training and testing sets to validate the models. This partitioning allows for evaluating how well the model generalizes to unseen data and prevents overfitting.

6. MODEL SELECTION:

Choose appropriate machine learning models, such as Logistic Regression for predicting employee turnover and Random Forest for variable importance.

7. MODEL FITTING, VALIDATION ACCURACY, AND TEST ACCURACY:

Fit the selected models to the training data and evaluate their performance using validation techniques. Assess the models' accuracy on both the training and test sets to ensure they perform well on new, unseen data.

8. REPORT MODELS PERFORMANCE:

Generate detailed reports on the performance of each model, including metrics like accuracy, precision, recall, and F1-score. These reports help compare models and select the best one for predicting employee turnover.

9. MODEL EVALUATION (OF THE SELECTED MODELS):

Conduct a thorough evaluation of the final selected model to ensure it meets the business goals. This includes interpreting the model's predictions and understanding the impact of key variables on employee turnover.

10. OBSERVATION AND CONCLUSION:

Summarize the findings from the data analysis and model evaluations. Provide clear, actionable recommendations for HR strategies based on the insights gained, focusing on reducing turnover and improving employee retention.

# DATA PREPROCESSING:

**Attributes Definition:**

|  |  |
| --- | --- |
| **Variables** | **Descriptions** |
| satisfaction\_level | Represents the employee's satisfaction level at work on a scale from 0 to 1, where higher values indicate higher satisfaction. |
| last\_evaluation | Represents the score of the last evaluation given to the employee, also on a scale from 0 to 1, with higher values indicating better performance. |
| number\_project | Indicates the number of projects an employee has been involved in. |
| average\_montly\_hours | Represents the average number of hours an employee works per month. |
| time\_spend\_company | Indicates the total number of years the employee has spent in the company. |
| Work\_accident | A binary variable indicating whether the employee has had a work-related accident. 1 represents "Yes," and 0 represents "No." |
| left | A binary variable indicating whether the employee has left the company. 1 represents "Yes," and 0 represents "No." |
| promotion\_last\_5years | A binary variable indicating whether the employee has been promoted in the last 5 years. 1 represents "Yes," and 0 represents "No." |
| sales | Represents the department in which the employee works, such as sales, technical, HR. |
| salary | Represents the salary level of the employee. The possible values are low, medium, and high, with low being the lowest salary and high being the highest. |

## Missing Values:

There are no missing values in the dataset



## Null values:

There are no Null values in the dataset



## Zeros in dataset:

No zeros found in numerical variable’s data, zeros are observed in few columns which are binary.

A computer code with numbers and letters

Description automatically generated

# DATA EXPLORATION:

The dataset (Employee.csv) contains 14,999 observations and 10 variables. Each row corresponds to an individual employee, and each column represents different attributes related to the employee's work environment and performance. All the variables are numeric datatype except sales and salary which are characters.

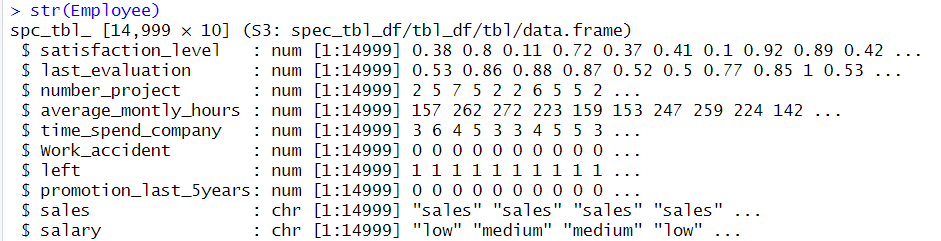


Figure 1Structure of the Employee data

The first few observations and variables of the Employee dataset.

