INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Applying Machine Learning Models to Predict Voluntary Employee Attrition in a Portuguese IT Organisation

Sofia de Almeida Correia

Master in Management

Supervisors:
PhD, Rui Francisco Pereira Moital Loureiro da Cruz, Invited Associate Professor,
Iscte – Instituto Universitário de Lisboa

PhD, Daniel Monroe Corry, Invited Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2025

BUSINESS
SCHOOL

Department of Marketing, Operations and General Management

Applying Machine Learning Models to Predict Voluntary Employee Attrition in a Portuguese IT Organisation

Sofia de Almeida Correia

Master in Management

Supervisors:
PhD, Rui Francisco Pereira Moital Loureiro da Cruz, Invited Associate Professor,
Iscte – Instituto Universitário de Lisboa

PhD, Daniel Monroe Corry, Invited Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2025

Acknowledgments

This is the moment to express my heartfelt gratitude to all those who made this possible, not only the completion of this thesis but my entire academic journey leading up to it.

I am sincerely grateful to my supervisors, Professor Francisco Cruz and Professor Daniel Corry, for their continuous guidance, constructive feedback, and encouragement throughout this research. Their combined expertise, Professor Cruz in the technical domain of Machine Learning (ML), and Professor Corry in the field of Human Resource Management (HRM), was essential in shaping both the direction and quality of this work, ensuring that it successfully bridges the gap between ML and HRM in a meaningful and coherent way.

I would also like to express my gratitude to the organisation for providing access to the anonymized human resources (HR) data that made this research possible, and in particular to the HR team for their availability, collaboration, and valuable contributions during the course of this study.

I am grateful to the faculty and staff at ISCTE – Instituto Universitário de Lisboa for their academic support and resources. Over the past seven years, I have gained invaluable knowledge, critical thinking skills, and professionalism that have shaped my academic, professional and personal development.

Finally, I wish to express my deepest gratitude to my family for their endless patience, understanding, and love — their support has been my foundation every step of the way. To my friends, for your patience, encouragement, and respect for my absence during this journey. And to my partner, for believing in me unconditionally and motivating me to persevere until the very end.

Obrigada!

Resumo

Esta tese explora a aplicação de um modelo preditivo de *machine learning* (ML) para prever a saída voluntária de colaboradores numa empresa portuguesa de tecnologia da informação (TI). Num contexto de elevadas taxas de rotatividade no setor tecnológico e dos custos significativos associados às saídas voluntárias, o estudo propõe o desenvolvimento de uma ferramenta de apoio à decisão baseada em dados, destinada a otimizar a gestão de recursos humanos.

Seguindo a metodologia CRISP-DM, a investigação recorre a dados reais anonimizados de recursos humanos (HR) de uma organização para conceber, treinar e avaliar modelos de classificação, destacando-se o LightGBM como o preditor mais preciso da rotatividade. Os resultados revelam que fatores como progressão na carreira (intervalo entre promoções funcionais), compensação variável, unidade organizacional, nível de lealdade e aumento salarial são os mais influentes na previsão de saídas voluntárias de colaboradores. O modelo preditivo demonstra um desempenho robusto (F1 = 0.578; AUC = 0.92), oferecendo *insights* acionáveis para estratégias de retenção direcionadas e para o planeamento da força de trabalho baseado em evidências.

Adicionalmente, o estudo discute as implicações mais amplas da análise preditiva na gestão de RH, salientando a importância da monitorização contínua e da atualização do modelo para garantir a sua fiabilidade e relevância, ao longo do tempo. Esta investigação contribui para a literatura académica ao evidenciar a eficácia da análise de RH baseada em ML e apresenta recomendações práticas para organizações que pretendam abordar a rotatividade de colaboradores através de métodos analíticos avançados.

Palavras-chave: Rotatividade de Colaboradores; Aprendizagem Automática; Análise Preditiva; Gestão de Recursos Humanos; CRISP-DM; LightGBM; Retenção de Colaboradores.

JEL Classification System: C55 Large Data Sets: Modeling and Analysis; J63 Turnover, Vacancies, Layoffs; M12 Personnel Management; M15 IT Management; O15 Human Resources and Development.

Abstract

This thesis explores the application of a machine learning (ML) predictive model to forecast voluntary employee attrition in a Portuguese information technology (IT) company. The study proposes the development of a data-driven decision-support tool aimed at optimizing human resource management, in the context of high turnover rates in the technology sector and the significant costs associated with voluntary departures.

Following the CRISP-DM methodology, the research makes use of real anonymized human resources (HR) data from an organization to design, train, and evaluate classification models, with LightGBM emerging as the most accurate predictor of attrition. The findings reveal that factors such as career progression (functional promotion interval), variable compensation, organisational unit, loyalty level, and salary increase are the most influential in predicting employee departures. The predictive model demonstrates robust performance (F1 = 0.578; AUC = 0.92), providing actionable insights for the development of targeted retention strategies and evidence-based workforce planning.

Additionally, the study discusses the wider implications of predictive analytics for HR management, emphasizing the importance of continuously monitoring and updating the model to ensure its reliability and relevance over time. This research contributes to academic literature by illustrating the effectiveness of ML-based HR analytics and offers practical recommendations for organisations seeking to address employee turnover through advanced analytical methods.

Keywords: Employee Attrition; Machine Learning; Predictive Analytics; Human Resource Management; CRISP-DM; LightGBM; Employee Retention.

JEL Classification System: C55 Large Data Sets: Modeling and Analysis; J63 Turnover, Vacancies, Layoffs; M12 Personnel Management; M15 IT Management; O15 Human Resources and Development.

Index

Chapter I - Introduction 1

Chapter II - Literature Review 3

2.1. The Technology Industry and Employee Turnover 3

2.1.1. Overview of Turnover in the Information Technology Sector 3

2.1.2. Turnover Metrics in *Tech* 7

2.2. Employee Retention and Turnover 8

2.2.1. The Importance of Employee Retention 8

2.2.2. Factors Influencing Employee Turnover 9

2.3. Employee Engagement 10

2.3.1. Role of Engagement in Retention 10

2.3.2. Measuring Employee Engagement 11

2.4. Human Resource Management through Analytics 12

2.4.1. Predictive Modeling in Human Resources 12

2.4.2. Strategic HR Management through Analytics 14

2.4.3. Ethical Considerations in HR Analytics 14

Chapter III - Approach and Methodology 16

3.1. Business Understanding 17

3.2. Data Understanding 19

3.2.1. Dataset Description 19

3.2.2. Exploratory Data Analysis 21

3.3. Data Preparation 33

3.3.1. Feature Selection 34

3.3.2. Feature Engineering 34

3.3.3. Data Cleaning 35

3.3.4. Transformation of Non-Numerical features 35

3.4. Modeling 37

3.4.1. Dataset Partitioning 38
3.4.2. Dataset Balancing 39
3.4.3. Model Selection 40
3.5. Evaluation 42
3.6. Deployment and Model Management 44
Chapter IV - Results Presentation and Discussion 47
4.1. Comparative Results Across Models 47
4.2. Evaluation of the Selected Model's Performance 49
4.3. Feature Importance Analysis of the Selected Model 51
4.4. Predicted Attrition Analysis 52
4.4.1. Bivariate Analysis 53
4.4.2. Employee Attrition Risk Profiles 55
4.5. Study Limitations 58
Chapter V - Conclusion and Recommendations 60
References 63
Annexes 68

Index of Figures

Figure 1 – Industries with above-average turnover rates (LinkedIn, 2022) 7
Figure 2 – CRISP-DM Stages 16
Figure 3 – Correlation Heatmap for the Numerical Features 33
Figure 4 - Class Distribution of the Target Variable (Attrition, 2022–2024) 39
Figure 5 - Confusion Matrix of the LightGBM Classifier on the Test Dataset 50
Figure 6 - ROC curve of the LightGBM Classifier on the Test Dataset 51
Figure 7 - Top 20 Feature Importances Identified by the LightGBM Classifier Model 52
Figure 8 - Distribution of Predicted Attrition Rate (Test Set) 53

Index of Tables

Table 1 - Annual Attrition Rates (2022–2024) 18
Table 2 - Representation of a Confusion Matrix 44
Table 3 - Comparative F1 scores by Classification Model (Validation vs. Train) 47

Index of Equations

Equation 2.1 – Attrition Rate 4
Equation 3.1 – Pearson Correlation Coefficient 31
Equation 3.2 – F1 Score Formula 42
Equation 3.3 – Precision Formula 42
Equation 3.4 – Recall Formula 43
Equation 3.5 – True Positive Rate (TPR) 43
Equation 3.6 – False Positive Rate (FPR) 43
Equation 3.7 – Area Under the Curve (AUC) ROC 43
Equation 3.8 – Specificity 44

Glossary of Acronyms

AI – Artificial Intelligence

AUC – Area Under the Curve

AUC-ROC – Area Under the Receiver Operating Characteristic Curve

CRISP-DM – Cross-Industry Standard Process for Data Mining

CV – Cross-Validation

F1 – F1 Score (harmonic mean of Precision and Recall)

FPR – False Positive Rate

HR – Human Resources

HRM – Human Resources Management

IT – Information Technology

ML – Machine Learning

ROC – Receiver Operating Characteristic

TPR – True Positive Rate

Chapter I - Introduction

In today's rapidly evolving technological landscape, human capital has emerged as a central driver of organisational competitiveness and sustainability. Nowhere is this more evident than in the information technology (IT) sector, where the demand for specialized talent is matched by a pronounced challenge: high employee turnover (Ang & Slaughter, 2004; Lewis & Soroñgon, 2022). For Portuguese technological companies, retaining skilled professionals is not only a matter of operational continuity but also a key strategic priority that directly influences productivity, client satisfaction, and long-term organisational resilience (Baron et al., 2001; Shaw, 2011). Within this context, the ability to anticipate and mitigate voluntary employee departures has become an imperative for effective human resources (HR) management.

This research project investigates the application of a machine learning (ML) predictive model in human resources management at a Portuguese information technology (IT) company. Given its direct application to a real-world organisation, the project is categorized as a practical, solution-oriented research endeavour. The project aims to address an important organisational problem—voluntary employee attrition—by taking advantage of predictive analytics to make informed data-driven decisions. As an applied, in-company research project, this study bridges academic research and practical organisational application, aiming to provide actionable insights to enhance employee retention strategies in the company.

Although the average attrition rate in the company across the three years under analysis was relatively moderate (approximately 8%), the costs associated with voluntary departures remain disproportionately high. These costs extend beyond direct recruitment and training expenses to include productivity losses, project interruptions, diminished team cohesion, and the erosion of organisational knowledge (Hom et al., 2019; Hancock et al., 2011). The central research problem is to determine how predictive models can be applied within this company to detect early indicators of potential voluntary attrition and to inform targeted strategies for strengthening employee retention. In other words, the research problem statement can be framed as: *How can machine learning predictive models be effectively used for analysing factors influencing employee attrition*?

Accordingly, the primary objective is to develop a predictive ML model that accurately forecasts voluntary employee attrition in this company. The specific objectives of this research are as follows:

i. To construct a ML classification model that predicts whether an employee will voluntarily leave the organisation, using the company's HR data. The model analyses key employee characteristics such as age, job position, department, compensation, job satisfaction, performance ratings, and work-life balance to make accurate predictions.
ii. To identify the most significant factors influencing employee attrition, by examining feature importance within the selected ML model. This analysis will reveal which attributes are most strongly correlated with an employee's likelihood of leaving.
iii. To evaluate the model's performance metrics, including f1 score, precision, recall, and AUC-ROC, ensuring the model's effectiveness and reliability in predicting attrition.
iv. To use the CRISP-DM methodology to ensure a structured approach in developing the predictive model, including the stages of business understanding, data understanding and preparation, modeling and evaluation.
v. To provide the company with a predictive model that serves as a practical decision-support tool for HR professionals, enabling both the identification of employees at high risk of attrition, and the recognition of the most influential factors driving employee exits. The model also facilitates the analysis of attrition distribution across key organisational variables, thereby offering actionable insights to guide targeted retention strategies and evidence-based workforce planning.
vi. To contribute to the academic debate on HR Analytics by demonstrating the effectiveness of predictive and descriptive analytics techniques in addressing employee turnover, specifically within the technology industry.

The present research project is structured into five chapters. Chapter one introduces the study by outlining the background, problem statement, objectives, and overall purpose. The second chapter provides a comprehensive literature review, discussing the theoretical foundations of employee turnover, retention, and engagement, as well as the application of predictive analytics and machine learning in human resource management. Subsequently, chapter three describes the methodological framework, specifically the CRISP-DM process used for developing the predictive model, including data understanding and preparation, modeling, and evaluation stages. Chapter four presents and interprets the results obtained from the different models tested, including comparative algorithm performance, final evaluation of the selected model, feature importance analysis, prediction results, and study limitations. Finally, chapter five summarizes the conclusions, discusses the practical and academic implications of the findings, and offers recommendations for future organisational practice, followed by the annexes containing supporting materials and additional analyses.

Chapter II - Literature Review

2.1. The Technology Industry and Employee Turnover

2.1.1. Overview of Turnover in the Information Technology Sector

2.1.1.1. Employee Turnover

One of the foundational studies on employee turnover centres around its historical significance, the associated costs, and the implications for organisational productivity. Hom, Allen, & Griffeth (2019) provide an overview of the subject:

> Employee turnover—or voluntary termination of employees from employing organizations (Mobley, 1982)—has attracted interest from employers and scholars alike for a century (Hom, Lee, Shaw, & Hausknecht, 2017). As early as 1917, management consultants and academicians have written about turnover's costs, its amelioration, and its causes (Hom et al., 2017). Managers have long focused on employee turnover because of the personnel costs incurred when employees quit, such as those for hiring and training replacements (Cascio, 2000). To illustrate, turnover costs can range from 90% to 200% of annual pay (Allen, Bryant, & Vardaman, 2010). Though difficult to compute for some occupational fields, turnover can impair productivity because short-staffed workforces may underproduce, while inexperienced replacements may produce inefficiently during early employment (Hom & Griffeth, 1995). (Hom, Allen, & Griffeth, 2019, p. 1).

Mobley (1982, as cited in Hom et al., 2019, p. 3) classically defined employee turnover as "an employee's voluntary severance of his or her current employment ties." Scholars have since expanded on this notion, focusing on self-initiated employee departures to study individual choice behaviour, as these reflect autonomous decisions (Campion, 1991, as cited in Hom et al., 2019). Voluntary turnover typically excludes internal movements or non-paid roles (Hom & Griffeth, 1995, as cited in Hom et al., 2019). In line with this definition, many researchers and practitioners conceptualize voluntary turnover as a dichotomy, differentiating *employee-initiated* resignations from *employer-initiated* separations (e.g., dismissals, layoffs). "Despite their simplicity, the types of leaving that constitute voluntary self-determined quits often vary

across employers and scholars, making determination of whether turnover is voluntary or not difficult." (Mobley et al., 1979, as cited in Hom et al., 2019, p. 4).

