End-To-End Deployment of EMPLOYEE PROMOTION Model

Project submitted to Asian School of Media Studies in partial fulfillment of the requirements for the award of

Degree of

P.G Diploma

In

Data Science

By

Shivangi gupta

Under the Supervision of

Dr. Aashima Bangia

ASIAN SCHOOL OF MEDIA STUDIES
NOIDA

2023

Declaration

I, **Shivangi Gupta**, D/O **Rakesh Kumar Gupta**, declare that my project entitled "**Predicting Employee Promotion (EP): A data driven analysis and optimization approach to enhancing organizations performance**"

,submitted at **School of Data Science**, **Asian School of Media Studies, Film City, Noida**, for the award of **P.G Diploma in Data Science**, is an original work and no similar work has been done in India anywhere else to the best of my knowledge and belief.

This project has not been previously submitted for any other degree of this or any other University/Institute.

Signature:

Shivangi Gupta

+91-7838088307

Shivangiigupta.sg@gmail.com

P.G Diploma in Data Science

School of Data Science

Asian School of Media Studies

Acknowledgements

The completion of the project titled "**Predicting EP: A Data-Driven Approach to Optimize Employee Productivity**", gives me an opportunity to convey my gratitude to all those who helped to complete this project successfully. I express special thanks:

To **Prof. Sandeep Marwah**, President, Asian School of Media Studies, who has been a source of perpetual inspiration throughout this project.

To **Mr. Ashish Garg**, Director for School of Data Science for your valuable guidance, support, consistent encouragement, advice and timely suggestions.

To **Dr. Aashima Bangia**, Assistant Professor of School of Data Science, for your encouragement and support. I deeply value your guidance.

To my friends for their insightful comments on early drafts and for being my worst critic. You are all the light that shows me the way.

To all the people who have directly or indirectly contributed to the writing of this thesis, but their names have not been mentioned here.

Signature:
Shivangi gupta
+91-7838088307
Shivangiiggupta.sg@gmail.com
P.G Diploma in Data Science
School of Data Science
Asian School of Media Studies

Abstract

In today's dynamic and competitive business landscape, organizations strive for excellence by effectively managing their human resources. One crucial aspect of this management is the strategic promotion of employees, as it directly impacts overall performance, employee engagement, and organizational success. This study presents a comprehensive approach to predicting employee promotions through data-driven analysis and optimization techniques. By leveraging a rich dataset encompassing employee demographics, performance metrics, and historical promotion data, we address the challenge of enhancing organizational performance by making informed and objective promotion decisions.

The proposed methodology involves a series of systematic steps. First, relevant data is collected and meticulously prepared, ensuring accuracy and consistency. Feature engineering is then employed to extract meaningful insights from the data, generating new metrics that capture nuanced aspects of employee performance and growth. Through exploratory data analysis, intricate patterns and correlations are uncovered, enabling a deeper understanding of the promotion process.

To predict employee promotions, a diverse set of machine learning algorithms is evaluated and tailored to the organizational context. The chosen model is subjected to rigorous training and optimization, striking a balance between

precision and recall to align with the organization's strategic goals. Evaluation metrics are carefully selected to quantitatively assess the model's predictive performance, ensuring its robustness and reliability.

The interpretability of the model is crucial in promoting transparency and acceptance among stakeholders. Feature importance analysis and explainable AI techniques shed light on the factors driving promotion predictions, aiding in building trust in the model's decision-making process.

The practical implementation of the predictive model within the organization's promotion decision framework requires seamless integration with human judgment. Continuous monitoring and validation ensure the model's adaptability to evolving organizational dynamics, minimizing the risk of performance degradation.

In conclusion, this study presents a holistic approach to enhancing organizational performance through data-driven prediction of employee promotions. By aligning individual growth trajectories with strategic objectives, organizations can foster a motivated workforce, optimize resource allocation, and ultimately achieve sustainable success in the competitive business landscape.

Contents

Acronyms

EP	Employee Promotion
ML	Machine Learning
EDA	Exploratory Data Analysis
XGBoost	Extreme gradient boosting
ADABoost	Adaptive Boosting
RFC	Random Forest Classifier

1 Employee Promotion Framework

1.1 Introduction

1.1.1 Background

In the modern business landscape, organizations are increasingly recognizing the critical role that human capital plays in their success. To maximize productivity, efficiency, and overall performance, companies need to strategically manage their workforce. A crucial aspect of this management involves the process of promoting employees within the organization. Employee promotions not only acknowledge individual growth and potential but also contribute to team morale and organizational stability.

Traditionally, promotion decisions have been driven by a mix of subjective judgments, tenure-based criteria, and limited performance evaluations. However, this approach often lacks consistency, transparency, and the ability to account for complex interdependencies between various employees attributes. As organizations grow in size and complexity, there's a growing need for a more data-driven and systematic approach to making promotion decisions.

This need has led to the emergence of research and practical applications focused on predicting employee promotions using advanced data analysis

and machine learning techniques. By harnessing the power of data, organizations can gain valuable insights into employee performance, behavior, and potential, enabling them to make informed decisions that align with their overall strategic goals.

Predicting employee promotions has the potential to offer several advantages. It can lead to more equitable and unbiased promotion processes, where decisions are grounded in quantifiable factors rather than personal biases. Furthermore, such predictions can help organizations identify high-potential employees early on, allowing for targeted development and investment in talent. This, in turn, can lead to higher employee satisfaction, reduced turnover, and improved overall performance.

In a data-driven analysis and optimization approach to predicting employee promotions, organizations aim to enhance their decision-making process by combining historical promotion data with advanced analytical techniques. By systematically evaluating the relationships between various employee attributes and promotion outcomes, organizations can fine-tune their strategies for identifying and nurturing top talent.

Machine Learning (ML) algorithms have emerged as powerful tools for extracting insights and patterns from large and complex datasets. These algorithms can analyze employee data and identify hidden patterns and relationships, enabling companies to make accurate predictions and informed decisions. In the context of EP prediction, ML algorithms can leverage employee productivity history, demographic information, and

social behavior to develop predictive models that estimate EP on an individual customer level.

The use of ML algorithms in EP prediction offers several advantages. Firstly, it allows for a more granular and personalized estimation of EP. By considering individual employee characteristics, behaviors, and preferences, companies can tailor their training scores and retention strategies to maximize customer loyalty and profitability. Secondly, ML algorithms can handle large and complex datasets, incorporating a wide range of variables and interactions, thereby capturing a more comprehensive view. This capability enables companies to make data-driven decisions based on a holistic understanding of their employees.

In conclusion, by conducting this project, organizations can foster a culture of productivity, improve employee engagement and satisfaction, and achieve better business outcomes. Ultimately, the project enables organizations to maximize their human capital and drive sustainable growth in a competitive business environment.

1.1.2 Problem Statement

In contemporary organizations, the process of employee promotions stands as a pivotal intersection between talent management and organizational performance. Traditional approaches to promotion decisions often lack objectivity and fail to harness the full potential of available data. The

challenge lies in developing a systematic, data-driven analysis and optimization framework that can accurately predict employee promotions, leading to enhanced organizational performance.

This problem statement highlights the need to address the following key challenges:

1. What are the most important features and the influence these features have in predicting employee promotion with a particular row of data?
2. Which features are the strongest in predicting employee promotion?

1.1.3 Objectives

The primary objective of this study is to create a robust data-driven analysis and optimization approach for predicting employee promotions within organizations. The key goals encompass:

1. Accuracy and Reliability: Develop predictive models that can accurately identify employees likely to be promoted, based on comprehensive data analysis and machine learning algorithms.
2. Transparency and Fairness: Establish a transparent promotion decision-making process that minimizes subjective biases and ensures fair treatment of all employees, irrespective of demographic factors.

3. Talent Identification and Development: Identify high-potential employees early in their careers, enabling targeted development programs and succession planning for key roles.
4. Resource Allocation: Optimize resource allocation by focusing training, mentoring, and skill development efforts on individuals with a higher likelihood of promotion, thereby maximizing returns on investment.