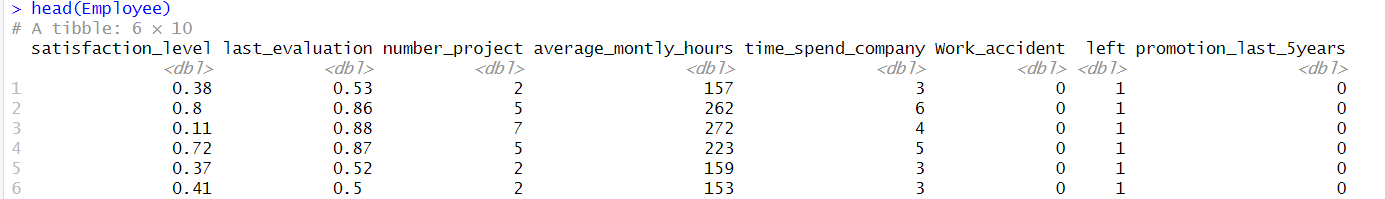


Figure 2 Head of the Employee data

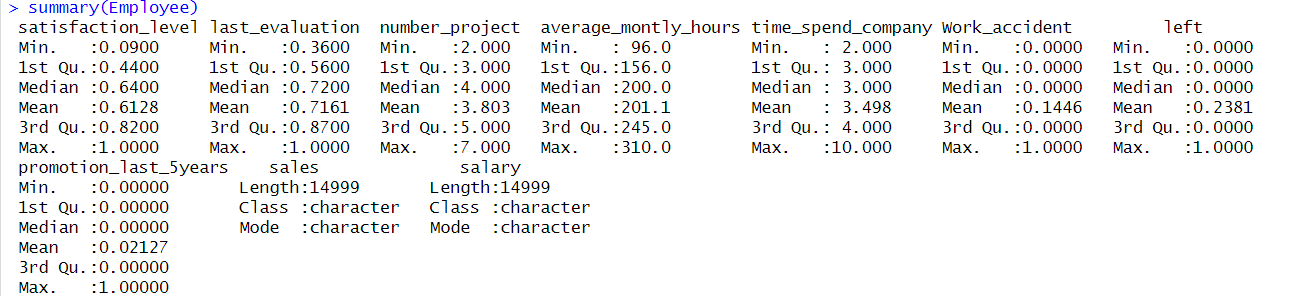


Figure 3 Summary of the Employee data

1. SATISFACTION LEVEL:

The mean satisfaction level among employees is 0.6128, which is slightly lower than the median of 0.64. This suggests that the satisfaction levels are somewhat skewed to the left, that means a small number of employees might have significantly lower satisfaction levels, that pulled the mean down.

2. LAST EVALUATION:

The mean and median values for the last evaluation are quite close (0.7161 and 0.72), indicating that the distribution of evaluation scores is likely symmetric. Most employees seem to receive relatively high evaluations indicating that they are performing better.

3. NUMBER OF PROJECTS:

The median number of projects handled by employees is 4, with a mean slightly lower at 3.803. This suggests that while most employees work on 4 projects, some work on fewer, possibly indicating a small group of employees with a lighter workload.

4. AVERAGE MONTHLY HOURS:

The mean (201.1 hours) and median (200 hours) for average monthly hours worked are very close, indicating that employees typically work around 200 hours per month. The range of hours (from 96 to 310) shows significant variability, suggesting that while some employees work fewer hours, others may be working a lot more, possibly indicating overtime which might cause burnout.

5. TIME SPENT AT COMPANY:

Employees have a median tenure of 3 years, with a mean of 3.498 years. The maximum time spent at the company is 10 years, which might indicate a group of long-term employees. The slight difference between the mean and median could imply a few employees with longer tenures than most.

6. WORK ACCIDENTS:

The mean value for work accidents is 0.1446, which implies that approximately 14.46% of employees have experienced a work accident. Since the median is 0, it shows that most employees have not had any accidents.

7. EMPLOYEE TURNOVER (LEFT):

The mean value for the `left` variable is 0.2381, meaning that around 23.81% of employees have left the company. This could be an area of concern if the company aims to reduce turnover.

8. PROMOTION IN THE LAST 5 YEARS:

The mean value for promotions in the last 5 years is very low (0.02127), indicating that only about 2.13% of employees received a promotion in this period. This could suggest limited opportunities for advancement, which might be linked to other factors like satisfaction levels and employee turnover.