On the other hand, according to Susanto et al. (2023, as cited in Trisna Wijayanthi et al., 2024) employee turnover refers to the departure of employees from an organization, either voluntarily or involuntarily. Reasons for turnover can include job dissatisfaction, better career opportunities, or organisational decisions such as layoffs. High turnover rates often indicate underlying challenges within an organisation, such as a negative work environment, limited opportunities for career growth, or dissatisfaction with company policies.

To clearly distinguish between related concepts, Trisna Wijayanthi et al. (2024) provide the following categorization:

> The indicators or dimensions contained in the Employee Turnover variable include: 1) Voluntary turnover: Employees who leave the organization for numerous reasons, such as pursuing better career chances, job unhappiness, or personal reasons; 2) Involuntary turnover: Employees who are terminated by the company for performance reasons, restructuring, or staff cutbacks; 3) Internal turnover: Employees who change roles or departments inside the same organization; 4) Turnover rate: The percentage of employees who leave the organization in a certain time period, which can indicate workforce stability; and 5) Turnover expenses: The costs of replacing personnel, such as recruitment, training, and lost productivity (Widodo, 2023a). (Trisna Wijayanthi et al., 2024, p.291)

For the purposes of this study, voluntary attrition, which is conceptually aligned with voluntary turnover, is quantified using the following formula:

$$\text{Attrition Rate (\%)} = \frac{\text{Number of Voluntary Departures During a Period}}{\text{Average Number of Employees During the Same Period}} \times 100 \qquad (2.1)$$

In line with the literature (e.g., Hom, Allen, & Griffeth, 2019), this study adopts the term *attrition* to refer exclusively to voluntary employee departures, excluding involuntary exits such as dismissals or layoffs. This definition ensures conceptual alignment between the HR context and the predictive modeling framework developed in later chapters.

2.1.1.2. Why is turnover particularly high in the technology industry?

IT employees often shift jobs to seek challenging roles, innovative projects, and faster career progression, contributing to the industry's high turnover. "The apparent high employee turnover rates in the Information Technology industry suggests that the factors influencing turnover in the industry may not be well understood" (Abii et al., 2013, p. 740).

The results of a study conducted by Abii et al. (2013) indicate that employee compensation and workplace relationships are factors that influence turnover in the IT industry. Thozhur et al. (2006) observed that employees tend to be dissatisfied with their jobs if they do not receive adequate compensation. Additionally, Shields (2006) suggested that job satisfaction is connected to personal income (as cited in Abii et al., 2013).

Horn et al. (1992) emphasized that turnover intention would lead to actual turnover (as cited in Abii et al., 2013). Organisations do not necessarily have to pay employees high salaries to encourage them to remain with organisations (Abii et al., 2013). Other factors, such as good relations with supervisors, trust, and confidence in organisational leaders, are paramount to job satisfaction (Abii et al., 2013).

Anuradha (2005) stated that turnover intentions can indicate employees' plans to leave but do not reveal if those who remain are motivated to stay (as cited in Abii et al., 2013). Even if organisational leaders know why their employees leave, they cannot effectively retain employees until they understand the factors that encourage employees to remain (Abii et al., 2013).

Other studies have shown that while money and salary are important to most employees, remuneration is often not the primary reason employees choose to stay with organisations (Chandler, 2004; Muck, 2006, both as cited in Abii et al., 2013). Chandler (2004) identified several factors that are important to employees, including career growth, challenging and meaningful work, working with good people, recognition and appreciation for work well done, opportunities to contribute ideas on work processes, flexible work hours, and fair pay, good benefits, and job security. Although Chandler's study was conducted at an extension agency in Texas, this research focused on the IT industry in the DC metropolitan area (as cited in Abii et al., 2013).

A study by Lock (2003, as cited in Abii et al., 2013) found that IT workers are motivated by several workplace factors, including effective leadership, positive workplace relationships, challenging work, respect, and a balanced work-life. Additionally, DeMers (2002, as cited in Abii et al., 2013) suggested that IT professionals do not consider compensation the most

important factor when deciding to accept job offers. These findings suggest that organisational leaders should carefully examine the factors that motivate employees (as cited in Abii et al., 2013).

This leads us to the conclusion that turnover in the IT industry is intensified by competitive pressures for talent, stress linked to demanding job roles, and a trend towards frequent job changing driven by the need for career advancement and skill development.

2.1.1.3. Impact of Employee Turnover on the Performance of IT Companies

Employee turnover represents a significant loss for the high-*tech* business as the company must suffer from substantial indirect costs. "These costs include recruitment cost for finding replacements, expenses related to training and development of new employees, productivity loss on account of experienced employees leaving and morale loss to the organization" (Cascio & Boudreau, 2010; Matthew & Kung, 2007, both as cited in Khera & Divya, 2019, p. 1). Additionally, when new employees are hired, they undergo specialized training for a considerable amount of time that consumes extensive time and resources of the organisation. Meanwhile, businesses face low productivity as employees are unable to fully meet customer expectations until they thoroughly understand the entire process of the company (Bapna, Langer, Mehra, Gopal, & Gupta, 2013, as cited in Khera & Divya, 2019).

Conversely, the voluntary departure of employees who do not meet performance expectations may, in some cases, result in enhanced overall productivity for the company. Nonetheless, a drastic increase of the attrition rate can clearly indicate poor IT sector performance (Purohit, 2016; Shanmugam & Giri Babu, 2016, as cited in Khera & Divya, 2019). Furthermore, according to Kamalanabhan et al. (2009), "client satisfaction is also affected by employees, where trained and skilled employees can positively lead to satisfaction, but newly joined employees that lack experience may not be able to" (as cited in Khera & Divya, 2019, p. 13). This indicates that high employee turnover can directly impact business productivity, as client requirements may not be adequately met.

In conclusion, research suggests that:

> High employee turnover is a problem that organizations cannot ignore because of the financial burden, negative impact on employee morale, and employee relationships, and adverse impact on the quality of services and products that organizations deliver to customers. (Abii et al., 2013, p. 740).

2.1.2. Turnover Metrics in *Tech*

The technology industry experiences higher turnover rates compared to other sectors, with various factors contributing to this trend. IT professionals, particularly those in technical roles, are more likely to change jobs for better opportunities (Ang & Slaughter, 2004). "In fact, the tech sector has the highest turnover rate at 13.2% out of every single business sector, according to a recent turnover report from LinkedIn." (Johnson, 2018).

According to Johnson (2018), companies in the information technology (IT) industry are facing substantial costs due to high turnover, as it takes 51 days on average to fill technical vacancies. This results in considerable expenses not only for recruiting specialized talent but also for replacing that same talent within a short timeframe, as mentioned by Johnson (2018), “when they leave in one to two years”. Such high turnover can impact project timelines and financial performance, especially when key employees leave during critical phases like product launches (Johnson, 2018).

A LinkedIn study conducted by Lewis & Soroñgon (2022) revealed that, following the professional services industry, which holds the highest turnover rate (13.4%), the technology and media sector experiences the second-highest turnover rate at 12.9% (Figure 1). “Workers with *tech* skills remain in high demand, and employees in engineering roles have an above-average turnover rate.” (Lewis & Soroñgon, 2022). In this study, the authors further suggest that “improving compensation and expanding work flexibility may be effective strategies for reducing turnover.”

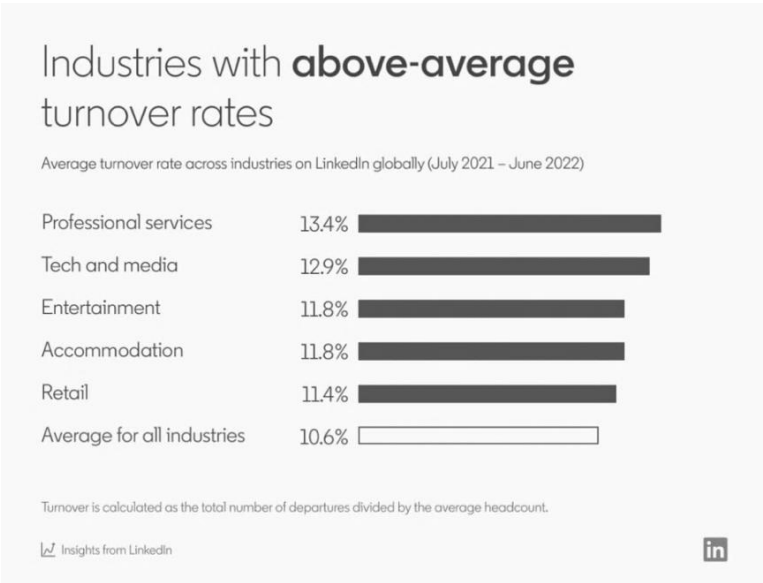


Figure 1 – Industries with above-average turnover rates (LinkedIn, 2022)

2.2. Employee Retention and Turnover

2.2.1. The Importance of Employee Retention

After examining the challenges associated with employee turnover in the IT sector, the discussion now addresses employee retention as a crucial element for ensuring organisational stability and long-term success. "Employee retention refers to a company's capacity to keep qualified staff for the long term. Retention methods frequently include creating programs that attract and fulfil employee requirements, such as professional advancement possibilities, competitive salary, and a work environment that promotes work-life balance. High retention rates reflect employee satisfaction and organizational stability, as well as the potential to save money on hiring and training new staff" (Susanto & Rony, 2023, as cited in Trisna Wijayanthi et al., 2024, p. 290).

"Long-term success and health of the company are based on the retention of skilled employees that helps in meeting the expectations of client and increasing the productivity of the business" (Vasantham & Swarnalatha, 2013, as cited in Khera & Divya, 2019, p. 13). The resource-based view of firms recognizes human capital as a scarce and valuable resource that can provide competitive advantage (Mackey et al., 2013; R. Castanias & C. Helfat, 2001). Human resources are considered strategic assets due to their unique skills, knowledge, and social architecture that develop over time (Mueller, 1996).

Conversely, it’s also important to assess the contributions of employees to the organisation's overall profitability. Such an exercise can help determining which employees are essential to retain and where it might be more beneficial to part ways. Measuring employee profitability is crucial for businesses to optimize their human capital investments; one of the approaches proposed is the return on investment (ROI) of human capital (Cnor, 2004). Several studies have explored different approaches to quantifying this ROI. Ulrich & Smallwood (2005) introduced the concept of "return on intangibles" as a new HR ROI metric. Ben-Gal (2019) provided an ROI-based review of HR analytics, identifying workforce planning and recruitment as high-ROI areas. Baldi & Trigeorgis (2020) applied real options methodology to quantify flexible human capital, particularly in career development programs.

In this context, predictive analytics plays a pivotal role by identifying the characteristics of employees who demonstrate high performance or possess strategic value, as well as those at greater risk of voluntary departure. Rather than applying blanket retention measures to the entire workforce, organizations can use such insights to target retention initiatives more precisely, prioritizing resources for top-performing or critical talent. By helping to determine exactly who

should be the focus of more costly or intensive retention strategies, this work contributes to optimizing the effectiveness and efficiency of human capital management.

2.2.2. Factors Influencing Employee Turnover

According to Abii et al. (2013), examining turnover factors in the technology industry is essential to identify the factors that influence employees to leave organisations and help company leaders predict turnover trends. Many researchers have examined the factors that influence employees to remain in organisations (Capps, 2007; Carson, Carson, Birkenmeier, & Toma, 2006; Chang & Lee, 2006; Chiaburu & Marinova, 2006). According to Lock (2003, as cited in Abii et al., 2013, p. 740), "challenging work, respect, a balanced work life, effective leadership, and a workplace that fosters risk-taking strengthens social bonds among IT employees".

In this context, it's essential for leaders and decision-makers, such as managers and department heads, to cultivate a work environment where employees feel safe and encouraged to explore new ideas, take on challenging tasks, or make decisions without fear of consequences or judgment if things do not go as anticipated. In the case of IT workers, where rapid technological advancement is usually associated with innovation and flexibility, such a risk-taking culture could contribute to employee motivation and a better interpersonal relationship - as employees can support each other in a variety of creative trials and failures.

Horn and Griffeth's (1995) turnover model illustrated job satisfaction and organisational commitment as two antecedents of voluntary turnover (as cited in Abii et al., 2013). Additionally, Horn et al. (1992) noted that employees who are dissatisfied with their jobs or who have lost organisational commitment would have intentions or make decisions to quit (as cited in Abii et al., 2013).

On the other hand, Storey (2016, as cited in Khera & Divya, 2019) suggests that the success of a business is based on the retention of employees and job satisfaction, as these two factors are closely connected. Job satisfaction often serves as a foundation for employee retention, yet various issues—such as ineffective work practices, lack of management support, inadequate compensation, and poor work-life balance—can drive employees to leave. "These factors influence an employee to leave the current workplace with the aim to search for better opportunities. In case the current workplace is interesting and meeting the requirement of employees, then they extend their service period. Owing to this, management is tasked with

creating a comfortable work environment and appropriate pay scale in the line of the ability of hired employees" (Perryer et al., 2010, as cited in Khera & Divya, 2019).

Pourkhodabakhsh, Mamoudan & Bozorgi-Amiri (2022) suggest that employee turnover is often the result of poor hiring decisions and ineffective management, directly affecting the organisation's strategic plans. Specifically, poor management and managerial performance gaps in decision making can increase the employee turnover rate, pushing many companies to adjust, improve and upgrade their management systems to reduce turnover rates (Zhu et al., 2017, as cited in Pourkhodabakhsh et al., 2022).

"It can be said that maintaining and optimizing expensive and scarce resources such as human recourses is one of the most important concerns of managers, which requires more advantage and more innovative tools to improve decisions. This is the reason why identifying the factors affecting employee turnover is a very valuable measure that can lead to corrective measures to prevent employee turnover and helps managers to making better decisions." (Pourkhodabakhsh et al., 2022, p. 16310).

2.3. Employee Engagement

2.3.1. Role of Engagement in Retention

"Job engagement is the degree to which individuals are psychologically and emotionally invested in their jobs. Engaged employees are more likely to be motivated, perform well, and stay with the organisation. High job involvement minimizes voluntary turnover because employees are happy with their positions and perceive meaning and purpose in what they do. They are also more likely to stay with the company and engage in constructive internal turnover, which is when employees seek new roles or responsibilities inside the organisation rather than leaving." (Trisna Wijayanthi et al., 2024, p. 296).

Employee engagement is a critical concept in HR management within the IT industry, linked to job satisfaction, performance, and organisational commitment. Studies show that engaged employees are more productive and less likely to leave their jobs (Kamalanabhan et al., 2009; Aggarwal et al., 2022). Factors influencing engagement include organisational culture, transformational leadership, and flexible HR management (Sekhar et al., 2017; Bhuvanaiah & Raya, 2016). Workplace spirituality, organisational justice, and perceived organisational support also positively affect engagement (Sharma & Kumra, 2020; Aggarwal et al., 2022).