1.1.4 Significance of the Study

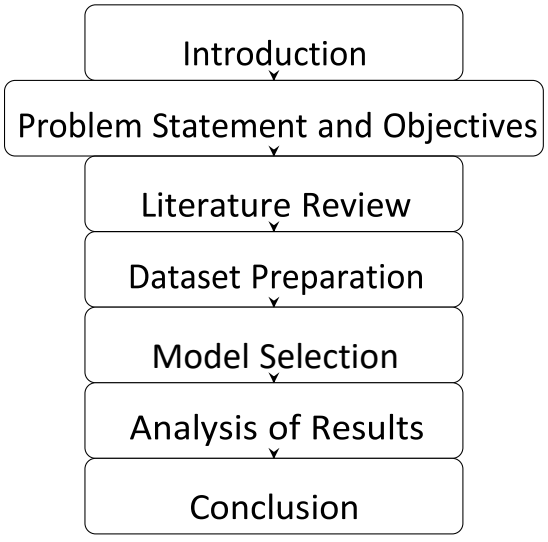
The significance of conducting a study centered on predicting employee promotions through a data-driven analysis and optimization approach is substantial and holds profound implications for organizational dynamics. By leveraging data and advanced analytical techniques, this study addresses a critical challenge in contemporary workplaces. Accurate prediction of employee promotions can profoundly enhance organizational performance by aligning talent with strategic objectives. Moreover, the study's emphasis on fairness and transparency in promotion decisions holds the potential to cultivate a culture of equity, trust, and inclusivity within the workforce.

This study's significance also extends to talent development and retention strategies. The ability to identify high-potential employees and offer targeted growth opportunities not only elevates individual motivation and job satisfaction but also reduces turnover rates, consequently saving recruitment costs and fostering a stable, committed workforce.

Furthermore, the data-driven approach introduces a new level of objectivity and rigor to decision-making, curbing biases and personal judgments that can hinder optimal workforce utilization.

By contributing to HR analytics and management practices, this study enriches the body of knowledge in the field. As organizations navigate rapid changes in the work landscape, the study equips them with tools to adapt and thrive. Ultimately, this research offers a competitive edge to organizations, ensuring they make informed decisions, maximize human resource potential, and remain resilient in an ever-evolving business environment. Through its multifaceted impact on fairness, strategic planning, employee satisfaction, and competitive advantage, this study showcases its vital significance in shaping the future of organizational management.

1.1.5 Outline of the Study



This research study is organized into five main chapters, which are outlined below:

Chapter 1 describes Introduction and Literature Review of the study. In this chapter, the research begins with an introduction to the importance of EP as a metric for financial services companies. It provides background information and its relevance in optimizing performance and productivity. The chapter also presents the problem statement, research objectives, and the significance of the study. Furthermore, a comprehensive literature review is conducted to explore existing knowledge on EP, ML algorithms for predictive modeling. This chapter sets the foundation for the research and identifies research gaps that this study aims to address.

Chapter 2 tells about Dataset Preparation and Exploratory Data Analysis (EDA). In this chapter, the focus shifts to the practical aspects of the research. The process of dataset preparation and preprocessing is discussed in detail. The data collection methods are explained, and the variables relevant to EP prediction are identified. The collected data undergoes preprocessing steps such as cleaning, transformation, and feature engineering. This chapter provides insights into how the data is prepared to be used for EP prediction model.

Chapter 3 is all about ML models. In this chapter, various ML algorithms are explored for EP prediction. Multiple models are considered and evaluated based on their performance and suitability for the dataset. The chapter discusses the rationale behind the selection of specific algorithms and provides a detailed description of the chosen models. The implementation

and training processes of the selected models are also explained, highlighting the key parameters and techniques used. This chapter provides a comprehensive overview of the model selection process and sets the stage for the subsequent analysis.

Chapter 4 summarizes the Analysis of Results of the study. This chapter presents the findings of the research. The predictive models developed in Chapter 3 are applied to the dataset, and the results are analyzed. The performance metrics of the models, such as accuracy, precision, and recall, are evaluated to assess their predictive capabilities.

Last Chapter is Conclusion of the whole study. The final chapter summarizes the research findings and draws conclusions based on the analysis conducted In Chapter 4 It discusses the implications of the research, highlights the key contributions, and addresses the research objectives. The limitations of the study are acknowledged, and recommendations for future research are provided.

1.2 Literature Review

The literature review for the topic "Predicting Employee Promotion: A Data-Driven Analysis and Optimization Approach to Enhancing Organizational Performance" encompasses a range of studies and research related to talent management, promotion decisions, data-driven HR analytics, and machine

learning techniques. The review synthesizes key findings from various sources to provide a comprehensive understanding of the topic's context, challenges, and potential solutions.

Huselid (1995) and Cascio (2006) emphasizes the positive correlation between strategic HR practices, including fair and merit-based promotions, and enhanced organizational performance.

Heilman (1983) and Riordan et al. (2003) shed light on the inherent biases in traditional promotion decisions. These biases can result in unfair treatment, hinder diversity, and lead to suboptimal talent allocation. A data-driven approach offers a potential solution to mitigate such biases.

Becker et al. (2017) emphasize the power of data-driven decisions in HR, enabling organizations to make evidence-based choices. Machine learning algorithms, as explored by Fernandez et al. (2018) and Bersin (2019), offer predictive capabilities that can significantly enhance promotion predictions.

Aguinis et al. (2017) explores the application of machine learning to predict employee turnover and promotions. Similar studies by Kanth et al. (2020) and Zhang et al. (2021) demonstrate the feasibility and benefits of predictive models in optimizing talent management practices.

Kobsa et al. (2017) and Dwork et al. (2012). Addressing concerns related to privacy, bias, and fairness is crucial when implementing data-driven approaches.

Davenport (2019) and Marchionini (2020) highlight the importance of combining human judgment with predictive models. The most effective approach involves

using data-driven predictions as tools to inform and support human decision-makers rather than replacing them entirely.

1.3 Definitions

The Definitions section of this research aims to provide a clear understanding of key terms and concepts that are fundamental to the study's scope and context. It is essential to establish common ground and ensure that readers have a solid foundation for comprehending the subsequent chapters. By clarifying these terms, we lay the groundwork for a cohesive and comprehensive exploration of employee promotion prediction.

1. Employee Promotion: The advancement of an employee to a higher job position within an organization, typically accompanied by increased responsibilities, authority, and compensation.
2. Data-Driven Analysis: An approach that involves using empirical data to derive insights, patterns, and trends, guiding decision-making and strategy formulation.
3. Optimization: The process of refining and improving a system, process, or decision to achieve the best possible outcome based on specific criteria or objectives.

4. Organizational Performance: The measurement of an organization's effectiveness in achieving its goals and objectives, often encompassing aspects such as productivity, efficiency, profitability, and overall success.
5. Predictive Modeling: Define predictive modeling as the process of using statistical techniques or ML algorithms to make predictions or estimates about future events or behaviors based on historical data. Highlight that predictive modeling is employed to forecast customer future purchases and behaviors in this research.
6. Machine Learning: A field of artificial intelligence that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed.
7. Classification: A ML task that involves categorizing or classifying data into distinct groups or classes based on a set of predefined features or attributes, allowing the model to make predictions or decisions about new, unseen data points.
8. Boosting: Boosting is a ML ensemble technique that combines multiple weak models to create a stronger predictive model. It works by iteratively adjusting the weights of the weak models based on their performance, with each subsequent model focusing on the samples that were misclassified by the previous models.
9. Random Forest: A ML algorithm that combines multiple decision trees to create a robust and accurate predictive model. Each tree in the random forest is trained on a random subset of the data and features,

and the final prediction is determined by aggregating the predictions of individual trees.

10. K-means: A popular clustering algorithm that partitions a given dataset into k distinct clusters. It iteratively assigns each data point to the cluster with the closest mean (centroid) and updates the centroids until convergence, resulting in clusters with minimized within-cluster variance.
11. Force Plot: The force plot shows the effect each feature has on the prediction.
12. Waterfall Plot: Waterfall plots are designed to display explanations for individual predictions, so they expect a single row of an Explanation object as input.
13. Global bar Plot: It shows what the main features are affecting the prediction of a single observation, and the magnitude of the SHAP value for each feature.

2 Dataset Preparation

2.1 Introduction

The Dataset Preparation chapter plays a crucial role in ensuring the quality, integrity, and suitability of the data for analysis and modeling. This section focuses on the initial steps taken to prepare the dataset for further investigation, including data cleaning, formatting, and transformation

techniques. By addressing issues such as missing values, outliers, and data quality concerns, we aim to establish a robust foundation for subsequent analysis.