# FEATURE EXPLORATION:

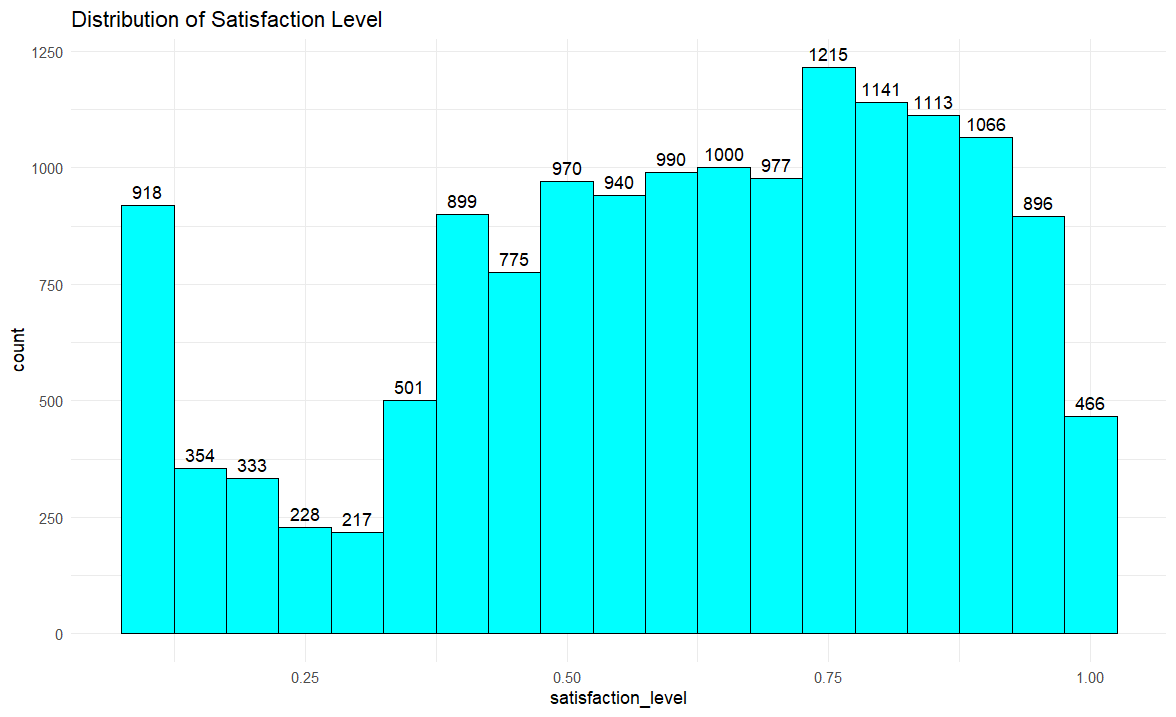
1.satisfaction\_level:

Figure 4 Satisfaction level of employees

Employee satisfaction levels show a generally positive sentiment, with a significant number of employees reporting high satisfaction, particularly between 0.7 and 0.8. Lower satisfaction levels around 0.1 are still notable, but fewer employees fall within the mid-range satisfaction levels (0.25 to 0.5). Most employees are satisfied as indicated by a sharp increase in counts above 0.5, peaking between 0.7 and 0.8. There is a gradual decline in the number of employees as satisfaction levels approach 1.0, with fewer reporting the highest levels of satisfaction. This pattern suggests overall satisfaction is high, but there's potential to further increase the number of highly satisfied employees.

**2.** last\_evaluation:

A graph of a number of blue bars

Description automatically generated with medium confidence

Figure 5 Last evaluation of employees

Employee evaluation scores are most concentrated around 0.5 and 0.6, with 1,631 and 1,687 employees respectively, suggesting average performance is common. After this peak, there is a gradual decline in the number of employees with higher scores, with only 804 employees having near-perfect evaluations. A secondary peak occurs around 0.8, where 1,478 employees are rated, indicating a group that performs above average. The lower end of the scale, around 0.3, has very few employees (77), indicating that poor performance is rare. Overall, most employees fall within mid to high evaluation scores, reflecting a generally average to above-average performance level across the workforce, with fewer at the extreme ends.

3. number\_project:

A graph with numbers and a number of projects

Description automatically generated

Figure 6 Number of projects an employee involved in

The most frequent number of projects handled by employees is 4, with 4,365 employees working on exactly 4 projects. This suggests that 4 projects might be the standard or average workload assigned to employees in the dataset. There are 2,388 employees who have worked on only 2 projects. While this is less common than working on 3 or 4 projects, it still represents a significant portion of the workforce, possibly indicating newer employees or those with lighter workloads. A smaller number of employees handle 6 or 7 projects, with 1,174 and 256 employees, respectively. This suggests that handling such a high number of projects is less common, likely reserved for more experienced employees or those in specialized roles.

4. average\_montly\_hours:

A graph of blue bars

Description automatically generated

Figure 7 Average monthly hours of employees

The histogram displays the distribution of average monthly hours worked by employees, with the x-axis representing the range of average monthly hours and the y-axis showing the count of employees within each range. The distribution is notably bimodal, with two peaks around 150-200 hours and 250-300 hours, indicating that a significant number of employees fall within these hour ranges. The highest count is observed at around 150-200 average monthly hours, where over 1,200 employees are concentrated. There are fewer employees working either very low or very high average monthly hours, as seen in the lower counts at both extremes of the distribution. The overall shape suggests that most employees tend to work within a moderate to high number of hours monthly, with fewer working extremely low or high hours.

5. time\_spend\_company:

A graph of a number of blue bars

Description automatically generated with medium confidence

Figure 8 Time spent at the company by the employees in years

This histogram shows the distribution of time spent at the company by employees, with the x-axis representing years of service and the y-axis showing the number of employees within each category. The data is heavily right-skewed, with most employees having spent between 2.5 and 5 years at the company. The largest group is employees with around 3 years at the company, totaling 6,443 individuals. There are fewer employees with more extended tenures, and the number significantly drops for those who have spent more than 7.5 years at the company. The sharp decline in the number of employees after 5 years suggests a high turnover rate after a few years of service.