Engagement is associated with organisational citizenship behaviours and can differ between genders and company types (Saradha & Patrick, 2010). In downsized organisations, where workforce reductions frequently increase uncertainty and stress, effective internal communication, knowledge sharing, and continuous learning can contribute greatly to maintaining employee engagement (Tiwari & Lenka, 2020). HR professionals play a crucial role in developing engagement strategies to improve organisational performance and workplace culture (Shuck & Reio, 2011). Engaged employees demonstrate higher job performance and better mental health outcomes (Sekhar et al., 2017; Sharma & Kumra, 2020).

Kahn (1990) established that there is a positive association between an employee's engagement and the psychological impact their organisation has on them, as cited by Pandita & Ray (2018). "When employees find coherence in their work and a sense of assurance from their employers, they tend to be more engaged at their workplace, indicating that focusing on the simple things has become imperative at present. This sense of safety can be provided to employees by forging deep emotional and intellectual bonds with an employee, guaranteeing citizenship behaviour and retention in return" (Gibbons, 2006, as cited in Pandita & Ray, 2018, p. 187).

2.3.2. Measuring Employee Engagement

"One of the most effective tools in ensuring that employees stay engaged and committed to their work is talent management. This sense of engagement or commitment toward their work ensures in turn that these employees stay with the organization in the long run" (Pandita & Ray, 2018, p. 185). According to Pandita & Ray (2018), as organisations began to analyse the costs associated with recruitment, selection, and the opportunity cost related to attrition, they increasingly concentrated their talent management efforts in the direction of retention. One of the key findings of their study was that aligning talent management practices with employee engagement initiatives significantly enhances talent retention.

Employee engagement is a multifaceted construct that organisations measure using various tools and methods, such as surveys, feedback systems, and performance evaluations. Surveys are commonly used, with several validated scales available, including the Korean Employee Engagement Scale (Song et al., 2021) and a 3-factor/16-item scale developed for the Indian context (Shrotryia & Dhanda, 2020). The Management Standards Indicator Tool has also been used to assess engagement (Ravalier et al., 2015). Performance management processes, including performance agreements, ongoing feedback, and evaluations, can enhance engagement when focused on employee development, involvement, trust, and justice (Saratun,

2016). However, choosing the most appropriate employee measurement scale depends on organisational context and decision-maker preferences (Khodakarami et al., 2018).

Furthermore, the transformation of work contexts, including remote work and digital technologies, presents new challenges for engagement research and practice (Kabalina & Makarova, 2022), particularly within companies in the IT sector. Remote work often leads to longer work hours and increased work intensity (Felstead & Henseke, 2017; Hill et al., 1998), potentially blurring work-life boundaries and negatively impacting work-life balance (Sandoval-Reyes et al., 2021; Tejero et al., 2021). Organisations must adapt their engagement strategies, leveraging digital tools to foster collaboration, recognition, and emotional connection among remote workers (Gupta et al., 2021; Baptista et al., 2020).

2.4. Human Resource Management through Analytics

2.4.1. Predictive Modeling in Human Resources

Human Resource (HR) analytics involves gathering and analysing data related to an organisation's human resources to identify relationships between HR practices and overall organisational performance. This approach provides the organisation with valuable insights to support data-driven decision-making and strategic planning (Mansi & Goswami, 2024). "Analytics is mainly categorized into descriptive (what has happened), predictive (what will happen), and prescriptive analytics (what should be done)" (Nagpal & Mishra, 2021, as cited in Mansi & Goswami, 2024, p.87).

In human resource management, predictive analytics has become one of the most effective tools in forecasting employee behaviour, especially in predicting voluntary attrition and spotting high-risk employees. "Businesses are using data-driven techniques such as predictive analytics, driven by machine learning algorithms, to address the challenge of employee attrition. Machine learning is a subcategory of AI technology that uses statistical algorithms and computational models to enable computers to learn from data without explicit programming. Employee attrition is a significant concern for companies as it can negatively impact productivity, financial performance, and workforce morale." (Krishna & Borah, 2024, p. 15).

Machine learning models such as decision trees, random forests, and gradient boosting are found to be highly accurate in predicting employee turnover (El-Rayes et al., 2020; Jain et al., 2020). These models consider a variety of variables, such as compensation, organisational culture, and management performance, to identify potential attrition risks (El-Rayes et al., 2020). The application of explainable AI techniques increases the interpretability of such

models, helping HR professionals to understand the underlying factors driving turnover (Marín Díaz et al., 2023). This shift from big data to deep data focuses on data quality rather than data quantity, hence increasing prediction accuracy (Yahia et al., 2021). The application of predictive analytics in HR can help organisations develop targeted retention strategies, improve employee satisfaction, and reduce attrition rates (Singh et al., 2022).

According to Krishna & Borah (2024), in the context of machine learning, predicting employee attrition can be considered a classification problem, where the goal is to categorize employees into two distinct groups: those who will probably quit the organisation and those who will not. The process of constructing an employee attrition prediction model typically follows a series of systematic stages (Krishna & Borah, 2024), listed below:

i. Data Collection: "The goal of data collection is to gather relevant information regarding employees and their job functional areas to create a model capable of predicting the probability of an employee departing the organisation." (Krishna & Borah, 2024, pp. 15-16).
ii. Data Analysis: "The principal objective of data analysis is to extract valuable insights from available data." (Krishna & Borah, 2024, p. 16). This stage involves data cleaning, randomization, as well as visualization.
iii. Feature Scaling: "Scaling aims to guarantee that numerical input variables are on an equivalent scale, which can lead to increased machine-learning algorithm performance." (Krishna & Borah, 2024, p. 16). Common scaling techniques include MinMaxScaler, Z-score (StandardScaler), Robust Scaler, and Quantile Transformer (Ahsan et al., 2021).
iv. Machine Learning Model Development: Utilizing ML algorithms such as Logistic Regression, Random Forest, and Support Vector Machines to build predictive models (Krishna & Borah, 2024).
v. Performance Evaluation: Assessing the models' effectiveness using metrics like the confusion matrix and AUC-ROC curve (Krishna & Borah, 2024).

"Data visualization is a methodology for visually representing and summarizing data to gain more information and insight. In the context of employee attrition prediction, data visualization helps to understand the dataset's various features and their impact on the target variable." (Krishna & Borah, 2024, p. 16). The target variable in these models typically refers to whether an employee will leave the company or not (Jain et al., 2020). This variable is typically binary, with values such as "1" indicating that the employee is predicted to leave and "0" indicating that the employee is predicted to stay.

2.4.2. Strategic HR Management through Analytics

HR analytics significantly enhances decision-making across various human resource management areas, including recruitment, talent management, and employee development (McCartney & Fu, 2022). This data-driven approach provides valuable insights that enable more informed and strategic HR decisions. “In other words, HR analytics not only centres on investigating and improving elements of human capital but also applying analytical techniques coupled with people data to inform organizational strategy and improve performance.” (McCartney & Fu, 2022, p. 26).

In the realm of recruitment, HR analytics can significantly improve the selection process and outcomes. McCartney & Fu (2022) note that HR analytics enables organisations to identify the most effective recruitment sources, predict candidate success, and assess the return on investment of recruitment channels. El-Rayes et al. (2020) provide a concrete example of how analytics can enhance recruitment. They describe a case where HR analytics was used to build a profile of high-performing agents by analysing factors such as age, gender, performance, and previous work experience. This profile was then used to inform the recruitment and interview process, with the goal of improving retention rates.

On the other hand, HR analytics also plays an essential role in talent management by providing insights that can inform strategic decisions. One specific application of HR analytics in talent management is workforce planning. El-Rayes et al. (2020) describe a sophisticated workforce management tool used for both day-to-day management and strategic capacity planning. This tool utilizes simulation modeling to predict service levels based on various workforce parameters, allowing for informed decisions on workforce reorganisation and performance management (El-Rayes et al., 2020).

Concerning employee development, HR analytics can help identify skill gaps and development needs within the workforce. Mansi & Goswami (2024) observe that HR analytics allows organisations to determine the most effective training initiatives, anticipate problems before they arise, and recognize factors influencing employee turnover. This information can be used to make strategic decisions about employee development programs and interventions.

2.4.3. Ethical Considerations in HR Analytics

HR analytics presents both opportunities and challenges for organisations. While it can enhance decision-making and improve organisational performance, as previously explained, there are significant ethical concerns regarding privacy, transparency, and fairness (Chen et al., 2017).

Employees often lack oversight of how their data is used, which can lead to potential misuse or hidden biases (Zieglmeier & Pretschner, 2023). To address these issues, companies should implement governance structures that include stakeholders from the earliest stages of development and provide transparency into how inferences are drawn from employee data (Chen et al., 2017). The concept of "inverse transparency by design" has been proposed as a promising approach to responsible people analytics (Zieglmeier & Pretschner, 2023). However, there are concerns that current HR analytics practices may fail to deliver transformational change and could potentially harm employee interests (Angrave et al., 2016). Additionally, firms must be cautious of constructing "predatory predictions" that unfairly categorize individuals for organisational benefit (Martin, 2023).

In line with recent regulatory developments, the European Union's Artificial Intelligence (AI) Act — formally adopted as Regulation (EU) 2024/1689 — classifies AI systems used in human resources as *high-risk applications*. This includes AI tools used in recruitment, performance evaluation, promotion, and termination processes, since these can greatly impact individuals' rights, privacy, and access to employment. Therefore, organisations using predictive or analytical models in HR must comply with strict requirements related to transparency, accountability, and data governance to ensure that these systems operate fairly and without discrimination (Regulation (EU) 2024/1689).

As HR analytics continues to evolve and play a more significant role in organisational decision-making, addressing ethical considerations becomes increasingly important. Mansi & Goswami (2024) emphasize that organisations must address ethical concerns, establish robust safety and security measures, and consider the psychological consequences of HR analytics prior to its implementation, as it will lessen the negative impact of analytics.

Chapter III - Approach and Methodology

This study adopts a predictive research methodology, utilizing actual human resources data to forecast employee departures and assess loyalty trends. This approach aligns with the thesis topic, as predictive analytics relies on identifying patterns and relationships within data to generate informed forecasts. A combination of quantitative and qualitative methods is utilized**.** Quantitative data include historical employee information such as compensation, performance ratings, and attrition rates, whereas qualitative insights were obtained through sentiment analysis conducted by the company's internal analytics teams based on a structured employee feedback mechanism. By integrating these data sources into a predictive model, the research analyses how effectively machine learning models can anticipate voluntary attrition events. This methodological approach not only fulfils the predictive nature of the study but also allows for a comprehensive analysis of both quantitative metrics and qualitative nuances impacting employee retention.

The scope of this project encompasses several key activities aimed at developing a predictive model for employee attrition in an information technology (IT) company, following the CRISP-DM (*Cross-Industry Standard Process for Data Mining*) methodology. CRISP-DM is a robust and widely used framework for structuring data mining and ML projects, providing a systematic and comprehensive approach consisting of seven stages (Figure 2): Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment and Model Management (Wehrstein & Bachmann, 2020).

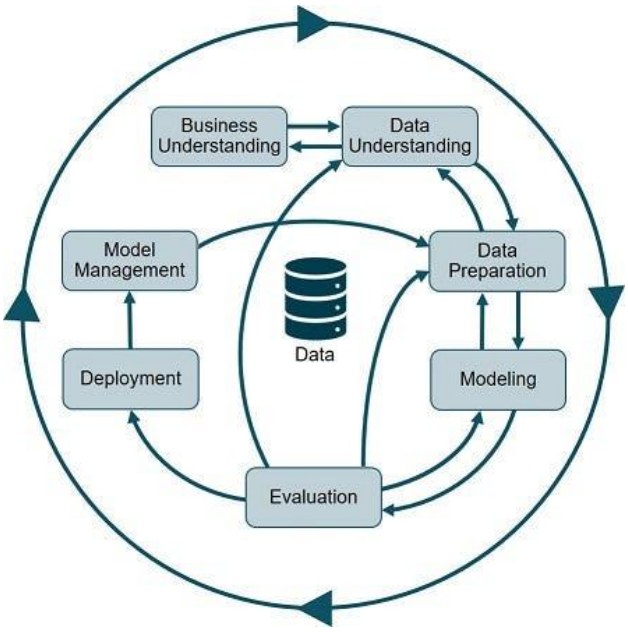


Figure 2 – CRISP-DM Stages

From a technical perspective, the Data Understanding, Data Preparation, Modeling and Evaluation phases were conducted using Python within the JupyterLab environment, which was made available by the company via Google Cloud Storage. The implementation utilized essential machine learning libraries including scikit-learn, pandas, matplotlib, XGBoost, LightGBM, and CatBoost. This technological infrastructure provided the computational resources and methodological flexibility necessary to implement the CRISP-DM process effectively.

The following subchapters address all seven stages of the CRISP-DM framework: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, and Model Management. The first five stages collectively form the methodological foundation of the study by defining the problem, exploring and preparing the data, developing the predictive models, and assessing their performance. The final two stages, Deployment and Model Management, are included only at a conceptual and planning level, as their practical implementation lies outside the scope of this research.

3.1.Business Understanding

The Business Understanding phase is the first stage of the CRISP-DM methodology, and it plays a key role in aligning machine learning objectives and analytical activities with organisational needs. According to the framework, this stage involves five main activities: (1) determining the business objectives, (2) determining the analytics outcome, (3) assessing the situation, (4) assessing the IT infrastructure, and (5) producing a project plan (Wehrstein & Bachmann, 2020). In the context of this study, each of these activities has been adapted to the business problem under investigation—voluntary employee attrition in a Portuguese information technology (IT) company.

This challenge was explored through a collaborative project conducted in partnership with professionals from the Human Resources and People Analytics departments of a large Portuguese organisation operating in the IT sector. The organisation provides a range of technology-driven services, including software development, IT consulting, and digital transformation solutions for clients across multiple industries. With offices in several cities across Portugal, the company maintains a geographically distributed workforce that supports both national and international operations. Due to confidentiality constraints and a non-disclosure agreement (NDA) signed between the researcher and the company, specific details such as the company's name, size, and internal structure cannot be disclosed. Nevertheless, the

contextual information provided is sufficient to understand the organisational environment in which this study was conducted.

The primary business objective of this research is to address the challenge of voluntary employee departures - defined in this dataset as employees who have willingly left the company, as captured by the *Attrition* feature (see Annex A, pp. 70–78, for further detail). Although attrition represents only 8.1% of the company's workforce, it is perceived internally as having a significant impact.

During the Business Understanding phase, several meetings were held with members of the company's HR teams to discuss the organisational context and challenges related to employee turnover. The perception shared in these meetings was that the loss of talented professionals in a knowledge-intensive sector such as IT often leads to project delays, increased recruitment and training expenses, and disruption of institutional expertise. Moreover, voluntary attrition was viewed as particularly concerning, as it reflects employees' dissatisfaction, lack of engagement, or the search for external job opportunities—factors that can, in principle, be mitigated through targeted HR interventions. As a result, employee retention emerged as a strategic priority for the organisation.