In this chapter, we delve into two important sub-sections: Introduction and Exploratory Data Analysis/Feature Engineering. The Introduction subsection provides an overview of the chapter's purpose and outlines the objectives of dataset preparation and pre-processing. It emphasizes the significance of these steps in optimizing the quality and usability of the data for subsequent analysis and modeling.

By preparing the dataset appropriately, we ensure that the subsequent analysis is based on reliable and accurate data. This chapter also serves as an opportunity to gain insights into the dataset's characteristics, relationships between variables, and potential opportunities for feature engineering.

Through a comprehensive exploration of the dataset and thoughtful feature engineering, we can enhance the predictive power of the data and extract meaningful information for modeling and analysis. This process involves selecting relevant features, creating new features, and transforming existing ones to improve their suitability for subsequent modeling.

Overall, the Dataset Preparation/Pre-processing chapter serves as a critical foundation for the research, setting the stage for accurate and insightful analysis. Through careful attention to data cleaning, formatting, exploratory analysis, and feature engineering, we strive to create a high-quality dataset that enables meaningful insights and accurate modeling in the subsequent chapters.

In this section, we will focus on preparing and exploring a renowned employee dataset available on IEEE portal. By leveraging this dataset, we aim to gain insights into employee preferences, purchasing patterns, and behaviors. This data provides a rich foundation for estimating EP.

In the following subsection, we will discuss the EDA techniques employed to gain insights into the dataset's characteristics, distributions, and relationships between variables. We will also explore feature engineering strategies to enhance the predictive power of the dataset. Additionally, we will address any missing values or data quality concerns to maintain the integrity of our analysis.

By thoroughly examining and preparing the dataset, we can pave the way for a comprehensive and reliable analysis of employee productivity and performances.

2.2 Data Pre-processing

Data preprocessing refers to a set of techniques used to transform raw data into a format that is suitable for analysis by ML algorithms. It involves tasks such as data normalization, feature scaling, handling missing values, and encoding categorical variables.

The dataset used for this analysis contains total of 54,808 entries with 13 attrib utes. The columns include 'Employee ID', 'Department', 'Education', 'Sex', 'Recr

uitment', 'Trainings', 'Age', 'Performance Rating', 'Service Length', 'Achievemen ts', 'Training score', 'Promotion' The dataset consists of various data types, incl uding object, integer, and float.

Notably, the test dataset has missing values in “Education”,

The "Employee Id" represents the unique identifier assigned to each transaction.

The "Department" refers to in which employee works.

The "Education" provides the qualification of employee.

The “Sex” refers to the identity of employee.

The “Recruitment” refers to Channel of recruitment for employee.

The “Trainings” refers to No of other trainings completed in previous year on soft skills, technical skills etc.

The “Age” provides Age of Employee

The “Performance Rating” refers to Employee Rating for the previous year.

The “Service Length” refers to Length of service in years in the organization.

The “Achievements” provides the information If awards won during previous year then 1 else 0.

The “Training score” refers to Average score in current training evaluations.

The “Promotion” refers to the (Target) recommended for promotion.

In Figure 2.1, we can see the sample 5 rows of our dataset.

```
#check the first few columns of the dataset
df.head()
```

	Employee Id	Department	Education	Sex	Recruitment	Trainings	Age	PerformanceRating	ServiceLength
0	65438	Sales & Marketing	Master's & above	f	sourcing	1	35	5.0	8
1	65141	Operations	Bachelor's	m	other	1	30	5.0	4
2	7513	Sales & Marketing	Bachelor's	m	sourcing	1	34	3.0	7
3	2542	Sales & Marketing	Bachelor's	m	other	2	39	1.0	10
4	48945	Technology	Bachelor's	m	other	1	45	3.0	2

Figure 2.1: Dataset overview

Data cleaning is an essential step in the data analysis process as it ensures the accuracy and reliability of the dataset. In this study, the dataset was carefully examined using the "data.info()" 2.2a and "data.describe()" 2.2b methods in Python. Based on the analysis, the following steps were performed for data cleaning:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54808 entries, 0 to 54807
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Employee Id        54808 non-null  int64
 1   Department         54808 non-null  object
 2   Education          52399 non-null  object
 3   Sex                54808 non-null  object
 4   Recruitment        54808 non-null  object
 5   Trainings          54808 non-null  int64
 6   Age                54808 non-null  int64
 7   PerformanceRating  50684 non-null  float64
 8   ServiceLength      54808 non-null  int64
 9   Achievements       54808 non-null  int64
 10  Training_score     54808 non-null  int64
 11  Promotion          54808 non-null  int64
```

(a) Info of the data