6. work\_accident:

A graph with a blue rectangle

Description automatically generated

Figure 9 Work accidents

This bar chart depicts the distribution of work accidents among employees, with the x-axis indicating whether an accident occurred (0 for no, 1 for yes) and the y-axis representing the number of employees in each category. The vast majority of employees, 12,830, have not experienced a work accident, while a smaller group of 2,169 employees have reported a work accident. The significant difference between these two groups suggests that work accidents are relatively rare in this organization.

7. left:

A graph with a blue rectangle

Description automatically generated

Figure 10 Did the employee leave the company?

This bar chart illustrates the distribution of employees who have left the company, with the x-axis indicating whether an employee left (0 for no, 1 for yes) and the y-axis representing the count of employees in each category. The majority of employees, 11,428, have not left the company, while 3,571 employees have left. This indicates a substantial retention rate, although a significant number of employees have also departed. The ratio suggests that while most employees stay, the turnover is not negligible and could be an area of concern for the organization.

8. promotion\_last\_5years:

A blue rectangular object with black text

Description automatically generated

Figure 11 Promotion in the last 5 years

This bar chart shows the distribution of employees who have received a promotion in the last five years. The x-axis indicates whether an employee was promoted (0 for no, 1 for yes), and the y-axis represents the number of employees in each category. The chart reveals that the overwhelming majority of employees, 14,680, did not receive a promotion, while only 319 employees were promoted during this period. This significant disparity suggests that promotions are rare within the company, potentially indicating limited upward mobility or stringent promotion criteria.

9. sales:

A graph with blue squares

Description automatically generated

Figure 12 Number of employees working the respective departments

This bar chart represents the distribution of employees across different departments within the company. The x-axis lists the various departments, while the y-axis shows the number of employees in each department. The sales department has the highest number of employees, with 4,140 individuals, followed by the technical and support departments, with 2,720 and 2,229 employees, respectively. Departments like IT, product management, and marketing have a moderate number of employees, ranging between 630 and 1,227. The departments with the fewest employees are HR, R&D, and management, each with fewer than 800 employees. This distribution indicates that the company has a strong focus on sales, technical, and support functions, while other departments are comparatively smaller in size.

10. salary:

A graph of a salary level

Description automatically generated with medium confidence

Figure 13 Salary level of employees

This bar chart shows the distribution of employees across different salary levels, categorized as high, medium, and low. The x-axis represents the salary level, while the y-axis shows the count of employees in each category. Most employees fall into the low salary category, with 7,316 employees, followed by the medium salary category with 6,446 employees. The high salary category has the fewest employees, with only 1,237 individuals. This distribution indicates that a large proportion of the workforce is compensated at the lower end of the salary scale, with fewer employees receiving higher salaries.

# PREDICTOR RELEVANCY:

## Satisfaction Level:

A graph of blue and purple lines

Description automatically generated

Figure 14 Satisfaction level of employees

The density plot shows how employee satisfaction levels differ between those who stayed with the company (represented by the blue line) and those who left (represented by the purple line). Employees who left tend to have two peaks in their satisfaction levels: one at very low levels (around 0.1 to 0.2) and another at moderate levels (around 0.5). In contrast, employees who stayed mostly have higher satisfaction levels, with a peak between 0.7 and 0.9. This pattern suggests that lower satisfaction is closely linked to employee turnover, while higher satisfaction levels are associated with retention. But the employees with satisfaction level between 0.2 and 0.3 and also between 0.6 and 0.7 did not leave the company figuring out the reason for their retention could help the company to know what factors are influencing them to stay in the company and from this understanding we can reduce or at least try to minimize the turnover rate for the other three peaks of the employees, Overall, the plot highlights that satisfaction plays a crucial role in whether employees choose to stay or leave the company.

## Last Evaluation:

A graph of a graph

Description automatically generated with medium confidence

Figure 15 Last Evaluation

This density plot displays the distribution of the last evaluation scores for employees, categorized by whether they left the company (1) or stayed (0). Employees who left (purple area) have two prominent peaks: one at lower evaluation scores (around 0.4) and another at higher scores (above 0.8). In contrast, employees who stayed (cyan area) have a more consistent distribution, peaking around moderate evaluation scores (0.6 to 0.7). This suggests that employees who received very low or very high evaluations are more likely to leave, while those with moderate evaluations tend to stay. The plot indicates that extreme performance ratings—either perceived as underperforming or exceptionally performing—are associated with higher turnover rates.