From an analytical perspective, the project aims to develop a ML classification model that can assess the probability of an employee voluntarily departing from the organisation. In addition to prediction, the model provides insights into feature importance, allowing the organisation to understand which variables have the most significant impact on employee attrition. This dual outcome, combining predictive capabilities with interpretability, ensures that the model is both technically robust and practically useful for HR decision-making.

The assessment of the company's situation regarding employee attrition revealed a clear class imbalance across the years under analysis (2022–2024), based on the data provided by the company. The annual attrition rates demonstrate a downward trend, with approximately 10.5% of employees leaving the organisation in 2022, followed by 7.1% in 2023, and 6.6% in 2024 (Table 1). This gradual decline suggests that employee retention efforts may have strengthened over time, although the overall imbalance remains substantial, with an average of 8.1% of employees voluntarily leaving (*Attrition = 1*) compared to 91.9% who stayed (*Attrition = 0*).

Table 1 - Annual Attrition Rates (2022–2024)

Year	**2022**	**2023**	**2024**	**Average Attrition Rate**
Attrition Rate (%)	10.46	7.13	6.58	8.06

Although this imbalance presents a methodological obstacle for supervised learning models, as algorithms might tend to predict the majority class, its business implications remain critical. Even a small number of voluntary departures can have a substantial effect on project continuity and talent availability in a competitive labour market such as the IT industry.

Regarding IT infrastructure, the dataset used in this research was originally stored within the organisation's internal Human Resources Information Systems, which maintain detailed records on employee demographics, professional trajectories, and engagement indicators. However, due to the organisation's strict confidentiality protocols and data protection policies, direct access to these internal databases was not granted. Instead, the company provided the data in an aggregated and anonymized format through Microsoft Excel files, ensuring compliance with internal security and privacy protocols. This approach, however, limits the level of detail, or granularity, available for certain variables, which may have a relevant impact on the model's performance, as further discussed in Chapter 4.3. Additionally, the dataset was composed of three separate sheets, each corresponding to a distinct year (2022, 2023, and 2024), representing annual snapshots of the workforce.

Lastly, the project plan included all the previously described stages of the CRISP-DM framework, with each phase executed iteratively to refine data quality, model performance, and interpretability. The final deliverable—an operational predictive model—serves as a decision-support tool for the company's HR department, offering actionable insights regarding the probability and factors influencing attrition. Furthermore, the company is responsible for monitoring and maintaining the model to ensure its ongoing relevance as workforce dynamics change.

3.2.Data Understanding

3.2.1. Dataset Description

For this thesis project, a real employee dataset was provided by the partner company, containing anonymized and encoded information due to confidentiality requirements. The dataset includes a combination of demographic, organisational, career development, compensation, leadership, and satisfaction variables, allowing for a multifaceted analysis of the factors influencing the voluntary employee attrition rate within the company. These variables encompass both categorical and numerical data, offering a comprehensive perspective on workforce characteristics, career trajectories, and organisational dynamics. For a detailed explanation of each variable, please refer to Annex A (pp. 70–78).

The dataset comprises 5728 records across 47 variables, including data from three consecutive years (2022, 2023, and 2024). Since the same employees are observed across multiple years, the dataset does not represent 5728 unique individuals, but rather repeated annual snapshots of the workforce. The target variable in this study is *Attrition*, which indicates whether an employee has voluntarily left the company during the corresponding year. This means that, for instance, an employee appearing in multiple yearly records may have a value of 0 ("No") in the earlier years and a value of 1 ("Yes") only in the year of their departure. The variable is therefore binary, where a value of 1 represents "Yes" (the employee has exited the organisation) and 0 represents "No" (the employee remains employed). During these 3 years, attrition was on average 8.1%, meaning that most employees remained in the organisation throughout the period in analysis.

Due to confidentiality restrictions, several sensitive attributes were aggregated and anonymized by the partner company before being shared for analysis. This ensured that no personally identifiable or sensitive organisational information could be inferred from the data. As a result, attributes such as employees' exact age, specific job level categories, department names, office locations, or detailed salary amounts were not provided in their raw form but instead delivered as generalized, anonymized categories.

For instance, *Generation* was not derived from exact birth years but instead provided in grouped cohorts (*Baby Boomer*, *Generation X*, *Millennial*, *Generation Z*). Similarly, *Tenure*, which would typically indicate the precise number of years an employee has worked at the company, was aggregated into experience-based categories such as *New Joiners*, *Pathfinders*, *Growing Talent*, *Core Contributors*, *Legacy Builders*, and *Experienced Veterans*. The attribute *Location* was anonymized into generic identifiers (*A*, *B*, *C*, *D*) instead of detailed office addresses. At the organisational level, *Board_Member_Oversight* was grouped into anonymized categories representing responsibility areas of board members (A–H), while *Department* and *Unit* were delivered as numeric identifiers rather than descriptive names.

Other structural attributes were also aggregated to protect confidentiality. *Job_Level*, for instance, was provided as hierarchical categories (e.g., *EL*, *ML*, *PL*, *SL*), while *Total_Annual_Compensation* was expressed in bands such as *Very Low*, *Low*, *Mid*, *High*, *Very High*, and *Ultra High*. Likewise, *Salary_Increase*, instead of reporting exact percentages, was delivered as predefined percentage intervals (e.g., 1–3%, 5–7%, 9–11%).

3.2.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. An EDA was conducted on the HR dataset to understand data distributions, identify patterns, and discover relationships between the variables.

First, a univariate analysis was undertaken to examine the distribution of both numerical and categorical features within the dataset. Then, a bivariate analysis was performed to explore the relationships between the target variable (*Attrition*) and the most influential predictors identified for the model. Lastly, a correlation analysis was performed to assess the strength and direction of linear relationships among numerical variables, supporting the identification of multicollinearity and inter-feature dependencies prior to model development.

3.2.2.1.Univariate Analysis

In this subchapter, a univariate analysis is conducted to examine the distribution of both numerical and categorical variables in the dataset, providing an initial understanding of employee-related characteristics and behavioural patterns.

Firstly, a grid of small-multiple histograms (also known as a faceted plot) was created to visualize the distribution of each numerical feature (see Annex B, pp. 79). This data visualization approach allows for a simultaneous and consistent comparison across variables, highlighting the shape, spread, and balance of each feature.

The following conclusions regarding the numerical continuous variables within the dataset can be derived from the faceted plot:

i. *Year*. The dataset covers three periods (2022, 2023 and 2024) with near-equal sample sizes per year. A gradual decline in headcount is noticeable from 2022 to 2024, which may reflect the cumulative effect of employee attrition or organisational restructuring over time.
ii. *Department* and *Unit*. Both variables, though numerically encoded, represent categorical identifiers for organisational divisions. Their uneven distributions indicate varying employee concentrations across departments and units, suggesting the need to treat them as categorical features during modeling.
iii. *Functional_Promotion_Interval*. This variable represents the time interval, in months, between each functional promotion of an employee. The distribution is

right-skewed, showing that most employees experience promotions within shorter intervals, while only a smaller portion have significantly longer promotion intervals. This suggests that career progression tends to occur relatively early for most employees, with fewer cases of delayed advancement.

iv. *Variable_Compensation*. This feature represents the percentage of variable remuneration within an employee's total annual salary. The distribution is right-skewed, with most employees receiving a variable compensation between 10% and 20% of their salary, while only a few receive higher proportions. This suggests that variable pay is generally modest for most employees, with higher incentives limited to a smaller group, likely associated with performance-based or leadership roles.

v. *Annual_Unemployment_Rate*. This variable was included in the dataset, at the company's request, to capture external economic conditions that could influence employee attrition. It represents the national unemployment rate corresponding to each year in the dataset, sourced from Statistics Portugal (INE, 2025). Although the chart displays only two distinct values due to identical rates in 2022 and 2023 (6.4%), the feature effectively captures annual economic conditions across the three years, with a slight improvement observed in 2024 (6.4% compared to 6.8% in 2022–2023). This stability suggests limited year-to-year variation in external labour market conditions during the analysed period.

vi. *Consumer_Price_Index.* Also added at the company's request to account for external economic influences that may affect attrition behaviour, this feature measures inflation in Portugal, expressed as an index reflecting the average change in prices of goods and services consumed by households. The distribution shows three distinct values, corresponding to the years 2022 (113.383), 2023 (118.271), and 2024 (121.129), indicating a steady increase in the cost of living over time and reflecting the overall inflationary trend during the period.

On the other hand, the faceted plot also reveals the presence of multiple binary features — variables that display values of "0" and "1" on the X-axis (see Annex B, pp. 79) — where "0" means *No* and "1" means *Yes*. Although these variables are conceptually categorical, Python automatically interprets them as numerical when generating the plots. These binary features capture the presence or absence of specific employee attributes or events (e.g., having children, leadership roles, promotions, company car, etc.), and their imbalanced distributions provide early evidence of potential class asymmetry that may influence model performance.

Across the binary features (*Leadership_Role*, *Internal_Mobility*, *Functional_Promotion*, *Change_of_Direct_Supervisor*, *Children*, *Promotion_in_the_Last_3_Years*, *Promotion_in_the_Last_5_Years*, *Company_Car*, *Career_Opportunity*, and *Change_of_Direct_Supervisor_due_to_Mobility*), class distributions are visibly imbalanced—most show a dominant "0" class with a minority of "1"s. This pattern suggests that only a minority of employees have experienced such events or hold specific roles, such as leadership positions, promotions, or internal mobility. With respect to the target feature, as previously explained, the distribution is significantly imbalanced, with the "No Attrition" class predominantly prevailing in the dataset. This confirms the necessity for imbalance-aware evaluation metrics and modeling approaches. Accordingly, the modeling phase incorporated class weighting to mitigate the bias toward the majority class, and model performance was primarily assessed using the F1 score, which provides a balanced measure of precision and recall under class imbalance conditions.

To analyse the categorical variables within the dataset, a grid of small-multiple bar charts (a faceted bar plot) was produced, where each subplot displays the frequency distribution of a single categorical feature (see Annex C, pp. 80-81). This visualization provides a detailed demographic and organisational profile of the employee population, highlighting key imbalances and dominant patterns across functional, generational, and structural dimensions:

i. *Gender* and *Generation*. The workforce is predominantly male, with a smaller female representation. In terms of age composition, Generation X and Millennials constitute most employees, while Generation Z and Baby Boomers account for only a small fraction. This reflects a mid-career-dominant population, suggesting potential stability but limited generational renewal.

ii. *Location*. Employees are unevenly distributed across locations, with Location B hosting the largest share of staff. This geographic concentration may indicate regional operational hubs or central business units.

iii. *Job Level* and *Functional Area.* The *Job_Level* feature represents the company's internal hierarchy, and it is sorted by seniority level as follows: *EL*, *ML*, *MLL*, *PL*, *AL, SL, ALL,* and *SLL* - with the first level being entry level, followed by mid-levels and finally strategic level. Lower-seniority levels (particularly *PL* and *MLL*) concentrate most employees, while higher-seniority levels become progressively less represented, reflecting a pyramidal organisational structure. Similarly, Technology, Product and Marketing, Corporate Support, Business Support and

Commercial are the most represented functional areas within the organisation, showing where most employees are concentrated.

iv. *Board Member Oversight*. This feature groups departments under the supervision of different executive board members. The distribution shows a clear imbalance, with a few board members (notably A and B) overseeing a much larger number of departments. This pattern suggests structural asymmetry in governance responsibilities, which could reflect differences in departmental scope, organisational structure, or strategic prioritization within the company.

v. *Performance Rating*. This feature assesses employees on a scale from 1 (insufficient) to 5 (excellent), shows a strong concentration around rating 3, followed by 4, while higher ratings (5) and lower ones (1–2) are relatively rare. Additionally, a considerable number of employees are categorized as *"No Data Available"*, indicating ineligibility for evaluation (e.g., due to recent hiring, leave, or seniority). This distribution suggests that most evaluated employees perform satisfactorily, though a portion lacks recent appraisal data.

vi. *Tenure.* This feature, which classifies employees based on length of service, displays a predominance of *Experienced Veterans* and *Legacy Builders*, followed by *New Joiners* and *Pathfinders*. This indicates a workforce with a strong presence of long-tenured employees, balanced by a moderate influx of younger staff, suggesting both institutional stability and ongoing renewal within the organisation.

vii. *Type of Cross-Functional Project* and *Work Model*. Most employees are not involved in cross-functional projects, which may imply limited interdepartmental collaboration. Furthermore, the hybrid work model is predominant, reflecting modern post-pandemic organisational practices, followed by a smaller group of remote workers.

viii. *Total Annual Compensation* and *Salary Increase*. Most employees fall into the lower-mid salary ranges and report modest annual salary increases (typically below 5%), indicating a conservative compensation growth pattern across the organisation.

Lastly, the dataset includes a wide set of survey-based features designed to capture employees' perceptions and attitudes across several organisational dimensions, such as engagement, enablement, commitment, loyalty, effectiveness, and satisfaction with various aspects of the work environment (e.g., autonomy, collaboration, compensation, leadership trust,

training, ways of working, performance management, among others). Overall, these variables display a similar right-skewed pattern, with most responses concentrated around mid-to-high ratings (typically "3" and "4"), and a substantial portion categorized as "No Data Available", meaning incomplete survey participation or ineligibility in some cases. This general pattern indicates that employees tend to report moderately positive experiences, though variability across items implies differing satisfaction levels depending on the specific organisational domain assessed.

3.2.2.2.Bivariate Analysis

In this subchapter, multiple bivariate analyses are performed to explore the relationships between selected dataset features and the target feature, *Attrition*. The features analysed were chosen based on two complementary perspectives: business interpretation and modeling relevance. From a business perspective, the variables *Generation*, *Job_Level*, *Department, Functional_Area*, *Work_Model* and *Performance_Rating* were examined, as they provided particularly distinctive patterns regarding employee behaviour and organisational structure. From a modeling perspective, attention was given to the top five most influential predictors identified by the final predictive model (LightGBM), specifically, *Functional_Promotion_Interval*, *Variable_Compensation*, *Unit*, *Loyalty_Level* and *Salary_Increase*, which demonstrated the highest feature importance scores (see Figure 7) and will be further discussed in Chapter 4.3. Regarding data visualisation, stacked bar charts were used to illustrate the differences in attrition rates across the categories or value ranges of each feature, thus enabling a deeper understanding of how both business-relevant and model-driven factors influence employee attrition.

Generation

The stacked bar chart entitled "*Attrition Rate by Generation*" (see Annex D.1, pp. 82) illustrates the relationship between employee attrition and generational group. The *Generation* attribute was categorized into *Baby Boomer*, *Generation X*, *Millennial*, and *Generation Z* for confidentiality reasons since individual ages could not be disclosed. As presented in this chart, Generation Z and Millennial employees exhibit the highest attrition rates, at approximately 17.3% and 13.2%, respectively. In contrast, Generation X (2.8%) and Baby Boomers (4.2%) show substantially lower attrition rates. This generational disparity suggests that younger cohorts are more likely to leave the organisation compared to older employees. The difference

may reflect variations in career expectations, job stability preferences, or organisational commitment across age groups.