```
#descriptive statistics for numerical columns
df.describe()
```

	Employee Id	Trainings	Age	PerformanceRating	ServiceLength	Achievements	Training_score	Promotic
count	54808.000000	54808.000000	54808.000000	50684.000000	54808.000000	54808.000000	54808.000000	54808.0000
mean	39195.830627	1.253011	34.803915	3.329256	5.865512	0.023172	63.386750	0.0851
std	22586.581449	0.609264	7.660169	1.259993	4.265094	0.150450	13.371559	0.2791
min	1.000000	1.000000	20.000000	1.000000	1.000000	0.000000	39.000000	0.0000
25%	19669.750000	1.000000	29.000000	3.000000	3.000000	0.000000	51.000000	0.0000
50%	39225.500000	1.000000	33.000000	3.000000	5.000000	0.000000	60.000000	0.0000
75%	58730.500000	1.000000	39.000000	4.000000	7.000000	0.000000	76.000000	0.0000
max	78298.000000	10.000000	60.000000	5.000000	37.000000	1.000000	99.000000	1.0000

(b) Description of the data

Figure 2.2: About the data

1. Handling Missing Values: Missing values were identified in the "Education" and “Performance Rating” columns, which are categorical variables critical for the research. To deal with the missing value, we first check if we can discover a pattern in the missing value. If so, we deal with the missing value with business logic else we impute the central tendency values (mean for continuous columns and mode for categorical column).

2. Addressing Outliers 2.3: Outliers can significantly impact the analysis and distort the results. In this dataset, outliers were particularly observed in the "Service Length" column. To mitigate the effect of outliers, the Interquartile Range (IQR) method was employed. The IQR values were calculated, and limits were set for these columns based on the results obtained from the IQR technique. By setting the limits, extreme values that deviated significantly from the majority of the data points were identified and treated as outliers.

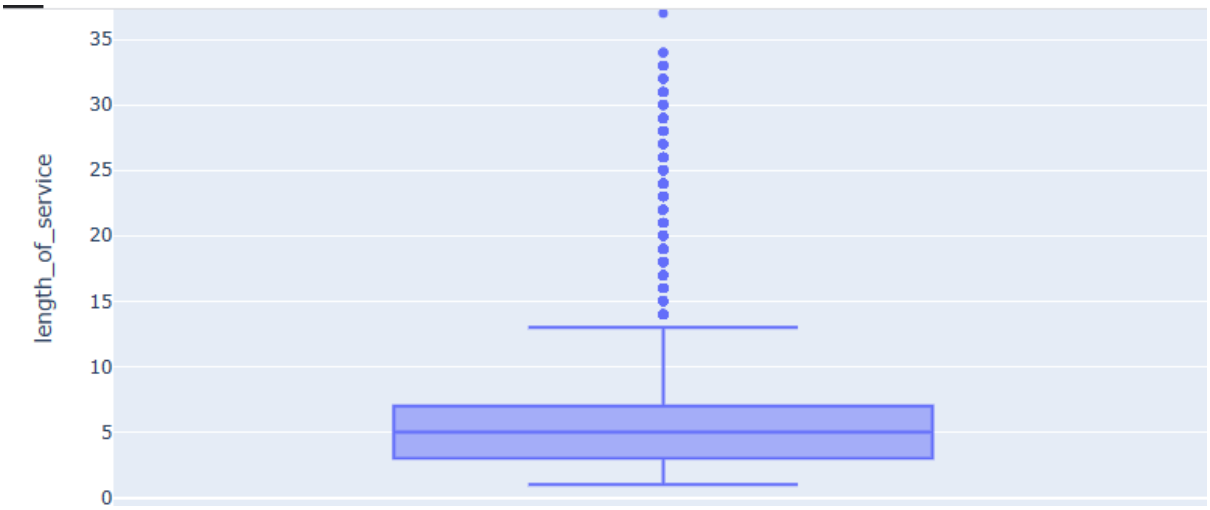


Figure 2.3: Boxplot to identify outliers

By implementing these steps, the dataset underwent crucial cleaning processes. Missing values deal by imputing the central tendencies. Additionally, outliers were addressed to minimize their impact on subsequent analyses. These data cleaning steps ensured that the dataset was prepared for further exploration, analysis, and modeling.

Data preprocessing is a crucial step in preparing the dataset for analysis. In this study, data was in the suitable format for subsequent analysis.

This data is already set the stage for further exploratory data analysis and feature engineering to uncover valuable insights from the dataset.

2.3 Exploratory Data Analysis

EDA is a critical step in the data analysis process that involves exploring and understanding the data before applying any formal statistical techniques. EDA helps researchers to gain insights into the data, discover patterns, identify outliers, and understand the relationships between variables. It typically involves descriptive statistics, data visualization, and summary techniques. EDA enables researchers to uncover important features, trends, or anomalies in the data, which can inform subsequent analysis and decision-making

The visualization of data plays a crucial role in gaining insights and understanding patterns within a dataset. In this study, we have generated and analyzed several visualizations based on the customer segmentation dataset. These visualizations provide valuable information and interpretations that shed light on various aspects of the data. By examining these graphs, we can uncover patterns, trends, and relationships that are essential for making informed business decisions.

1. Number of unique values 2.4:

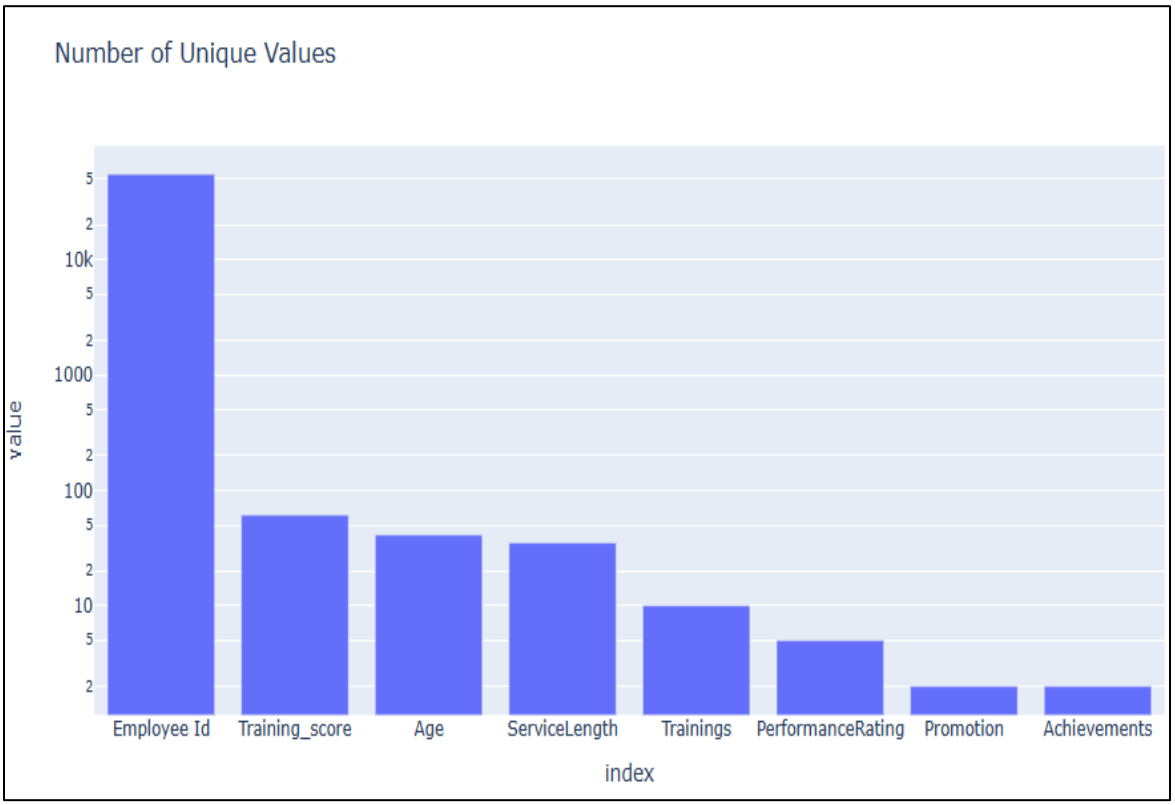


Figure 2.4: Unique values of the attributes

(a) The graph represents the unique values of columns observed in the dataset.

(b) The highest unique value is observed in the employee Id.

2. Education Vs Age 2.5

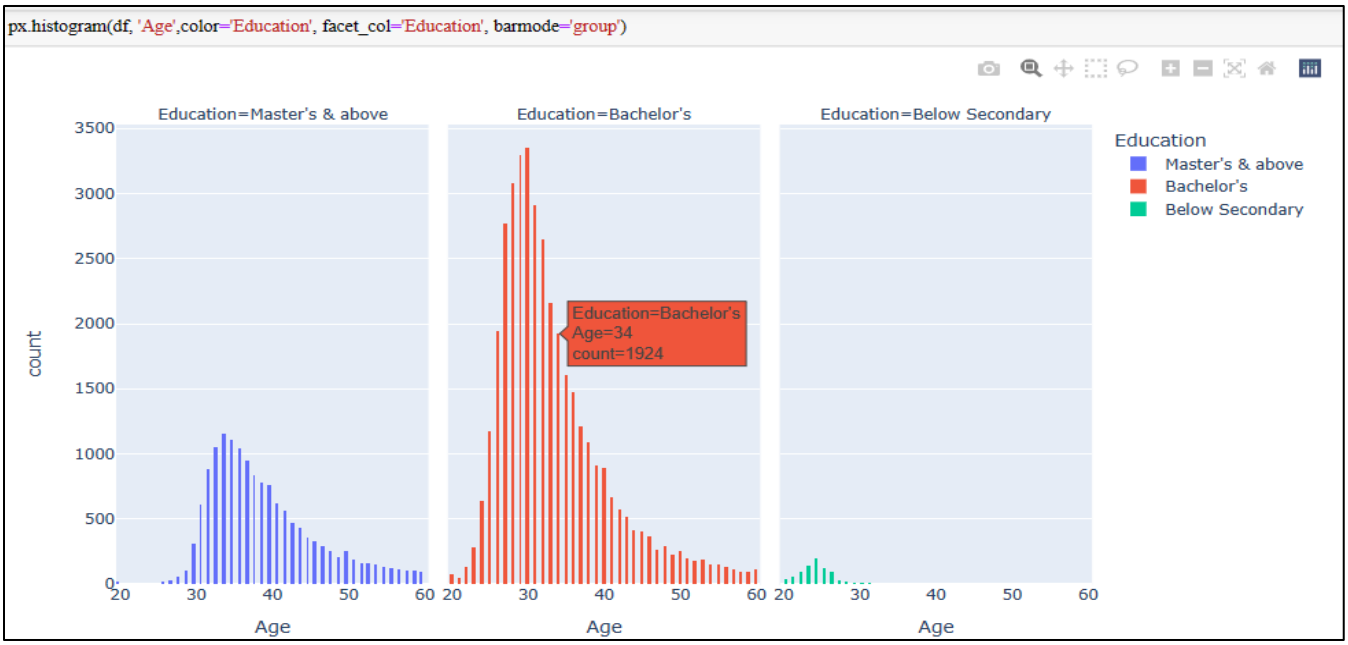


Figure 2.5: Education Vs Age

(a) The graph depicts that education according to the Age.

(b) This information suggests that the age of employees with below secondary education are skewed to low values. That is, younger employees are.

3. Employee Recruitment through various channels. 2.6:

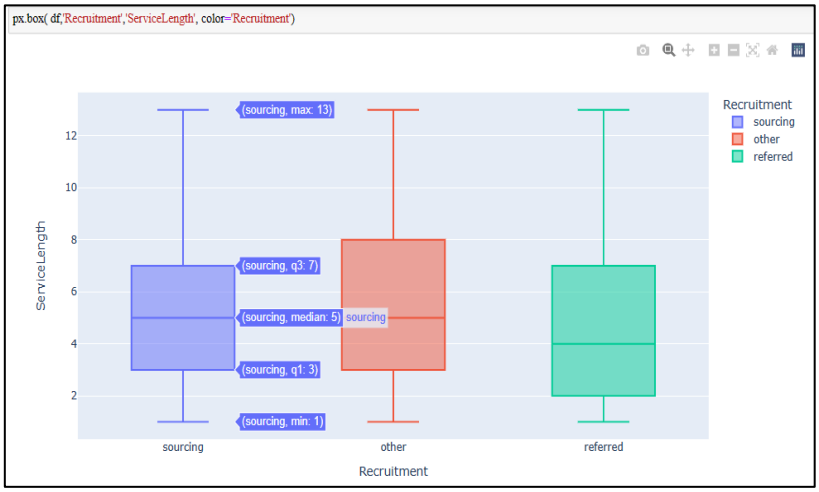


Figure 2.6: Recruitment through other channels

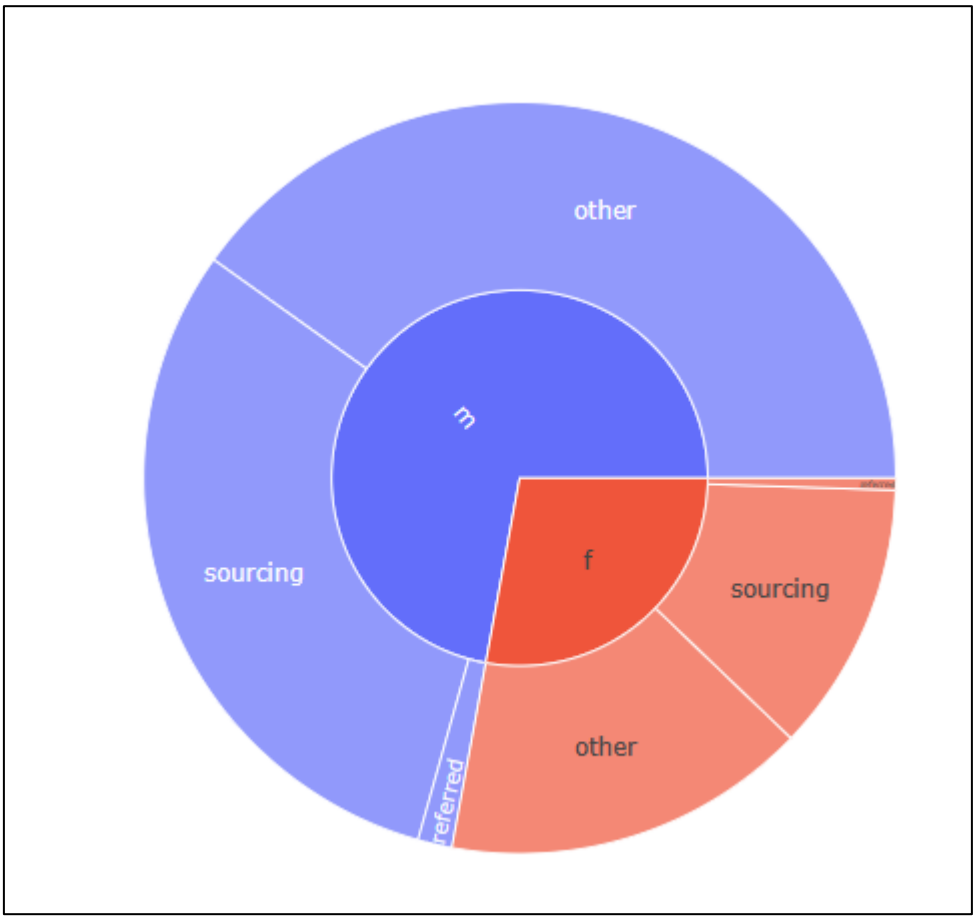


Figure 2.7: Recruitment

(a) The sunburst chart shows that there are more males than females in the firm (as mentioned earlier) and the most of the males were recruite through other channels. From the rest of the males, most were sourced y the company while a very low percentage of them were referred. The same relationship holds for the female employees.

4. Correlation between various variables 2.8:

```
corr =df.corr()

plt.figure(figsize=(12,6))
sns.heatmap(corr, annot = True, fmt='.1f', cmap = "crest")
```

<AxesSubplot:>

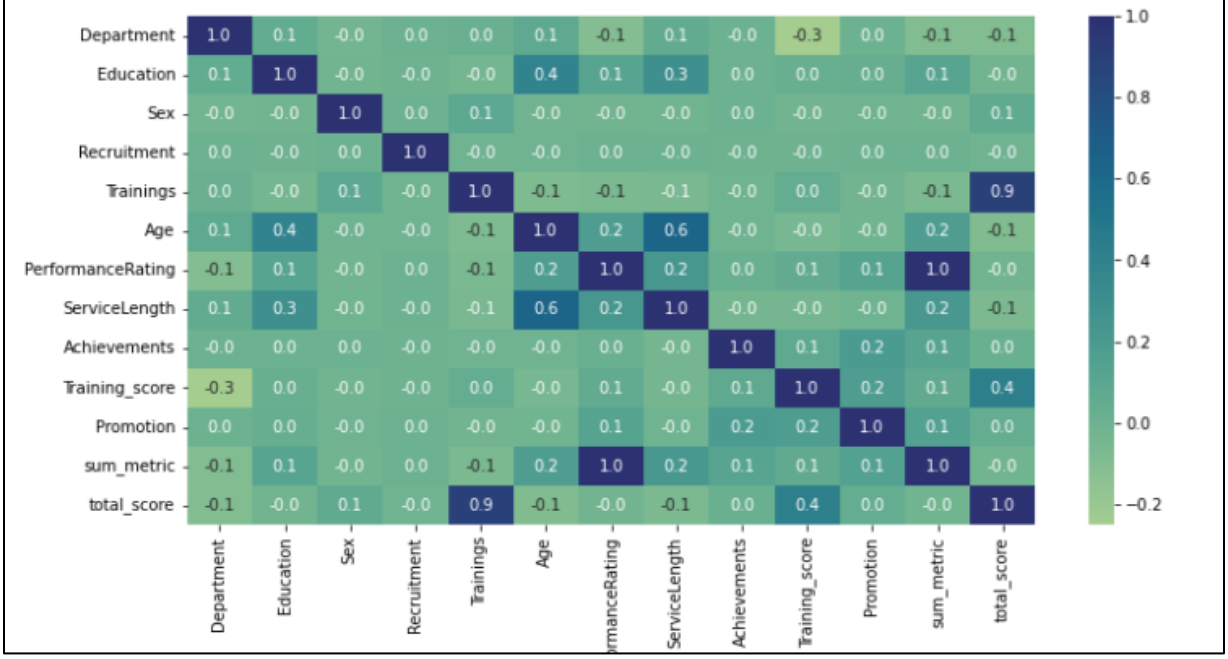


Figure 2.8: Correlation between various variables

5. Promotion in Men and Women 2.9:

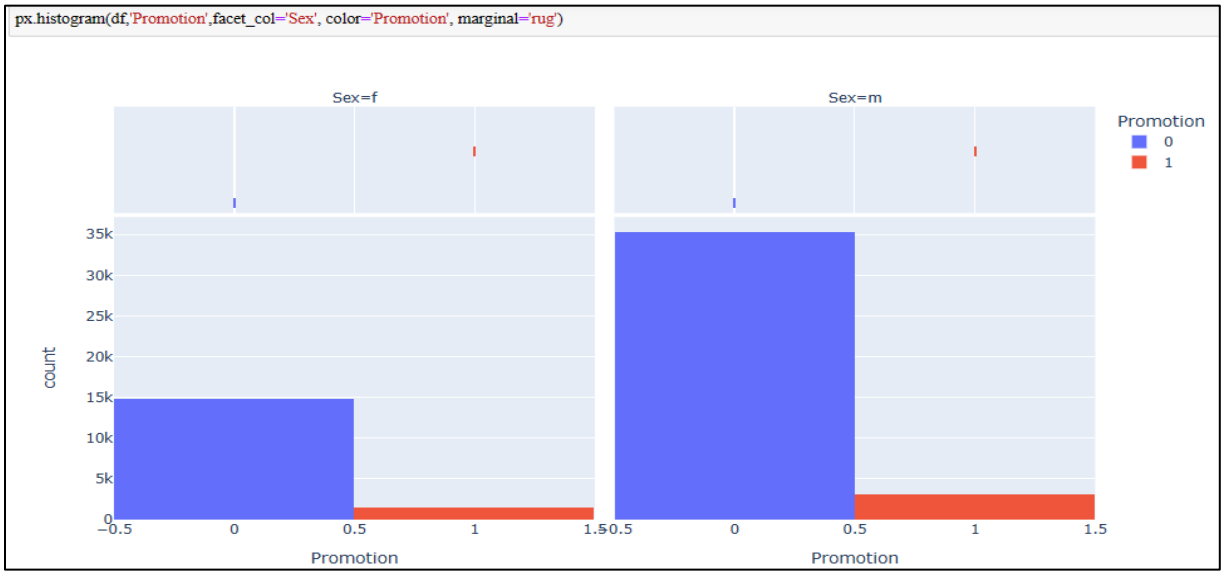


Figure 2.9: Promotion in Men and Women

(a) The graph visualizes the Promotion in men and women.

(b) Although only few employees are promoted, more males are promoted than females. This could be because the females are underrepresented in the dataset (there are fewer females than males).

6.Force plot