## Number of Projects:

A graph showing a number of projects

Description automatically generated

Figure 16 Number of Projects

This density plot illustrates the number of projects handled by employees, differentiated by whether they left the company (1) or stayed (0). Employees who stayed (cyan area) show peaks at 3 and 4 projects, suggesting that those who manage a moderate number of projects are more likely to remain with the company. Conversely, employees who left (purple area) have a broad distribution with noticeable peaks at both the lower end (2 projects) and the higher end (5 to 7 projects), indicating that both underworked and overworked employees are more prone to turnover. This pattern highlights that a balanced workload is key to employee retention, while extremes in project counts, whether too few or too many, may drive employees to leave.

## Average Monthly Hours:

A graph showing a blue and purple line

Description automatically generated with medium confidence

Figure 17 Average Monthly Hours

This density plot displays the distribution of average monthly hours worked by employees, categorized by whether they left the company (1) or stayed (0). Employees who left (purple area) show two prominent peaks: one at around 150 hours and another closer to 270 hours per month, indicating that both very low and very high workloads are associated with higher turnover rates. Conversely, employees who stayed (cyan area) tend to have a peak around 200 hours, suggesting that those with moderate workloads are more likely to remain with the company. The plot highlights that too few or too many working hours, contribute to employee turnover, while a balanced workload helps in retaining employees. This suggests that managing employee work hours carefully can be a key factor in improving retention rates.

## Time Spent in Company:

A graph of a company

Description automatically generated

Figure 18 Time Spent in Company

This density plot illustrates the distribution of time spent at the company by employees, categorized by whether they left (1) or stayed (0). Employees who left the company (purple area) show notable peaks around 3, 4, and 5 years, indicating higher turnover rates at these intervals. In contrast, employees who stayed (cyan area) show a broader distribution with less peaks, suggesting more consistent retention over different lengths of service. The sharp peaks for employees who left suggest that turnover is most common among employees with a few years of experience, particularly around the 3 to 5-year mark. This pattern indicates that mid-tenure employees may be at a higher risk of leaving, perhaps due to unmet expectations or lack of advancement opportunities.

## Work Accident:

A blue and black graph

Description automatically generated with medium confidence

Figure 19 Work Accident

This bar chart shows the proportion of employees who left the company based on whether they experienced a work accident (0 = No, 1 = Yes). Among employees who did not experience a work accident, 22.7% (3,402 employees) left the company, while 62.9% (9,428 employees) stayed. For those who did experience a work accident, only 1.1% (169 employees) left the company, whereas 13.3% (2,000 employees) stayed. This indicates that the likelihood of employees leaving the company is significantly lower among those who had a work accident, suggesting that work accidents may have a less direct effect on turnover compared to other factors.

## Promotion in Last 5 Years:

A blue and green rectangular shapes

Description automatically generated with medium confidence

Figure 20 Promotion in Last 5 Years

This bar chart shows the proportion of employees who left the company based on whether they received a promotion in the last five years (0 = No, 1 = Yes). Among employees who were not promoted, 23.7% (3,552 employees) left, while 74.2% (11,128 employees) remained. For those who were promoted, only 2.0% (300 employees) left, while 1.1% (19 employees) stayed. This data clearly indicates that the likelihood of leaving the company is significantly lower for employees who received a promotion, suggesting that promotions are an effective strategy for retaining employees. It highlights the value of recognizing and rewarding employee contributions to enhance retention rates.

## Department (Sales):

A graph of blue bars

Description automatically generated with medium confidence

Figure 21 Department (Sales)

This bar chart shows the proportion of employees who left the company by department, with "0" indicating employees who stayed (cyan) and "1" indicating those who left (blue). The highest turnover rates are observed in the HR department, with 29.09% of employees leaving, followed by the technical (25.62%), support (24.90%), and sales (24.49%) departments. In contrast, the management department has the lowest turnover rate at 14.44%, indicating higher retention in this group. More than 50% of the company employees are from Sales, Support and Technical departments, so having high turnover rates in these departments impacts the company.

## Salary:

A graph of a number of employees

Description automatically generated

Figure 22 Salary

This bar chart illustrates the proportion of employees who left the company based on their salary level, categorized as high, medium, and low. Employees with low salaries have the highest turnover rate, with 14.48% leaving the company. Employees with medium salaries also show a relatively high turnover rate of 8.78%. In contrast, employees with high salaries have the lowest turnover rate, with only 0.55% leaving the company. This suggests that lower salary levels are strongly associated with higher employee turnover, while higher salaries are more likely to retain employees. The data underscores the impact of compensation on employee retention, highlighting the importance of competitive pay to reduce turnover rates.