Consistent with this interpretation, the *Generation* distribution (see Annex C, pp. 80-81) reveals that Generation X and Millennial employees together account for the largest portions of the workforce, followed by Generation Z, while Baby Boomers represent a minimal share. Therefore, although attrition is more prevalent among younger generations, its overall organisational impact is amplified by the larger population of Generation X and Millennial employees. These findings emphasize the need for generation-specific retention strategies, particularly targeting younger employees who may prioritize career mobility, learning opportunities, and work-life balance.

Job Level

The stacked bar chart entitled *"Attrition Rate by Job Level"* (see Annex D.2, pp. 82) demonstrates the relationship between employee attrition and the hierarchical level within the organisation. The *Job_Level* attribute reflects the company's internal structure, ordered from entry to strategic positions — specifically, *EL, ML, MLL, PL, AL, SL, ALL,* and *SLL* — where the first levels represent entry and mid positions, and the latter denote higher strategic roles. As illustrated in this chart, attrition rates are higher among lower and middle job levels (notably EL, ML, and MLL), while senior and strategic levels (such as SL, ALL, and SLL) display markedly lower attrition. This pattern indicates that employees in the initial stages of their careers are more prone to leaving the company, potentially due to greater external mobility, pursuit of advancement opportunities, or shorter tenure within the company. In contrast, those in higher positions tend to demonstrate stronger organisational commitment, reflecting greater career stability and alignment with long-term strategic objectives.

Department

The stacked bar chart entitled "*Attrition Rate by Department (Top 10 by Headcount)*" (see Annex D.3, pp. 83) shows the relationship between employee attrition and the organisation's largest departments, focusing only on the top 10 by headcount to enhance visualization and ensure a more meaningful comparison. The *Department* attribute represents the specific divisions within the organisation responsible for particular functions and operations. For confidentiality reasons, departments were anonymized and identified numerically (e.g., Dept. 7, Dept. 10, Dept. 41), as their actual names could not be disclosed. As depicted in this chart,

attrition rates vary across departments, ranging from 3.6% to 14.1% among the ten with the highest headcounts. The highest attrition was observed in Dept. 7 (14.1%), followed by Dept. 19 (11.5%) and Dept. 10 (10.4%), while Dept. 44 (3.6%) and Dept. 18 (4.2%) recorded the lowest. These disparities suggest that certain divisions may experience greater employee departure pressures, potentially due to differences in work environment, job demands, or managerial practices.

Functional Area

The stacked bar chart entitled "*Attrition Rate by Functional Area*" (see Annex D.4, pp. 83) illustrates the relationship between employee attrition and the organisation's main functional clusters. The *Functional_Area* attribute represents groups of departments with related functions and expertise within the organisation, categorized according to the nature of the work performed, required skills, and overall functional objectives. The chart reveals that the highest attrition occurs in Product and Marketing (13.8%) and Corporate Support (12.2%), followed by Business Support (7.6%) and Commercial (7.0%), while Others (1.8%) and Information Systems (4.4%) display the lowest attrition. These variations suggest that employees in certain areas particularly those tied to client-facing, strategic, or administrative functions - may experience higher mobility or job transition tendencies.

Supporting this interpretation, the distribution of employees across functional areas (see Annex C, pp. 80-81) reveals that Corporate Support and Product and Marketing—the two areas with the highest attrition rates—also represent the second and third largest segments of the organisation, respectively. In contrast, Technology, which is the largest functional area, exhibits a comparatively low attrition rate (5.4%). This pattern indicates that while employee attrition is concentrated in high-impact areas such as Corporate Support and Product and Marketing, the organisation benefits from greater stability within its largest technical division.

Work Model

The stacked bar chart entitled "*Attrition Rate by Work Model*" (see Annex D.5, pp. 84) illustrates the relationship between employee attrition and the type of work arrangement within the organisation. The *Work_Model* feature specifies whether employees operate in a hybrid, on-site, or remote format, reflecting different levels of workplace flexibility. As evidenced in the chart, the *Remote* model exhibits the highest attrition rate (10.8%), followed by *Hybrid* (7.9%) and *On Site* (6.6%), while *On Site T* records the lowest (4.1%). It is important to note that *On*

Site refers to employees working within the company's facilities, whereas *On Site T* designates those working on-site but outside the company's premises. This pattern suggests that employees with greater location flexibility may be more prone to leaving the organisation, possibly due to broader employment opportunities or differing engagement levels compared to fully on-site staff.

Additionally, the distribution of employees by work model (see Annex C, pp. 80-81) indicates that the *Hybrid* model represents the majority of the workforce, followed by *Remote*, while *On Site* and *On Site T* comprise smaller portions. Consequently, although attrition is proportionally higher among remote and hybrid employees, its overall impact is amplified by their larger share of the organisation's workforce.

Performance Rating

The stacked bar chart entitled "*Attrition Rate by Performance Rating*" (see Annex D.6, pp. 84) illustrates the relationship between employee attrition and annual performance evaluation grades. The *Performance_Rating* feature reflects employee assessments on a 1–5 scale, where 1 indicates insufficient performance and 5 represents excellent performance. The category "*No Data Available*" includes employees who were not evaluated during the reference period, such as those who were recently hired, on extended leave or exempt due to seniority. As can be observed in this chart, employees with the best rating (5) exhibit the lowest attrition (5.1%), while those with ratings of 3 and 4 show higher departure rates (8.2%), similar to those with "No Data Available" (9.2%). Additionally, employees rated 1 recorded no attrition due to the extremely small group size. This pattern suggests that top-performing employees are more likely to remain within the organisation, whereas average-performing or unassessed employees demonstrate higher mobility.

In line with this pattern, the *Performance_Rating* distribution (see Annex C, pp. 80-81) indicates that most employees received grades of 3 or 4, followed by a smaller proportion without available evaluations. Therefore, even moderate differences in attrition among these mid-range groups can significantly influence the organisation's overall turnover dynamics.

Functional Promotion Interval

The stacked bar chart entitled *"Attrition Rate by Functional Promotion Interval (Grouped in Months)"* (see Annex D.7, pp. 85) demonstrates the relationship between employee attrition and the time elapsed since the last functional promotion. The *Functional_Promotion_Interval*

feature represents the number of months between each functional promotion, indicating a transition to a different Job Level. Its distribution is right-skewed (see Annex B, pp. 79), with most employees receiving promotions within the first 30 months, while a smaller proportion remains at the same level for extended periods (100–130 months).

According to the stacked bar chart (Annex D.7), attrition rates are highest among employees whose last promotion occurred between 13 and 36 months (approximately 15–16%), whereas recently promoted employees (9.4%) and those with longer intervals (≤3%) exhibit lower attrition levels. These results suggest that the likelihood of employees leaving the organisation increases when career progression slows, revealing the importance of promotion timing as a determinant of retention. Consistent with this interpretation, *Functional_Promotion_Interval* was identified as the most influential feature in the final selected model (LightGBM), contributing 24% of the total feature importance.

Variable Compensation

The stacked bar chart entitled "*Attrition Rate by Variable Compensation*" (see Annex D.8, pp. 85) illustrates how attrition varies with the proportion of an employee's annual remuneration that is contingent or performance-linked, such as bonuses or incentive pay. Since the *Variable_Compensation* feature is originally expressed as a continuous percentage (ranging from 0% to 47%), the values were grouped into percentage bands to facilitate clearer interpretation and comparison. The chart reveals that the relationship between variable compensation and attrition is non-linear. Employees who do not receive variable compensation (0%) exhibit a higher attrition rate (11.8%). Attrition decreases significantly in the 1–5% band (2.5%), increases in the 6–10% (8.4%) and 11–15% (11.0%) bands, and then declines for higher proportions of variable pay: 16–20% (5.1%), 21–30% (1.9%), and 31%+ (1.2%). These findings indicate that a higher variable compensation component is associated with lower attrition, while the absence of variable compensation and, to a lesser extent, mid-range bands (6–15%) correspond to higher attrition rates.

The corresponding distribution plot (see Annex B, pp. 79) shows that most employees fall between the 6% and 17% range of variable compensation, with far fewer observations in the extreme bands (> 20 %). Consequently, the very low attrition rates observed in the 21–30% and 31%+ ranges should be interpreted with caution, as they likely reflect smaller sample sizes. Finally, *Variable_Compensation* emerged as the second most influential predictor of employee attrition, contributing approximately 10% of total feature importance. This highlights the

relevance of performance-linked pay in shaping retention behaviour, suggesting that reward structure plays a meaningful role in employees' decisions to stay with or leave the organisation.

Unit

The stacked bar chart entitled *"Attrition Rate by Unit (Top 10 by Headcount)"* (see Annex D.9, pp. 86) depicts the variation in attrition rates across the organisation's main operational units. The *Unit* feature represents a specialized subgroup within a department, focusing on a specific function, task, or project, with each unit assigned a unique numerical identifier. As demonstrated in this chart, substantial disparities in attrition are observed between units. The highest attrition rates occur in *Unit 19.2* (20.6%) and *Unit 7.1* (14.5%), while several others—such as *Units 34.3* and *18.2*—report minimal attrition (below 1%). This imbalance indicates that attrition is not uniformly distributed across the organisation but may instead reflect differences in work environment, leadership, workload, or team dynamics.

The distribution histogram of *Unit* (see Annex B, pp. 79) further shows that employee representation varies widely across units, suggesting heterogeneous workforce sizes. Given these patterns, *Unit* was identified as the third most influential predictor in the LightGBM model, contributing 9% of total feature importance, reinforcing the relevance of intra-organisational context in explaining employee attrition.

Loyalty Level

The stacked bar chart entitled "*Attrition Rate by Loyalty Level*" (see Annex D.10, pp. 86) illustrates the relationship between attrition and employee loyalty, measured on a scale from 1 (low) to 5 (high), including a "No data available" category representing employees for whom loyalty data could not be recorded. The *Loyalty_Level* feature assesses how likely an employee is to remain with the company, based on job satisfaction, engagement, and long-term career alignment. The chart reveals a clear inverse relationship between loyalty and attrition: employees with lower loyalty levels show markedly higher attrition rates, with 16.7% attrition at level 1, compared to only 0.4% among those at level 5. The "No data available" group records an attrition rate of 12.8%, comparable to lower loyalty levels, indicating that this category includes employees with incomplete survey participation or, in some cases, those ineligible for the survey. This pattern indicates that employees who demonstrate higher levels of loyalty are considerably less likely to exit the organisation.

Furthermore, the distribution plot of *Loyalty_Level* (see Annex C, pp. 80-81) reveals that "No data available" represents the largest share of the dataset (approximately 49%), followed

by higher loyalty levels (5 and 4). These findings suggest that fostering loyalty in the organisation may play a critical role in employee retention, a conclusion supported by the feature's importance ranking in the LightGBM model, where it accounted for 5% of total predictive power.

Salary Increase

The stacked bar chart entitled "*Attrition Rate by Salary Increase Band*" (see Annex D.11, pp. 87) illustrates the relationship between annual salary growth and employee attrition. The *Salary_Increase* feature represents the percentage increase in an employee's annual salary during the year under analysis, expressed as a percentage interval. The chart reveals a non-linear association between salary adjustments and attrition rates. Employees receiving minimal or moderate increases (between 1% and 7%) display relatively stable attrition levels, ranging from 2.4% to 4.2%, while those with salary growth above 15% exhibit the highest attrition rate (7.5%). Conversely, employees with salary increases between 9% and 11% show the lowest attrition (1.3%). This pattern indicates that salary adjustments may not directly determine retention outcomes but rather interact with other factors such as promotions, role transitions, and performance-related recognitions, which together influence turnover behaviour.

Additionally, the corresponding distribution plot (see Annex C, pp. 80-81) indicates that most employees receive modest raises below 5%, reflecting a conservative compensation structure. Within the LightGBM model, Salary_Increase contributed 5% of total feature importance, highlighting its moderate yet meaningful role in explaining attrition dynamics.

3.2.2.3.Correlation Analysis

The term *correlation* is defined by Schober et al. (2018) as "a measure of a monotonic association between two variables. A monotonic relationship between two variables is one in which either (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the value of the other variable decreases" (p. 1763).

The Pearson correlation coefficient (r) quantifies the strength and direction of a linear relationship between two continuous numerical variables. It is calculated as the covariance of the two variables divided by the product of their standard deviations, as shown in Equation 3.1:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.1}$$

Where x_i and y_i represent individual observations of variables X and Y, and $\bar{x}$ and $\bar{y}$ are their respective means. The resulting coefficient ranges from –1 to +1, where values close to +1 indicate a strong positive linear relationship, values near –1 indicate a strong negative linear relationship, and values around 0 suggest little or no linear relationship (Schober et al., 2018). This measure was selected for the present study as it effectively captures linear associations between numerical features. Although several variables in the dataset are binary (e.g., *Children, Functional_Promotion, Internal_Mobility*, *Leadership_Role*), Pearson's *r* can be appropriately applied because, for dichotomous variables, it is mathematically equivalent to the point-biserial correlation coefficient, allowing consistent interpretation across the dataset (Field, 2013).

On the other hand, a correlation matrix is a table that displays these correlation coefficients between pairs of variables in a dataset. By visualizing these coefficients in a heatmap, it becomes easier to identify patterns, redundancies, or potential predictors that may inform subsequent modeling (Vasujoseph, 2023).

In this study, the correlation heatmap for the numerical features (Figure 3) highlights several meaningful relationships. Among the strongest observed relationships, *Consumer_Price_Index* and *Annual_Unemployment_Rate* demonstrate a very strong negative correlation (-0.93), reflecting underlying macroeconomic dynamics where lower unemployment is generally linked to periods of higher inflation. Another strong correlation is found between *Career_Opportunity* and promotion-related variables, such as *Functional_Promotion_Interval* (0.65) and *Internal_Mobility* (0.74). These correlations, however, were expected to occur, since *Career_Opportunity* was created during the feature engineering stage based on the combination of both *Internal_Mobility* and *Functional_Promotion.*

Moderate correlations are also observed, such as the association between *Leadership_Role* and *Variable_Compensation* (0.27), indicating that employees in leadership positions are more likely to receive higher levels of variable compensation. Similarly, *Company_Car* is moderately correlated with *Variable_Compensation* (0.45), suggesting that employees entitled to company-provided vehicles also tend to benefit from additional compensation elements, likely reflecting organisational seniority or role importance.

In contrast, most of the remaining features exhibit weak or negligible correlations, particularly with the target variable *Attrition*. Coefficients very close to zero, such as those observed between *Attrition* and features like *Internal_Mobility, Functional_Promotion* or *Change_of_Direct_Supervisor*, indicate the absence of strong linear associations. This low degree of multicollinearity is advantageous for subsequent predictive modeling, as it ensures

that the algorithms can capture complex, non-linear interactions across variables without redundancy (Kuhn & Johnson, 2013).