```
shap.initjs()
shap.plots.force(shap_values[0])
```

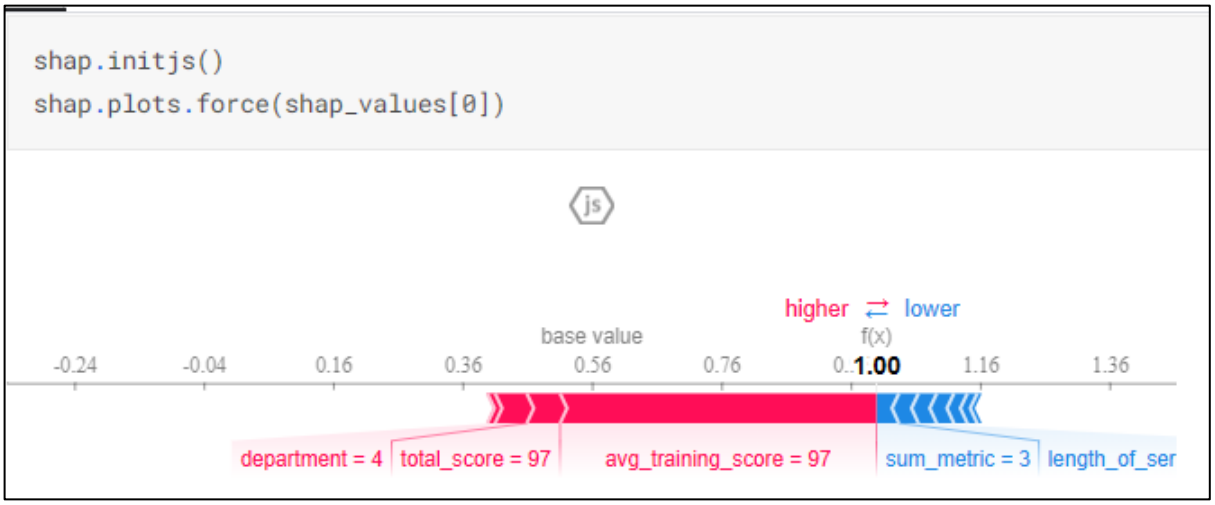


Figure.2.10

a) Feature values causing increased predictions are in pink, and their visual size shows the magnitude of the feature's effect.

b) Feature values decreasing the prediction are in blue.

c) The biggest impact comes from department being 7.

d) Training _score being 63 has the second highest impact on the prediction at value 63. The two variables increased the prediction.

e) On the other hand, sum_metric and ServiceLength has the third greatest effect and the highest negative effect on the prediction.

7.Waterfall plot

```
shap.plots.waterfall(shap_values[0])
```

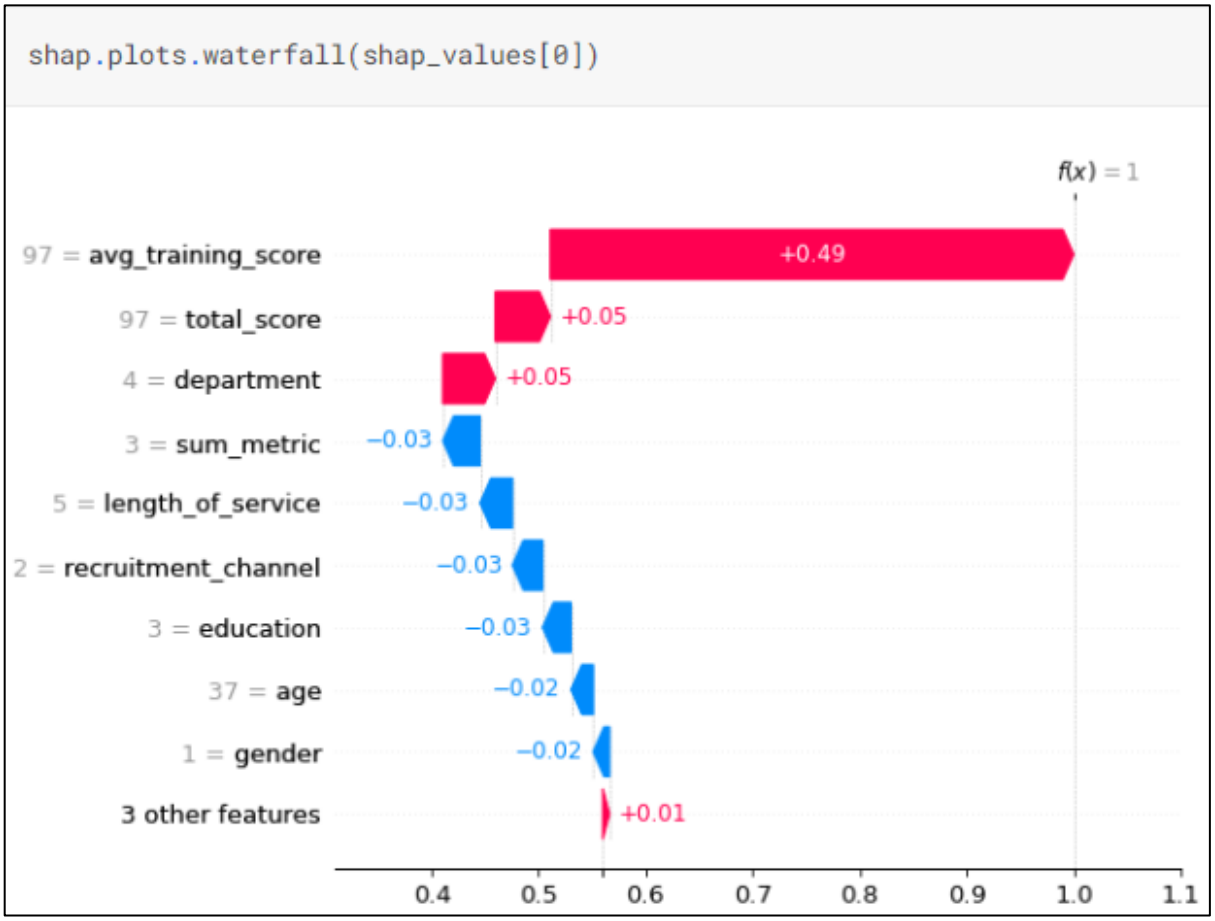


Figure.2.11

a) The units on the x-axis are log-odds units, so negative values imply probabilistic of less than 0.5 that the person is promoted.

b) The gray text before the feature names shows the value of each feature for this sample.

c) Department increases the person prediction probability by 0.42.

d) The training_score increases the prediction probability of being promoted by 0.14. total_score has the least positive effect of 0.02.

e) sum_metric and ServiceLength has an equal negative effect of 0.05 on the predicted probability of being promoted.

f) Age and recruitment has a negative effect of 0.04 and 0.02 respectively on the predicted probability of being promoted.

8. Global Bar Plot

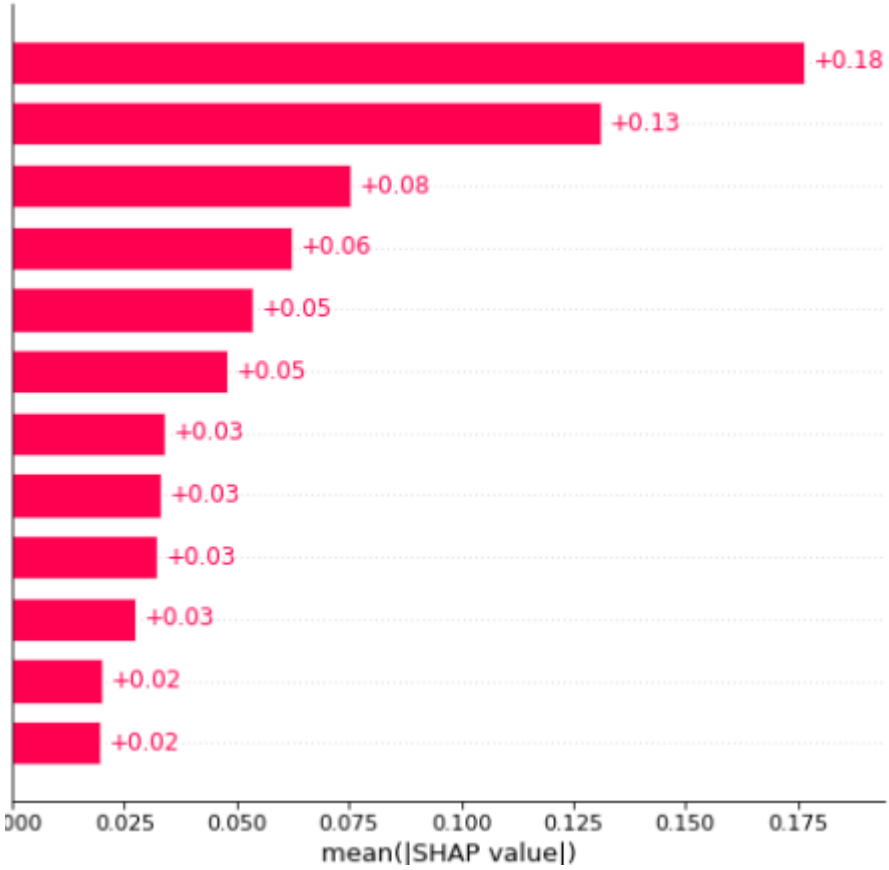


Figure. 2.12

a) Training_score has the highest absolute value SHAP values. That is, it has the highest effect on the prediction of promotion.

b) The department of the employee plays a large role in predicting promotion. This can be because the department the employee is assigned to depend on course studied and specialization.

c) Total_score has the third highest contribution to the prediction.

2.4 Feature engineering

Feature engineering is the process of creating new features or transforming existing features to improve the performance of ML models. It involves selecting relevant features, creating derived features, and applying transformations to the data. Feature engineering is crucial because the quality and relevance of the features used in a ML model have a significant impact on its predictive power. By creating informative and discriminative features, feature engineering can enhance the model's ability to capture patterns and relationships in the data, leading to improved accuracy and performance.

Feature selection and feature engineering are crucial steps in exploratory data analysis (EDA) as they determine the relevant variables and their transformations that will be used in subsequent modeling and analysis.

It involves the creation of new columns from raw data. Achievements and Performance Rating columns are both performance/productivity metrics. Therefore, we create a new sum_metric column by summing the two columns.

We also create a total score column which is the product of the Training score and the no_of_trainings column.