# DIMENSION REDUCTION:

RANDOM FOREST FOR DIMENSION\_REDUCTION:

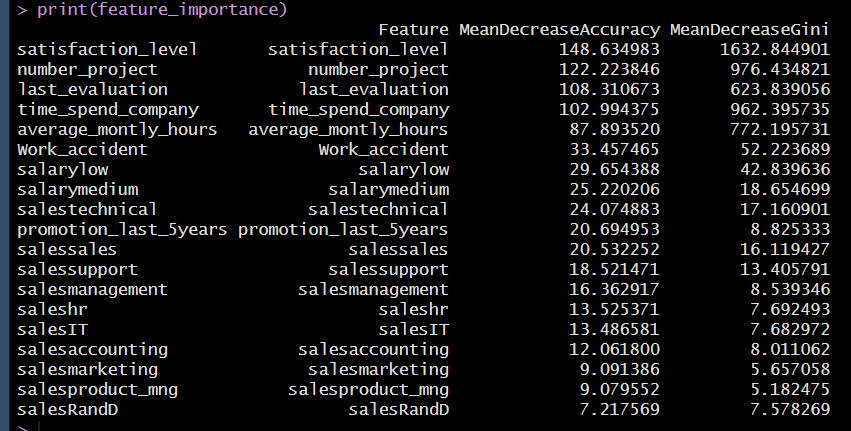


Figure 23 Dimension Reduction

From the above observation, no dimension is needed to be excluded from the analysis.

# DATA TRANSFORMATION:

No data transformation is needed.

# DATA PARTITION:

In a dataset consisting of 14,999 records, the data is partitioned into three key subsets to ensure effective model training, fine-tuning, and evaluation. Sixty percent of the data, or 8,999 records, is allocated for training the model, allowing it to learn from the majority of the available information. Another 20%, or 2,999 records, is designated for validation, where the model is fine-tuned and adjusted to avoid overfitting.

The final 20% of the data, or 2,999 records, is reserved as the test or holdout set. This portion remains untouched during training and validation and is only used at the end to evaluate how well the model generalizes to unseen data.

# TRANSLATING GOALS:

Business Goal 1: Identify Key Factors for Employee Turnover

* Model Used: Random Forest
* Reason for Using Random Forest:
  + Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their results, making it highly suitable for identifying the most important factors responsible for employee turnover. Unlike KNN, Random Forest is not solely focused on classification but also provides insights into *why* employees may be leaving by determining the importance of each feature in predicting turnover.
  + Random Forest provides a feature importance metric, ranking variables by their contribution to turnover prediction. For instance, it can highlight factors such as satisfaction level, time spent at the company, and salary as key drivers of turnover. This is essential for business decision-makers, as it offers actionable insights to inform HR and management strategies.
  + Additionally, Random Forest handles both numerical and categorical variables well and can manage large datasets without overfitting. The use of Random Forest enables both accuracy in predictions and transparency in understanding the key factors driving turnover.

Business Goal 2: Predict Employee Turnover

* Model Used: K-Nearest Neighbors (KNN)
* Reason for Using KNN:
  + The KNN algorithm is a commonly used classification model that works well when the objective is to predict an outcome based on historical data. In this case, KNN is appropriate for predicting employee turnover because it classifies employees based on their similarity to other employees who either stayed or left the company.
  + KNN operates by identifying the ‘k’ most similar instances (neighbors) in the dataset and making a prediction based on the majority label of those neighbors (whether they left or stayed). Given the dataset size of 15,000 records, KNN effectively leverages the similarity between employee data points such as satisfaction level, number of projects, and work hours to make predictions. It is particularly useful for problems where no assumptions about data distribution are required, making it a flexible method to predict turnover.
  + Additionally, KNN handles complex relationships between variables without needing to tune many parameters. However, ensuring that the data is pre-processed and scaled is crucial for KNN to avoid bias toward larger numeric values, such as 'average\_montly\_hours.'

# MODEL SELECTION:

The choice of models for determining key factors and predicting employee turnover in this project is driven by the specific strengths and relevance of each technique to the objectives.

**1. Models to Determine Key Factors for Turnover:**

**Random Forest:** Random Forest is selected for its ability to rank the importance of variables, which is critical for identifying the most significant factors influencing employee turnover. By building multiple decision trees and averaging their results, Random Forest can capture complex interactions between variables, providing a robust understanding of which factors, such as satisfaction level or time spent at the company, most strongly impact turnover decisions.

**Decision Tree:** The Decision Tree model is chosen for its clear interpretability and visualization capabilities. Its structure enables straightforward identification of the key factors driving turnover, allowing the business to understand the decision paths that lead to employees leaving the company. Decision Trees are also versatile, handling both categorical and numerical variables effectively, making them highly suited for exploring turnover causes.