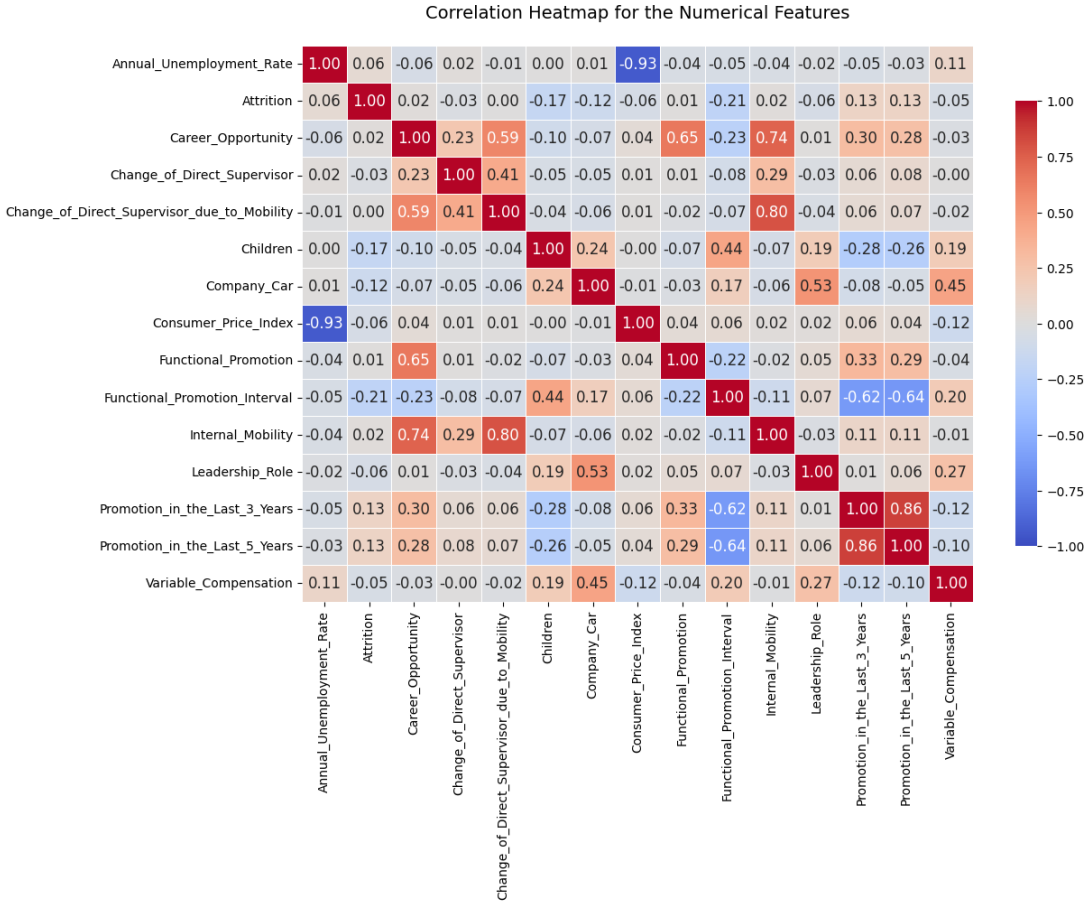


Figure 3 – Correlation Heatmap for the Numerical Features

Following the procedure outlined in Chapter 3.3.4. for converting categorical variables into numerical representations, a comprehensive correlation heatmap was produced that included every feature of the dataset, including variables that were initially numerical and those that were derived from categorical attributes. A threshold of 0.55 was used to filter the results and show only the most significant positive and negative correlations to simplify and improve interpretability. The strongest linear relationships in the dataset are thus highlighted in this filtered heatmap, which is shown in Annex E (pp. 88), while insignificant associations that add little to the analysis are left out.

3.3.Data Preparation

The preparation of the dataset was an essential step to ensure its suitability for predictive modeling algorithms. In this stage, several procedures were conducted to improve data quality, consistency, and analytical value. Importantly, all activities performed during this phase were

idealized and approved in collaboration with the partner company, guaranteeing both methodological alignment and practical relevance.

The process included four main components: (i) feature engineering, involving the creation of additional variables to capture relevant organisational and economic aspects; (ii) data cleaning, to address missing values and remove redundancies; (iii) transformation of non-numerical features into numerical representations suitable for ML algorithms; and (iv) splitting of the dataset into training, validation, and test subsets to enable robust model development and evaluation. Together, these steps provided a solid foundation for the subsequent modeling stage by ensuring that the dataset was accurate, consistent, and ready for analysis.

3.3.1. Feature Selection

The only feature excluded from the predictive modeling process was *Year*, which represents the reference year of analysis (2022–2024). This variable was removed because it does not contain intrinsic predictive information regarding employee attrition, functioning solely as a temporal identifier. Including it could introduce artificial temporal bias without meaningful explanatory value. All other dataset features were retained, as tree-based ensemble models—such as XGBoost, LightGBM, and CatBoost—are capable of performing implicit feature selection. These algorithms iteratively select features that maximize information gain at each node split, thereby prioritizing the most relevant predictors while down-weighting or disregarding those with low predictive contribution (Ke et al., 2017).

3.3.2. Feature Engineering

Feature engineering was used to generate derived variables that may be more informative on employee behaviour and organisational characteristics. The partner company had asked for the design of two new features - *Career_Oportunity* and *Change_of_Direct_Supervisor_due_to_Mobility* - based on the attributes already available in the dataset.

The *Career_Oportunity* feature was designed to indicate whether the employee has had a career opportunity within the organisation, determined by the combination of *Internal Mobility* and *Functional Promotion* attributes. If either internal mobility or a functional promotion had occurred, the feature assumed the value 1 (“Yes”); otherwise, the value was set to 0 (“No”). For instance, in a case where *Internal Mobility* = 1 and *Functional Promotion* = 0, the value of *Career Opportunity* would be 1.

On the other hand, the *Change_of_Direct_Supervisor_due_to_Mobility* feature was designed to indicate whether an employee experienced a supervisor / direct management change, either because of internal mobility processes or due to other organisational reasons. By incorporating this information, it was possible to account for changes in leadership that might influence employee perceptions and decisions. Determined by the combination of the variables *Internal_Mobility* and *Change_of_Direct_Supervisor*, if for instance, *Change_of_Direct_Supervisor* = 1 and *Internal_Mobility* = 0, then *Change_of_Direct_Supervisor_due_to_Mobility* = 0.

3.3.3. Data Cleaning

The data cleaning stage involved a set of adjustments aimed at ensuring consistency and eliminating redundancies in the dataset. First, for the variable *Salary_Increase*, records originally categorized as "No data available" represented only 0.7% of the overall dataset. These values were therefore replaced with the category "<1%", under the assumption - requested by the company - that no actual salary increase had occurred for the employees in question. Second, the variable *Cross-Functional_Project* was excluded from the dataset, since its information was already fully represented by the column *Type_of_Cross-Functional_Project* (see Annex A, pp. 70–78, for further detail). Retaining both variables would have introduced redundancy without adding value to the analysis. Finally, the presence of missing values in survey-based and performance rating features was addressed through data imputation, using the K-Nearest Neighbors (KNN) imputer, as further explained below. This allowed for the estimation and replacement of missing entries while maintaining the integrity of the dataset.

3.3.4. Transformation of Non-Numerical features

Transforming categorical (non-numeric) variables into numeric form is essential because most machine learning algorithms require numerical input to perform mathematical operations, calculate distances, and generate model coefficients. This transformation ensures that the data can be processed correctly by algorithms, enables consistent data representation, and allows for techniques like standardization and normalization. Encoding categorical variables numerically, through methods like label encoding or one-hot encoding, also helps preserve the inherent structure and relationships within the data, improving predictive model performance (Kuhn & Johnson, 2013; Ahsan et al., 2021).

To ensure that the encoded dataset provided by the company meets the specifications of ML algorithms, various transformation methods were utilized, such as label encoding, one-hot encoding, target encoding, and data imputation.

“Label Encoding is a process in machine learning where categorical data, represented as labels or strings, is converted into numerical format. In this encoding technique, each unique category is assigned a unique integer, effectively converting categorical data into numerical values. The main purpose is to prepare the data for machine learning algorithms that require numerical input” (Kumar, 2024, para. 2). In this study, Label Encoding was applied to several categorical attributes, namely *Gender*, *Generation*, *Tenure*, *Job_Level*, *Total_Annual_Compensation*, and *Salary_Increase*. For example, the feature *Tenure* originally consisted of descriptive categories such as “New Joiners”, “Pathfinders”, “Growing Talent”, “Core Contributors”, “Legacy Builders”, and “Experienced Veterans”. Through label encoding, these categories were mapped to integers from 1 to 6, respectively. Thus, an employee categorized as “Pathfinder” would be encoded as 2, while an “Experienced Veteran” would be encoded as 6. This transformation preserved the categorical distinctions while allowing the variables to be processed numerically by the predictive models.

On the other hand, the categorical features *Location*, *Functional_Area*, *Board_Member_Oversight*, *Work_Model*, and *Type_of_Cross_Functional_Project* were transformed using One-Hot Encoding. In this technique, for each category that a feature contains, a new column is added to the dataset. Each of these new columns takes the value 1 if the employee belongs to that category and 0 otherwise (Pramoditha, 2021). For example, the feature *Work_Model* originally contained the categories “Remote”, “Hybrid”, “On Site” and “On Site T”. After one-hot encoding, four new columns were created: 'WM_Remote', 'WM_Hybrid', 'WM_OnSite' and 'WM_OnSiteT'. If an employee’s work model was *Hybrid*, then the column 'WM_Hybrid' would take the value 1, while the columns 'WM_OnSite', 'WM_OnSiteT' and 'WM_Remote' would take the value 0.

In the case of the variables *Department* and *Unit*, the original values were simply numerical identifiers (e.g., Department 1, 2, …, 46; Unit 1, 2, …, 46.4), which did not carry any intrinsic meaning or ordinal relationship. Using these raw codes could mislead the model into assuming a hierarchy where none exists. To overcome this limitation, Target Encoding was applied. With this method, each category is replaced by a numerical value representing the average attrition rate observed within that category. For instance, if Department 5 had an attrition rate of 10%, then all employees belonging to that department were assigned the encoded value 0.10. Similarly, if Unit 12 showed an attrition rate of 16%, all records in that unit were encoded with

0.16. This approach ensures that the encoded values reflect the actual relationship of each department or organisational unit with attrition, while at the same time avoiding the creation of dozens of additional columns, which would have been the case with one-hot encoding. As a result, the data becomes both more informative and more efficient for predictive modeling.

Regarding the dataset's survey-based features, such as *Engagement_Level*, *Enablement_Level, Commitment_Level, Loyalty_Level*, among others, and *Performance_Rating*, these were originally measured on a 1–5 ordinal scale, where higher values reflect e.g., stronger satisfaction, engagement, or performance. Since the scales already represented ordinal information, these variables were stored directly as numerical features without further encoding. However, to address the presence of missing values in these features (see Annex F, pp. 89) data imputation was performed using the K-Nearest Neighbors (KNN) imputer. This method estimates each missing entry by considering employees with similar profiles across multiple features and assigning the average value of their responses. For instance, if an employee's *Engagement_Level* was missing, the imputation algorithm identified the most similar employees based on attributes such as *Tenure*, *Job_Level*, or other satisfaction indicators, and replaced the missing score with the average engagement level of those neighbours. Likewise, if *Performance_Rating* was unavailable, the imputed value was derived from comparable employees, ensuring that the replacement reflected realistic patterns observed in the workforce.

By applying this approach, the integrity of the dataset was preserved while avoiding the loss of valuable records. Compared to simple mean or median substitution, KNN imputation leverages multidimensional similarity, which provides more context-aware and reliable estimates. This was particularly important for variables such as *Loyalty_Level*, where almost half (49.2%) of the values were missing, and where discarding incomplete records would have substantially reduced the sample size.

3.4.Modeling

The Modeling stage of the CRISP-DM framework focuses on developing predictive models using the prepared dataset to identify meaningful patterns and generate reliable predictions of employee attrition. Building upon the data preparation phase, this stage transforms the processed data into tangible predictive tools (Kuhn & Johnson, 2013).

This chapter is structured as follows: first, the dataset partitioning strategy is described, detailing how the data were divided into training, validation, and testing subsets. Then, the

approach to handling class imbalance is presented, followed by the selection of appropriate modeling techniques and the experimentation process, which involved hyperparameter optimization and threshold-tuning strategies to identify the best-performing model configuration. The explanation of the metrics used to assess the performance of these models is explained in the chapter 3.5. Evaluation.

3.4.1. Dataset Partitioning

The dataset was divided into three distinct subsets: training, validation, and test sets. This division is critical for effective model training, appropriate hyperparameter tuning, and robust evaluation on previously unseen data to assess generalization performance (Kuhn & Johnson, 2013). This structure also mitigates *overfitting*, a common issue in predictive modeling where algorithms learn noise or random fluctuations from the training data instead of meaningful patterns, thus reducing their ability to generalize to new cases (Krishna & Borah, 2024). In this way, the model's predictive capacity can be more accurately evaluated based on its ability to generate correct predictions on unseen data.

The full dataset contained 5728 records, of which approximately 8% corresponded to employees who left the company (*Attrition* = 1) and 92% to employees who stayed (*Attrition* = 0). Given this imbalance, the splits were performed using a stratified sampling strategy, which preserved the original class distribution in all subsets. This was critical to avoid biasing the model training and evaluation towards the majority class (Kuhn & Johnson, 2013).

Following best practices in predictive modeling, a 70/15/15 split ratio was applied to the dataset (Kuhn & Johnson, 2013). The training set, comprising 70% of the data (4008 records) was used to train the models and capture patterns from the data. The validation set, representing 15% of the dataset (860 records), was used for hyperparameter tuning, model selection, and threshold adjustments, enabling a fair comparison across different algorithms. Finally, the test set — also 15% of the dataset (860 records) was kept aside and used exclusively for final model evaluation, providing an unbiased estimate of predictive performance. This structured partitioning ensured that all three subsets retained the same proportion of attrition cases (~8%), thus maintaining class balance and guaranteeing consistent evaluation across the modeling pipeline.

3.4.2. Dataset Balancing

As previously mentioned in Chapter 3.1., the target variable *Attrition* presents a significant class imbalance, with the proportion of employees who left the organisation (*Attrition = 1*) being considerably lower than those who remained (*Attrition = 0*). Specifically, only 8.1% of employees were recorded as leavers, while 91.9% were retained, highlighting the dominance of the majority class (see Figure 4).