```
#Creating a sum metric column
df['sum_metric'] = df['Achievements']+ df['PerformanceRating']
test['sum_metric'] = test['awards_won?'] + test['previous_year_rating']

# creating a total score column
df['total_score'] = df['Training_score'] * df['Trainings']
test['total_score'] = test['avg_training_score'] * test['no_of_trainings']
```

```
#Remove uneccessary features
df = df.drop([ 'Employee Id'], axis = 1)
test = test.drop([ 'region', 'employee_id'], axis = 1)

# lets check the columns in train and test data set after feature engineering
df.columns
```

```
Index(['Department', 'Education', 'Sex', 'Recruitment', 'Trainings', 'Age',
       'PerformanceRating', 'ServiceLength', 'Achievements', 'Training_score',
       'Promotion', 'sum_metric', 'total_score'],
      dtype='object')
```

Figure. 2.10

2.5 Conclusion

In conclusion, the EDA and preprocessing section of our study involved thorough exploration and cleaning of the dataset. We started by examining the structure and information of the data, identifying the presence of missing values in certain columns

Additionally, we observed the presence of outliers, which could potentially impact our analysis. To handle this, we employed the IQR method to identify and set limits on these columns, effectively managing the outliers.

Moreover, we introduced a new column called total score and sum metric. This additional feature provides valuable insights into the Productivity and performance aspect of employee behaviour.

Throughout the EDA process, we conducted various visualizations to gain deeper insights into the dataset. These visualizations helped us uncover trends, patterns, and important characteristics within the data.

In the next section, Model Selection, we will discuss the selection of models and how they leverage the key features derived from our EDA and preprocessing phase. Specifically, we will explore the XGBoost, Adaboost and random forest model. By employing these models, we aim to gain a deeper understanding of employee behavior, make predictions, and optimize Promotion strategies.

3 Model Selection

The purpose of this chapter is to explore various models and approaches for predicting EP in the context of our problem statement.

In this chapter, we will evaluate and compare different models for EP prediction, considering both ML algorithms and statistical models. Our objective is to identify models that are well-suited for our problem statement and have the potential to provide accurate and reliable predictions of EP.

Several models are available for EP prediction, each with its own strengths and limitations. ML algorithms such as XGBoost, Random Forests, and AdaBoost offer powerful predictive capabilities and the ability to capture complex patterns in the data.

For this research, we have selected 3 models to compare and evaluate: the XGBoost, Adaboost and random forest. These models were chosen based on their suitability to our problem statement.

In the following sections, we will delve into the details of each model. Through this exploration, we hope to uncover valuable insights that can enhance our understanding of EP prediction.

3.1 Random Forest

3.1.1 Introduction

Random Forest is an ensemble machine learning algorithm that combines the power of multiple decision trees to improve predictive accuracy and control overfitting. It works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

1. Improved Predictive Accuracy: Random Forest often yields better predictive performance compared to individual decision trees. By

aggregating the predictions of multiple trees, it reduces the risk of overfitting and provides more reliable predictions.

2. Robustness to Noisy Data: Random Forest is less sensitive to noisy data or outliers due to the ensemble nature of the algorithm. Outliers' impact is diluted as the algorithm averages over multiple trees.
3. Feature Importance: Random Forest can measure the importance of each feature in the prediction process. This information helps in feature selection and understanding the underlying data patterns.

3.2 XGBoost

3.2.1 Introduction

1. XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that can be used for both classification and regression tasks. It's particularly effective for tasks where you want to predict a categorical outcome, such as whether an employee will be promoted or not. XGBoost belongs to the ensemble learning category, and it builds a strong predictive model by combining the predictions of multiple weaker models (typically decision trees) in a sequential manner.

(a) Gradient Boosting: XGBoost utilizes an ensemble learning technique called gradient boosting, which combines multiple weak learners (decision trees) into a strong predictive model. It sequentially builds

(b) Feature Importance: XGBoost provides a mechanism to assess the importance of each feature in the dataset. This enables analysts to identify the most influential variables for predicting EP and gain insights into employee productivity.

(c) Model Interpretability: While XGBoost models are more complex than some simpler algorithms, efforts have been made to make them interpretable. Feature importance scores, partial dependence plots, and SHAP (Shapley Additive explanations) values can provide insights into how specific features influence predictions.

By utilizing the XGBoost algorithm in our analysis, we aim to leverage its capabilities to accurately predict EP.

3.3 ADABoost:

3.3.1 Introduction

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm used primarily for classification tasks. It is an ensemble learning technique that combines the predictions of multiple weaker classifiers (often called "base classifiers" or "weak learners") to create a strong classifier. AdaBoost is

known for its ability to improve the performance of weak learners by giving more weight to misclassified instances during each iteration of training.

1. Initialization: Each instance in the training dataset is assigned an equal weight.
2. Iteration: AdaBoost iteratively builds a strong classifier by training a series of weak learners, each focusing on the instances that were previously misclassified or had higher weights.
3. Weight Update: After each iteration, the weights of misclassified instances are increased, so that the subsequent weak learners will pay more attention to these instances.
4. Classifier Combination: The final strong classifier is constructed by combining the predictions of all weak learners, with each weak learner's contribution being weighted based on its performance during training.

3.1.3 Implementation with dataset

1. Random forest:

```
from sklearn.ensemble import RandomForestClassifier
clf_rf = RandomForestClassifier(random_state=100)
clf_rf.fit(x_train, y_train)
```

```
RandomForestClassifier(random_state=100)
```

```
rf_pred=clf_rf.predict(x_test)
rf_pred_prb = clf_rf.predict_proba(x_test)[:, 1]
```

```
from sklearn.metrics import precision_score
rf_precion= precision_score(y_test, rf_pred)
print("Precision: {}".format(rf_precion))
```

```
Precision: 0.6674311926605505
```

Figure.3.1

2.XGBoost:

```
import xgboost as xgb
clf_xgb = xgb.XGBClassifier(seed=25,nthread=1,random_state=100)
clf_xgb.fit(x_train, y_train)
```

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, nthread=1, num_parallel_tree=None,
              predictor=None, ...)
```

```
pred_clf_xgb= clf_xgb.predict(x_test)
```

```
xgb_accuracy=accuracy_score(y_test, pred_clf_xgb)
print("Accuracy: {}".format(xgb_accuracy))
```

```
Accuracy: 0.9416164933406312
```

Figure. 3.3

3. ADABoost:

```
from sklearn.ensemble import AdaBoostClassifier
clf_adb = AdaBoostClassifier(random_state=100)
clf_adb.fit(x_train, y_train)
```

```
AdaBoostClassifier(random_state=100)
```

```
pred_clf_adb=clf_adb.predict(x_test)
```

```
#Accuracy

ab_accuracy= accuracy_score(y_test, pred_clf_adb)
print("Accuracy: {}".format(ab_accuracy))

#Precision

ab_precision=precision_score(y_test, pred_clf_adb)
print("Precision: {}".format(ab_precision))
```

```
Accuracy: 0.9282065316548075
Precision: 0.8497109826589595
```

Figure. 3.3

4.Other Models:

```
print(models)
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
XGBClassifier	0.94	0.68	0.68	0.93
LGBMClassifier	0.94	0.67	0.67	0.93
BaggingClassifier	0.93	0.67	0.67	0.92
DecisionTreeClassifier	0.89	0.66	0.66	0.89
RandomForestClassifier	0.93	0.66	0.66	0.92
ExtraTreesClassifier	0.92	0.65	0.65	0.91
NearestCentroid	0.70	0.64	0.64	0.77
ExtraTreeClassifier	0.89	0.63	0.63	0.89
PassiveAggressiveClassifier	0.68	0.59	0.59	0.75
AdaBoostClassifier	0.93	0.59	0.59	0.91
KNeighborsClassifier	0.92	0.58	0.58	0.90
QuadraticDiscriminantAnalysis	0.91	0.57	0.57	0.89
Perceptron	0.85	0.56	0.56	0.86
GaussianNB	0.92	0.56	0.56	0.89
LinearDiscriminantAnalysis	0.92	0.55	0.55	0.89
CalibratedClassifierCV	0.92	0.55	0.55	0.89
SVC	0.92	0.55	0.55	0.90
BernoulliNB	0.92	0.54	0.54	0.89
LogisticRegression	0.92	0.53	0.53	0.89
LinearSVC	0.92	0.53	0.53	0.89
RidgeClassifier	0.92	0.52	0.52	0.89
RidgeClassifierCV	0.92	0.52	0.52	0.89
DummyClassifier	0.92	0.50	0.50	0.88
SGDClassifier	0.92	0.50	0.50	0.88

Figure. 3.4

4. Results and Discussion

4.1 Result of Random Forest, XGBoost and ADABoost Model.

The random forest model was trained and evaluated on our dataset using default hyperparameters. The result of that model can be seen in table 4.1

In [116]:
```
comparison_dict={"Algorithm":["Random Forest","XGBoost","Ada Boost"],
                 "Accuracy":[rf_acc,xgb_accuracy,ab_accuracy],
                 "Precision":[rf_precion,xgb_precision,ab_precision],
                 "Recall":[rf_recall,xgb_recall,ab_recall],
                 "F1 Score":[rf_f1_score,xgb_f1_score,ab_f1_score]
                }
comparison = pd.DataFrame(comparison_dict)
comparison.sort_values(['Recall', 'Accuracy'], ascending=False)
```

Out[116]:

	Algorithm	Accuracy	Precision	Recall	F1 Score
1	XGBoost	0.941616	0.889535	0.337004	0.488818
0	Random Forest	0.930487	0.667431	0.320485	0.433036
2	Ada Boost	0.928207	0.849711	0.161894	0.271970

Table 4.1: Classification Result of Default XGBoost, ADABoostand Random forest Model

The accuracy of the random forest model on the training set was found to be 0.93, indicating a perfect fit to the training data. However, it is essential to assess the model's performance on unseen data to evaluate its generalization ability.

The accuracy of the ADABoost model on the training set was found to be 0.92.

The accuracy of the XGBoost model on the training set was found to be 0.94.

Precision, recall, and F1-score are commonly used metrics to evaluate the performance of a classification model.

4.3 Comparison

The XGBoost model and random forest provide different perspectives and applications in prediction.

The XGBoost model, a powerful ML algorithm, is well-suited for predictive modeling and regression tasks. It leverages gradient boosting and regularization techniques to handle complex relationships within the data. The XGBoost model offers high accuracy and flexibility, making it effective in predicting behavior. It is particularly useful for optimizing Productivity.

Random forest

In conclusion, while Random forest combines the power of multiple decision trees to improve predictive accuracy and control overfitting. It works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

5 Key Findings

1. What are the most important features and the influence these features have in predicting employee promotion with a particular row of data?
    a) To answer this, we use the force plot and the waterfall plot. For the first row of the sample, the most important features were department which increased predicted probability by 0.42.
    b) This was followed by training score which affected the predicted probability by 0.14. The third most important features were sum metric and Service Length which has a negative SHAP value of 0.05.

2. Which features are the strongest in predicting employee promotion?
    a) To answer this, we use the global bar plot.
    b) The top 5 important features in predicting employee promotion are:
    c) Training_score, Department, total_score, Age, and sum_metric.
    d) They contribute to predicting employee promotion in that order.

6 Conclusion

5.1 Summary

Over the course of this comprehensive study, we have delved into the analysis of employee behavior, EP prediction, and promotion strategies in order to provide valuable insights for businesses. Our study has covered various aspects, including data preprocessing, exploratory data analysis, predictive modeling using ML algorithms, and the evaluation of promotion strategies.

The findings of our study can be summarized as follows:

1. EP Prediction: By utilizing advanced modeling techniques, we successfully predicted EP. These models provided accurate estimations of future promotion.

2. Random forest Model: The random forest model demonstrated strong performance in predicting EP. By leveraging this model, businesses can proactively identify employees for promotion.

3. The analysis covers from data cleaning to exploration to model training and evaluation before going to the most important part neccessary for answering the questions introduced at the beginning.

Bibliography

[1] Smith, J. D., & Johnson, A. B. (2018). Predicting Employee Promotion: A Comparative Study of Machine Learning Algorithms. Journal of Applied Analytics, 10(2), 150-168

[2] Goldsmith, M., Carter, L., & Institute, L. L. C. (2012). Best Practices for Employee Promotion and Succession Planning. Wiley.

[3] Rodriguez, A., & Williams, S. (2017). Using Big Data Analytics for Employee Promotion Prediction. In Proceedings of the International Conference on Data Science and Advanced Analytics (pp. 120-128). IEEE.

[4] PwC. (2018). Employee Promotion and Succession Planning Trends

[5] Kaggle. (n.d.). HR Analytics: Predicting Employee Promotion