**2. Models for Predicting Turnover:**

**Logistic Regression:** Logistic Regression is used to predict employee turnover due to its effectiveness in binary classification tasks. It not only provides accurate predictions about whether an employee will leave but also offers interpretable coefficients that indicate the strength and direction of influence for each factor, such as salary level or promotion status, on the likelihood of turnover. And it also gives the probability which can be used by HR to change the thresholds so that they can act early in retaining the employee.

**K-Nearest Neighbors (KNN):** KNN is applied in this project to capture potential non-linear relationships between variables affecting turnover. By comparing employees to their nearest peers based on similarities in their attributes, KNN offers a flexible approach to predicting turnover in situations where no clear linear relationship exists between the factors influencing employee retention and departure.

**Naive Bayes:** Naive Bayes is selected for its computational efficiency and ability to handle categorical data, which is prevalent in the dataset (e.g., department, salary). Despite its assumption of independence between features, Naive Bayes can provide strong predictive results, making it a useful model for quick and effective classification of turnover risk.

# MODEL FITTING:

1. Models to Determine key factors for turnover:
2. Random Forest

A screenshot of a computer screen

Description automatically generated

Figure 24 Random Forest for determining key factors for turnover

The most important predictor of employee turnover in the model is satisfaction level, indicating that how satisfied an employee is has the largest impact on whether they leave. The number of projects is another key factor, suggesting that workload plays a big role in turnover. Last evaluation scores and time spent at the company also significantly influence turnover, highlighting that performance and tenure are important considerations. Additionally, average monthly hours worked affects turnover, suggesting that work-life balance is a critical factor. These insights show the primary drivers behind employee retention and turnover.

1. Decision Tree

A diagram of a decision tree

Description automatically generated

Figure 25 Decision Tree for determining key factors for turnover

This decision tree for employee turnover identifies satisfaction level as the most significant factor influencing whether an employee leaves, as it is the first split in the tree. Employees with satisfaction levels below 0.47 have a higher likelihood of leaving the company, especially when they have fewer projects or a low last evaluation score. On the other hand, employees with a satisfaction level above 0.47 are more likely to stay, particularly if they have worked fewer than 5 years or have a high last evaluation. Time spent at the company and average monthly hours are additional factors influencing employee turnover, with employees working more hours or having longer tenures being less likely to leave. This tree visually highlights how various factors, particularly job satisfaction and performance metrics, interact to influence turnover decisions.

Observation: Both the models almost states the same predictors as important for analysis.

1. MODELS FOR PREDICTING TURNOVER:
2. Logistic Regression

A screenshot of a computer

Description automatically generated

Figure 26 Logistic Regression for Predicting Turnover

This logistic regression model summary highlights the significant predictors of employee turnover. The negative coefficients for satisfaction\_level and number\_project indicate that lower satisfaction and fewer projects increase the likelihood of an employee leaving the company. In contrast, a positive coefficient for last\_evaluation suggests that employees with higher evaluation scores are more likely to stay. Variables like average\_monthly\_hours and time\_spend\_company also play an important role, where more time spent at the company and working more hours increase the likelihood of turnover. Salary levels, specifically low and medium, have a significant positive relationship with leaving, indicating that employees with lower salaries are more likely to leave. The presence of a work accident or promotion in the last five years significantly reduces the probability of turnover.

And

A screenshot of a computer screen

Description automatically generated

Figure 27 Logistic Reg Confusion matrix

The model's accuracy, shown by the confusion matrix, stands at 79.17%, meaning it correctly classified 79% of the validation data. While this is a fairly solid result, the model's ability to identify employees likely to leave (sensitivity) is much lower at 35.71%. This indicates that the model struggles to catch those who may exit the company. On the other hand, the model does quite well at identifying employees who are likely to stay, with a specificity of 92.74%.

1. KNN

A screenshot of a computer screen

Description automatically generated

Figure 28 KNN Confusion matrix

In this model, the accuracy is 94.33%, which indicates that the model correctly classified the majority of the cases in the validation dataset. The sensitivity, which measures the model’s ability to correctly identify employees who will leave the company, is 89.92%. This is significantly better than the previous model, demonstrating that it is quite good at detecting employee turnover. The specificity, or the ability to correctly identify employees who will stay, is even higher at 95.71%, meaning the model is excellent at predicting those who will remain. The Kappa score of 0.8457 indicates strong agreement between the predicted and actual classifications. Overall, the model performs well on both classes, with a balanced accuracy of 92.81%.