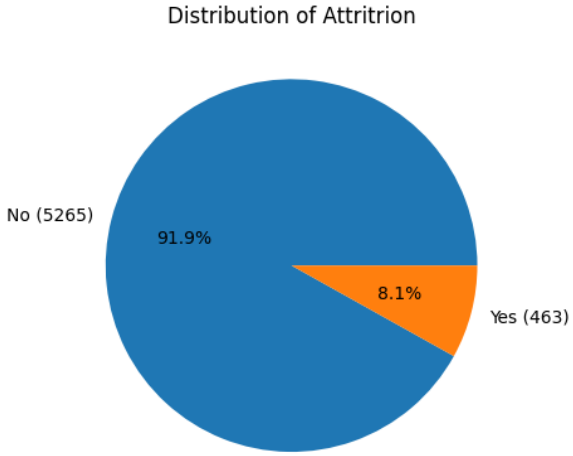


Figure 4 - Class Distribution of the Target Variable (Attrition, 2022–2024)

This imbalance, commonly referred to as target bias, can lead to a situation where predictive models become disproportionately influenced by the majority class (that is, employees who remained with the company). As a result, the model may achieve high overall accuracy while performing poorly in detecting minority-class cases — in this context, employees who are likely to leave the company.

To mitigate such effects, different dataset balancing strategies may be adopted, including resampling techniques or algorithmic weighting adjustments. Resampling techniques aim to rebalance class distributions by either increasing the representation of the minority class (*oversampling*) or reducing the dominance of the majority class (*undersampling*), thereby mitigating class bias. Oversampling generates or duplicates minority-class instances to improve class parity, while undersampling removes majority-class instances to achieve a more balanced dataset. These techniques, however, carry risks: oversampling may lead to overfitting and distort the underlying distribution, whereas undersampling may discard valuable information and weaken the model's generalization capability (Jafarigol & Trafalis, 2023).

Class weighting, by contrast, offers a principled algorithm-level solution that preserves the original dataset. This approach assigns higher weights to the minority class within the model's loss function, penalizing its misclassification more heavily and thereby emphasizing

underrepresented instances during training. Unlike resampling, class-weighting does not modify the data itself but rather adjusts how the model updates its internal parameters. Machine-learning frameworks such as scikit-learn, XGBoost, and LightGBM support this approach via parameters such as *class_weights*, *scale_pos_weight*, and *is_unbalance*, enabling internal rebalancing of the optimization process without synthetic or discarded samples (Jafarigol & Trafalis, 2023)

In this study, explicit resampling techniques, such as oversampling or undersampling, were deliberately avoided, as they may introduce artificial data or discard valuable information from the majority class. Instead, a class-weighting strategy was adopted, allowing the model to account for the imbalance internally during training. This method preserves the integrity of the original dataset, avoids the drawbacks of synthetic or discarded samples, and provides a principled way of addressing class imbalance directly during model optimization.

3.4.3. Model Selection

The classification problem addressed in this study is a binary classification problem, where the target variable indicates whether an employee leaves the company (*Attrition* = 1) or remains employed (*Attrition* = 0). In machine learning, such tasks are typically approached using supervised learning, where algorithms are trained on historical labelled data and learn to map input features (e.g., gender, department, tenure, promotions, compensation, etc.) to a binary outcome. This differs from unsupervised learning, which identifies latent structures in data without predefined labels. By framing attrition prediction as a supervised classification problem, labelled HR data are employed to train and evaluate models that can generalize to unseen cases (Kuhn & Johnson, 2013).

The objective of the model selection process is to identify the algorithms most suitable for predicting employee attrition while ensuring a balance between predictive performance, interpretability, and computational efficiency. To achieve this, a set of ML classifiers was selected, representing both single-model and ensemble-based approaches. The following algorithms were included in the experimentation: Decision Tree Classifier, Bagging Classifier, Random Forest Classifier, Gradient Boosting Classifier, XGBoost Classifier, CatBoost Classifier, and LightGBM Classifier. These models were chosen due to their strong performance in structured, tabular data contexts, their ability to handle non-linear relationships, and their proven applicability in HR analytics and attrition modeling (Yahia et al., 2021). In addition, tree-based models are particularly valued for their interpretability and explainability,

as they provide transparent decision structures that facilitate understanding of the factors influencing employee attrition.

The ensemble models (such as Random Forest, Gradient Boosting, XGBoost, CatBoost, and LightGBM) were included for their robustness against overfitting, high predictive power, and ability to capture complex interactions among predictors through the aggregation of multiple weak learners. In contrast, single-tree algorithms (e.g., Decision Tree Classifier) were retained for their interpretability and explainability, providing transparent decision rules that are particularly relevant in human resource management contexts where decision accountability is critical (El-Rayes et al., 2020).

All models were trained on the training subset (70% of the dataset) and tuned on the validation subset (15%) for hyperparameter optimization and model selection. The remaining 15% of the data was reserved for final testing of the selected model (LightGBM), to ensure an unbiased assessment of predictive performance. A systematic approach was used to identify the optimal model configuration through a series of experiments conducted across all classifiers (see Annex G, pp. 90). Each experiment combined various hyperparameter optimization and decision threshold tuning techniques. The remainder of this chapter provides a detailed explanation of the optimization process applied to the final selected model, LightGBM.

The Light Gradient Boosting Machine (LightGBM) is a gradient-boosting framework developed by Microsoft, designed to efficiently model complex, non-linear relationships in structured datasets. It builds an ensemble of decision trees sequentially, where each new tree focuses on correcting the errors of the previous ones. In contrast to level-wise tree growth, LightGBM adopts a leaf-wise growth strategy, which can improve predictive performance but requires careful regularization to avoid overfitting (Ke et al., 2017). The optimization process for the LightGBM model included three main phases: (i) broad hyperparameter exploration, (ii) fine-grained hyperparameter refinement, and (iii) decision threshold tuning.

i. The process began with the application of RandomizedSearchCV, which allowed the model to efficiently explore a wide hyperparameter space by sampling combinations at random, combined with stratified cross-validation to ensure stable performance assessment. This enabled the identification of promising regions within the parameter space without incurring the computational cost of an exhaustive search. The best-performing configuration from this stage served as the reference point for subsequent tuning.

ii. Based on the best hyperparameters identified by RandomizedSearchCV, GridSearchCV was then applied to the LightGBM Classifier, to conduct a more

focused and exhaustive hyperparameter search (Raj, 2023). This process refined the model parameters with greater precision, enabling LightGBM to achieve a more optimal balance between complexity and generalization. During this phase, the classifier was evaluated using the default decision threshold of 0.50, meaning that predicted probabilities for the target class (*Attrition)* greater than or equal to 0.50 were classified as *Attrition* = 1.

iii. Although hyperparameter tuning improved predictive performance, the default probability threshold of 0.50 does not necessarily maximize the F1 score, particularly in contexts with class imbalance such as this dataset. To address this, the best LightGBM model from the previous GridSearchCV was wrapped within a TunedThresholdClassifierCV, which performs cross-validated evaluation across a range of probability thresholds to identify the cutoff that maximizes F1. This procedure adjusts only the classification threshold and does not modify the underlying model parameters.

The comparative analysis of the algorithms' performance will be presented in Chapter IV – Results Presentation and Discussion, where each model is evaluated, and the most suitable algorithm for employee attrition prediction is identified.

3.5.Evaluation

For model evaluation, the F1 score on the validation dataset was selected as the primary performance metric for comparing models. This decision was made based on both methodological considerations and feedback from the organisation, which emphasized the importance of identifying employees at higher risk of attrition without excessively increasing false positives. The F1 score balances these two aspects by combining *precision* (the accuracy of positive predictions) and *recall* (the ability to detect actual attrition cases), making it particularly suitable for imbalanced datasets where the minority class, in this case, employees who left the company, is of greater analytical and managerial interest (Kuhn & Johnson, 2013).

The F1 score is mathematically defined as (Equation 3.2):

$$F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{3.2}$$

Where:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

Here, *TP* represents true positives, *FP* false positives, and *FN* false negatives. Precision measures the proportion of correctly predicted attrition cases among all predicted attrition cases, while recall indicates the proportion of actual attrition cases correctly identified by the model. By combining these two measures, the F1 score provides a single, interpretable value that penalizes extreme imbalances between precision and recall.

To complement the F1 score, the AUC–ROC (*Area Under the Receiver Operating Characteristic Curve*) metric was also employed after the final model selection, using the test dataset. This allowed for a threshold-independent assessment of the final selected model's ability to discriminate between attrition and non-attrition cases. The AUC–ROC is a widely used performance metric for evaluating binary classification models, as it measures their ability to discriminate between positive and negative outcomes. According to Krishna & Borah (2024), the AUC represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative one, providing a threshold-independent assessment of predictive performance.

The ROC curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels. These rates are mathematically defined as:

$$TPR = \frac{TP}{TP + FN} \tag{3.5}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.6}$$

Each point on the ROC curve corresponds to a specific threshold combination of these two rates. The Area Under the Curve (AUC) quantifies the model's overall ability to distinguish between the positive and negative classes and is computed as the integral of the ROC curve:

$$AUC = \int_0^1 TPR(FPR)\ d(FPR) \tag{3.7}$$

An AUC value of 1.0 indicates perfect discrimination, while 0.5 suggests no discriminative power, equivalent to random guessing (Krishna & Borah, 2024).

Furthermore, a confusion matrix was generated for the final selected model using the test dataset to visualize the distribution of predictions across true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This provided a detailed understanding of the model's predictive behaviour and potential trade-offs. As noted by Khera & Divya (2019),

a confusion matrix offers an effective means of assessing classification performance by comparing predicted outcomes with actual labels, thereby revealing both the accuracy and the types of misclassifications made by the model.

The confusion matrix (Table 2) provides a tabular representation of prediction outcomes, summarizing the counts of TP, TN, FP, and FN:

Table 2 - Representation of a Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

From this matrix, several relevant performance metrics such as Precision, Recall, Specificity, and the F1 score can be derived, providing a comprehensive view of the model's predictive capability and potential trade-offs between sensitivity and specificity. While recall (or *sensitivity*) focuses on correctly identifying employees who left, *specificity* focuses on correctly identifying those who stayed, making it an important measure of reliability in imbalanced datasets where the majority class dominates. Formally, specificity is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.8}$$

In contrast, accuracy was considered unsuitable for this study due to the dataset's class imbalance. Since approximately 92% of employees remained in the company and only about 8% left, a model that predicts all employees as non-attrition cases would still achieve 92% accuracy while failing to detect any actual leavers. Consequently, accuracy alone does not provide a meaningful assessment of predictive performance in this context.

Finally, the comparative performance results and model evaluation outcomes are presented in Chapter IV – Results Presentation and Discussion, where the implications of these metrics for employee attrition prediction are discussed in detail.

3.6.Deployment and Model Management

Within the CRISP-DM framework, the Deployment and Model Management stages address the operationalization of the analytical solution and its continuous monitoring after implementation. However, within the scope of this research project, these two final stages are not implemented. Instead, the study focuses on the design, development, and evaluation of the predictive model. Additionally, it prioritizes the transfer of knowledge to the company

regarding the model's insights to maintain continuous alignment with the organisational context. Therefore, only the intended deployment context and organisational considerations provided by the company are described below, rather than a comprehensive implementation plan.

With respect to Deployment, the organisation indicated that the predictive attrition model developed would function as a standalone analytical asset that can also be integrated into existing internal analytical use cases related to workforce planning. According to the information provided, the model would require the involvement of one Data Scientist and one HR Analyst. The HR Analyst would be responsible for extracting employee data from the Human Resource Management System in Excel format and applying their understanding of the variables that characterize employees and their career trajectories to provide strategic insights aligned with organisational objectives. The Data Scientist, in turn, would load and process the data in the company's JupyterLab environment hosted via Google Cloud Storage. When new data become available, the Data Scientist would repeat the required data preparation steps, following the workflow previously described in Chapter 3.3, including feature selection and engineering, data cleaning, and transformation to numerical features. The JupyterLab notebook developed for this project contains all required preprocessing and transformation procedures, enabling efficient preparation of the data prior to running the model.

The model’s execution would generate an Excel file containing the predicted attrition outcomes (coded as 1 for attrition and 0 for retention) along with the associated probabilities for each employee. This output could then be used by the HR department to identify employees at high risk of leaving the company, support proactive retention strategies, and monitor changes in predicted attrition risk over time.

The model is intended to be executed on an annual basis, particularly before relevant workforce events such as promotion cycles or salary review periods, and may also be executed during exceptional circumstances, such as organisational restructuring or periods of unexpectedly high voluntary turnover. Currently, integration with the company’s Human Resource Management System via API is not planned. Instead, the employee dataset will continue to be manually extracted and processed in the JupyterLab notebook, similarly to the workflow adopted in this research. This is particularly relevant because several features in the dataset used for this research were encoded for confidentiality, implying that, in practice, the organisation will need to reproduce the original full data preparation process using raw operational data.

Additionally, it is important to note that the evaluation presented in this study is based on predictions generated on a hold-out test dataset, where the actual attrition outcomes were already known. While this approach is standard for estimating generalization performance, it still reflects past employee behaviour and organisational conditions. In practical applications, the model will forecast future attrition cases before the outcomes occur, under potentially changing business environments, workforce dynamics, and policy conditions. As a result, the model's performance may vary over time, requiring periodic retraining and monitoring to maintain predictive reliability. In deployment, the model will be applied to new, unseen employee data, generating predicted probabilities and risk labels without immediate access to evaluation metrics, which require actual outcomes. Real-world performance assessment will involve comparing predictions with observed attrition events over time, enabling continuous monitoring, validation, and recalibration of the model as business conditions and employee behaviours evolve.

The following outlines a proposed approach for model management. To sustain model accuracy and relevance, the organisation should establish a structured monitoring and recalibration pipeline post-deployment. The model should be executed on the most recent employee dataset at regular intervals (in this case, annually), with predicted probabilities and risk categories securely recorded along with relevant metadata. After a designated observation period, the dataset should be updated to reflect actual attrition outcomes, enabling comparison between predicted risks and observed results and facilitating the calculation of performance metrics such as F1 score, precision, recall, and AUC. Tracking these metrics over time will support systematic performance evaluation.

Beyond predictive performance, the organisation should monitor data drift—changes in the statistical distribution of input variables or model outputs. Comparing new data with the original training distribution will reveal when model assumptions may no longer hold. Retraining triggers should be clearly defined, such as significant performance degradation or detection of drift in key features, with retraining scheduled periodically or as needed. Quantitative monitoring should be supplemented by a business feedback loop, allowing HR practitioners to assess the interpretability and relevance of predictions.

Chapter IV - Results Presentation and Discussion

This chapter provides a comprehensive analysis of the results obtained from the predictive modeling stage, in which several machine learning classifiers were used to forecast employee attrition. It begins by comparing the performance of multiple classification algorithms, emphasizing their strengths, weaknesses, and generalization capabilities. The discussion then focuses on the evaluation of the final selected model, including detailed interpretation of its performance metrics, feature importance, and predictive reliability. In addition, the chapter explores the business implications of the prediction outcomes and presents representative employee risk profiles. Finally, it concludes with a critical reflection on the methodological limitations and issues encountered throughout the study, ensuring a robust interpretation of the results.