1. Naïve Bayes

A screenshot of a computer screen

Description automatically generated

Figure 29 Naive bayes confusion matrix

In this confusion matrix, the model's overall accuracy is 78.57%, meaning it correctly classified about 79% of the validation data. The sensitivity, or the model's ability to correctly identify employees who are likely to leave, is 78.43%, showing a solid performance in this regard. Specificity, which measures how well the model identifies those who will stay, is also close at 78.61%, indicating balanced accuracy in predicting both outcomes. The positive predictive value (precision) is lower at 53.38%, implying that while the model catches most of those who leave, it also misclassifies some who actually stay. Overall, the balanced accuracy of 78.52% suggests that this model performs reasonably well for both predicting turnover and retention but could still benefit from further fine-tuning.

# MODEL EVALUATION:

Comparison of accuracy, sensitivity and specificity on validation data

Logistic Regression, KNN, and Naïve- bayes.

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluation Metrics | Logistic Regression | KNN | Naïve Bayes |
| Accuracy | 79.83 | 93.5 | 71.12 |
| Sensitivity | 35.85 | 86.69 | 79.97 |
| Specificity | 93.56 | 95.62 | 68.36 |

**KNN** has performed well predicting employees left as employees left and employees not left as employees not left with an accuracy of 93.5 and sensitivity of 86.69 and specificity of 95.62. So, the further analysis is carried out using KNN on hold out set and the results are

A screenshot of a computer screen

Description automatically generated

Figure 30 KNN confusion matrix on holdout set

# HYPOTHESIS ANALYSIS:

**First Hypothesis**

***The first hypothesis: salary is the reason why the employees left the company. Let's see if is this correct***

***A graph of a number of employees

Description automatically generated***

Salary might be the reason, that is why employees with the low salary had high attrition rate.

**Second Hypothesis**

A blue and black graph

Description automatically generated with medium confidence

***The second hypothesis is: employees leave the company because work is not safe.***

It’s not true, because 22.7% of the employees left the company even though the work is safe.

**Third Hypothesis**

A blue and green rectangular shapes

Description automatically generated with medium confidence***The third hypothesis is: this company is a good place to grow professionally.***

Company has a promotion rate of 2.01% which indicates that it’s not a good place to grow professionally.

# CONCLUSION AND FUTURE WORK:

The analysis conducted provides valuable insights into employee turnover and the factors driving it. By using Random Forest and Decision Tree models, it is clear that satisfaction level, number of projects, and last evaluation are the key factors influencing employee turnover. These findings suggest that employees with lower satisfaction and higher work demands are more likely to leave the company. Addressing these factors through initiatives like workload management and employee engagement programs can help improve retention rates. For predicting employee turnover, Logistic Regression, Naive Bayes, and KNN were used to classify potential leavers with reasonable accuracy. Among these, KNN provided clear insights into the likelihood of turnover based on various predictors.

# BUSINESS RECOMMENDATIONS:

Based on the findings from this analysis, the following business recommendations are made to improve employee retention and manage workforce stability effectively:

**1. Focus on Employee Satisfaction:** The analysis shows that satisfaction level is a key factor influencing turnover. Regularly monitor employee satisfaction through surveys or feedback systems, and address areas of concern promptly to maintain a positive work environment.

**2. Balance Workload and Project Management:** Employees handling a higher number of projects and working long hours are more likely to leave. It is recommended to assess and distribute workloads more evenly across teams, ensuring that no single group is overwhelmed.

**3. Offer Career Growth Opportunities:** Employees with higher evaluations and more time spent at the company tend to stay. Developing clear career paths, offering promotions, and supporting professional development can encourage long-term employee commitment.

**4. Implement Targeted Retention Programs:** For departments with higher turnover rates (as identified in the analysis), implement specific retention programs, such as mentorship, team-building activities, and recognition programs to boost morale.

**5. Salary Adjustments and Benefits:** Employees in lower salary bands showed higher turnover rates. Consider revisiting compensation structures, particularly for roles with high turnover, and ensure that benefits packages remain competitive to retain talent.

**6. Proactive Employee Turnover Monitoring:** Use predictive models, such as the ones developed, to continuously monitor turnover risk. Identifying employees likely to leave early allows HR teams to intervene and address concerns before they result in resignations.

**7. Promote Work-Life Balance**: Employees working long hours or spending too much time at the company without adequate breaks were prone to leave. Establish policies promoting work-life balance, such as flexible work hours, remote work options, and wellness programs.

**8. Engage with High-Risk Departments**: Departments like sales or technical roles may need additional attention. Regularly check in with these teams and offer them resources for stress management and skill development.

**9. Continuous Performance Evaluation:** Employees with lower evaluation scores are more likely to leave. Providing regular, constructive feedback and opportunities for improvement can keep employees engaged and help them grow within the organization.

**10. Use Data-Driven HR Strategies:** Continue leveraging data and machine learning models to track turnover trends, identify areas for improvement, and measure the effectiveness of retention strategies over time.