4.1.Comparative Results Across Models

Table 3 presents the best F1 scores for each classifier on the validation and test datasets, derived from the various experiments conducted during the model selection phase. The models are listed from highest to lowest validation F1 score since this metric served as the primary criterion for comparing predictive performance. These results provide a comparative assessment of the models' predictive performance and generalization capacity. The scores shown correspond to the optimal configurations identified through hyperparameter and threshold-tuning procedures. Overall, the results reveal distinct variations across models, reflecting differences in complexity, learning strategy, and sensitivity to the dataset's imbalance.

Table 3 - Comparative F1 scores by Classification Model (Validation vs. Train)

Classification Model	**F1 score (Validation Set)**	**F1 score (Train Set)**	**Generalization Gap (Train − Validation)**
LGBMClassifier	0.657	0.823	0.166
XGBClassifier	0.643	1.000	0.357
CatBoostClassifier	0.628	0.829	0.201
GradientBoostingClassifier	0.621	0.989	0.368
BaggingClassifier	0.574	0.874	0.300
DecisionTreeClassifier	0.564	0.571	0.007
RandomForestClassifier	0.537	0.612	0.075

Among all models, the LGBMClassifier achieved the highest F1 score on the validation dataset (0.657), demonstrating strong predictive capacity and a moderate degree of overfitting (generalization gap = 0.166). The XGBClassifier and CatBoostClassifier followed, with validation scores of 0.643 and 0.628, respectively. Although these models achieved higher training F1 scores, the large gap between training and validation results (0.357 and 0.201) indicates reduced generalization and increased overfitting.

The GradientBoostingClassifier also demonstrated strong predictive performance (validation F1 = 0.621), but exhibited the highest generalization gap (0.368), suggesting significant overfitting and reduced reliability when applied to unseen data. A similar pattern was observed with the BaggingClassifier, which achieved reasonable validation performance (0.574) but also showed high overfitting (gap = 0.300), reflecting its sensitivity to variance in the training data.

In contrast, the DecisionTreeClassifier exhibited the most stable performance, with nearly identical F1 scores for training and validation (0.571 and 0.564), reflecting a simpler structure with limited variance but lower predictive strength. The RandomForestClassifier also demonstrated a relatively small generalization gap (0.075), suggesting good consistency though at the cost of lower predictive power. Overall, ensemble-based models such as LightGBM, CatBoost, and XGBoost revealed superior performance on the validation dataset, confirming their capacity to capture complex patterns in the data. However, their high training performance indicates a risk of overfitting.

Considering the trade-off between predictive power and generalization capacity, LightGBM was selected as the final model to generate attrition predictions on the test dataset, as it offered the most favourable balance among the tested classifiers. This model selection is also supported by recent empirical evidence in the field, as gradient-boosted decision tree models, particularly LightGBM, CatBoost, and XGBoost, continue to dominate predictive modeling competitions involving tabular data, consistently delivering state-of-the-art performance (ML Contests, 2024).

The following analysis examines the model's performance in greater detail, including the comparison of F1 scores across training, validation, and test sets, as well as the interpretation of the confusion matrix, the AUC-ROC curve, and the most influential features contributing to prediction outcomes.

4.2.Evaluation of the Selected Model's Performance

Before evaluating its predictive performance, it is important to recall that the LightGBM classifier was optimized through a two-stage hyperparameter tuning process combining RandomizedSearchCV and GridSearchCV. The final parameter configuration was as follows: *learning_rate* = 0.045, *max_depth* = 4, *num_leaves* = 16, *min_child_samples* = 25, *min_split_gain* = 0.55, *n_estimators* = 1000, *subsample* = 0.6, *colsample_bytree* = 0.9, *reg_alpha* = 0.5, *reg_lambda* = 1, and *scale_pos_weight* = 3.5. This optimized configuration achieved an average F1 score of approximately 0.53 during cross-validation and was used as the baseline model for subsequent threshold tuning and final testing.

The LightGBM model produces a probability score between 0 to 1, indicating the likelihood that an employee will voluntarily leave the organisation (*Attrition* = 1). To convert these probabilities into a binary classification (high risk vs. low risk), a decision threshold is required. While the default threshold of 0.50 is commonly used, it is not always optimal for imbalanced datasets such as this, where only approximately 8% of employees left the company. Therefore, as previously explained in chapter 3.4.3. Model Selection, the LightGBM classifier was subjected to an additional threshold-tuning process to optimize the decision cut-off for classifying attrition cases. This process identified an optimal threshold of approximately 0.515, which is slightly higher than the default value. The selected threshold maximizes the F1 score by achieving a more effective balance between precision (i.e., the correct identification of true high-risk employees) and recall (i.e., capturing as many potential leavers as possible). Employees with predicted probabilities of 51% or greater are classified as high attrition risk, while those with lower probabilities are classified as low risk.

The performance of the LightGBM model was further evaluated on the independent test dataset, which was withheld from all stages of training and hyperparameter tuning to provide an unbiased assessment of generalization. The model achieved an F1 score of 0.578 on the test set, compared to 0.657 on the validation set, representing a performance decrease of approximately 0.079. This reduction is expected, as test performance typically reflects real-world predictive conditions more closely. Importantly, the relatively small drop between validation and test F1 scores indicates that the model retains much of its predictive ability when applied to unseen data, demonstrating robust generalization.

To complement the F1 score evaluation, the confusion matrix for the LightGBM Classifier on the Test Dataset was also generated (Figure 5). The LightGBM model correctly identified 769 employees who stayed (True Negatives) and 37 employees who left (True Positives), while

misclassifying 21 employees who stayed as leavers (False Positives) and 33 employees who left as stayers (False Negatives). Based on these counts, the model achieved a precision of 0.637, indicating that when the model predicts high attrition risk, this prediction is correct approximately 64% of the time. The recall (sensitivity) for the attrition class was 0.529, meaning the model identified about 53% of actual leavers. In contrast, the specificity was considerably higher at 0.973, demonstrating that the model correctly recognised 97% of employees who stayed.

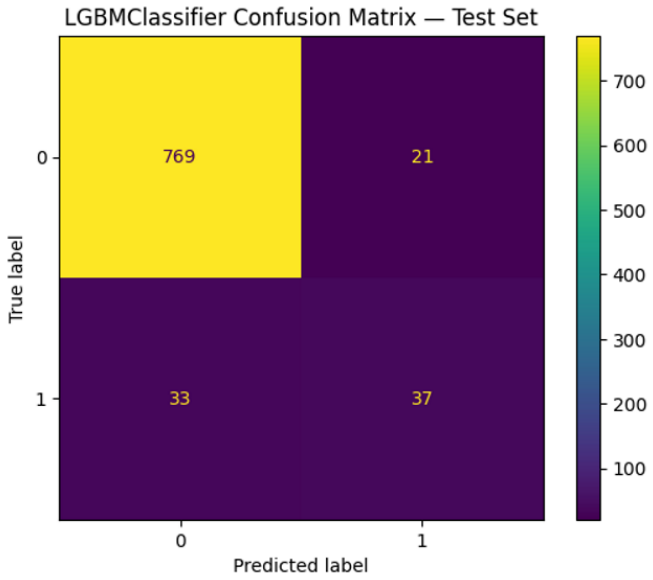


Figure 5 - Confusion Matrix of the LightGBM Classifier on the Test Dataset

These results indicate that the model is highly reliable at identifying employees who will remain in the company (strong specificity and low false-positive rate) but has moderate effectiveness in detecting those who will leave (moderate recall and a noticeable number of false negatives). This behaviour aligns with expectations for class-imbalanced datasets, where the attrition class represents a small proportion of the population. For practical HR decision-making, this means the model is helpful in flagging a prioritized set of high-risk employees, though additional monitoring or supplementary tools may be beneficial to further reduce missed attrition cases.

The Receiver Operating Characteristic (ROC) curve presented in Figure 6 demonstrates the model's capacity to discriminate between both classes, i.e., employees who voluntarily leave the organisation and those who remain. The LightGBM classifier achieves an Area Under the Curve (AUC) of 0.92, which indicates a high level of separability between the positive (*Attrition* = 1) and negative (*Attrition* = 0) classes. An AUC value approaching 1.0 reflects strong

discriminative performance, as the model is able to correctly assign higher predicted probabilities to true attrition cases than to non-attrition cases across a range of classification thresholds. The curve's consistent distance above the diagonal reference line (which represents random classification) confirms that the model performs substantially better than chance. This result provides further evidence of the model's robustness and generalization capability when applied to previously unseen data.

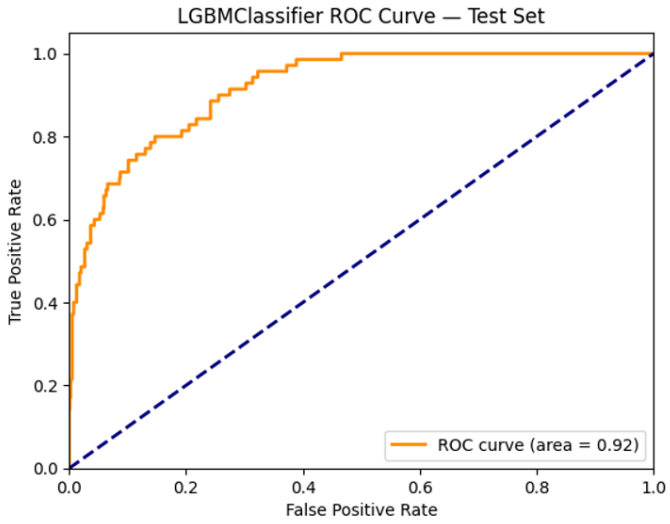


Figure 6 - ROC Curve of the LightGBM Classifier on the Test Dataset

4.3.Feature Importance Analysis of the Selected Model

Figure 7 presents the top 20 predictors ordered by their relative contribution to the LightGBM model. The feature *Functional_Promotion_Interval* emerges as the most influential variable, accounting for approximately 24% of the model's importance. The length of time (in months) since an employee's last functional promotion substantially distinguishes between those who remain and those who leave, indicating that delays or stagnation in career progression may be a key driver of voluntary attrition.

The subsequent most impactful predictors are *Variable_Compensation* (10%) and *Unit* (9%). The prominence of *Variable_Compensation* highlights the influence of pay variability and financial incentives on employee retention. The significance of *Unit* suggests that attrition risk is not uniform across organisational divisions, potentially reflecting heterogeneity in work culture, managerial practices, or workload demands. Additional notable features include *Loyalty_Level* and *Salary_Increase* (both approximately 5%), as well as *Tenure* (approximately 4%), demonstrating that both historical employment experience and recent compensation adjustments contribute to attrition risk.

Several features related to employee satisfaction and perception of the work environment, including satisfaction with development opportunities, performance management, leadership trust, and autonomy, have lower individual importance values (approximately 2% each). Although their individual influence is smaller, their presence among the top predictors reinforces the notion that employee experience and managerial support collectively play a meaningful role in shaping retention.

Collectively, the observed pattern of feature importance suggests that career progression, compensation, and organisational context are the primary determinants of voluntary attrition, while satisfaction-related variables offer secondary yet still pertinent explanatory value. These findings provide actionable insight for HR management within the company. Interventions focused on improving promotion transparency, equitable compensation, and supportive departmental environments may be particularly effective in reducing employee departures.

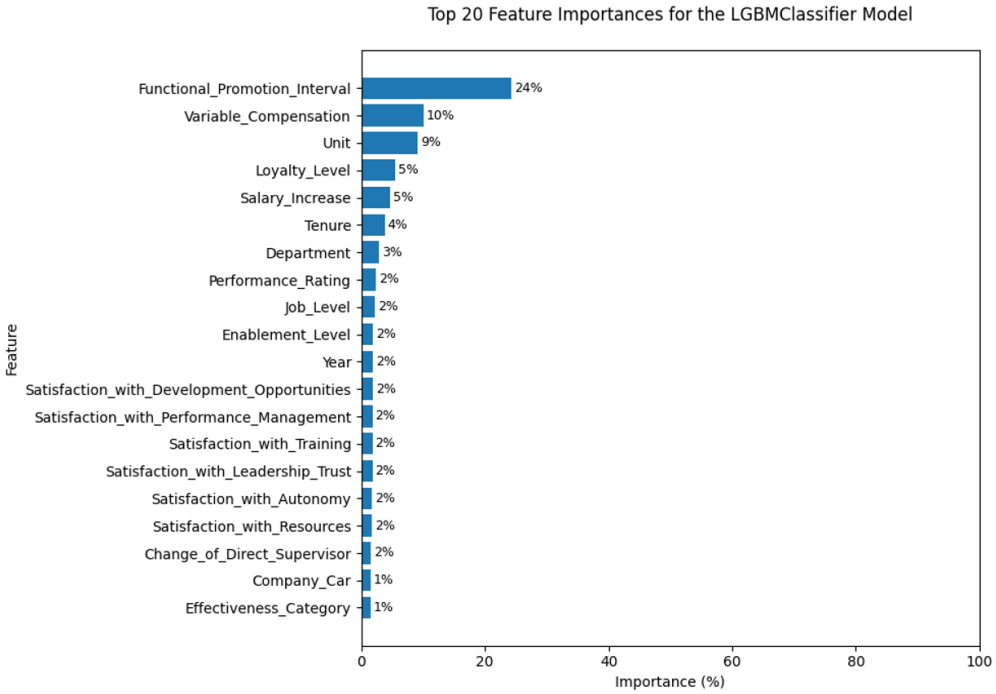


Figure 7 - Top 20 Feature Importances Identified by the LightGBM Classifier Model

4.4.Predicted Attrition Analysis

The distribution of predicted attrition outcomes on the test dataset indicates that the model classified approximately 6.7% of employees (58 out of 860) as high attrition risk, while the remaining 93.3% (802 employees) were classified as likely to stay. The narrow margin of 1.4 percentage points indicates that the model is slightly conservative, tending to under-identify potential leavers rather than overestimate attrition. This distribution aligns with the dataset's

class imbalance, where about 8% of employees had historically left the organisation. The consistency between observed and predicted class proportions demonstrates that the model does not artificially inflate attrition cases and instead preserves the underlying structure of the data, supporting its calibration and reliability when applied to unseen records.

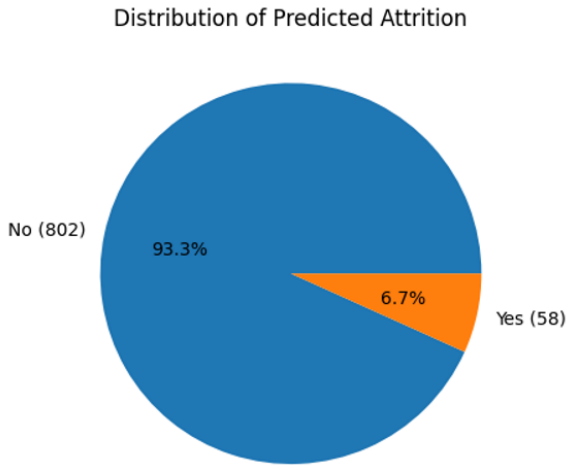


Figure 8 - Distribution of Predicted Attrition Rate (Test Set)

4.4.1. Bivariate Analysis

Complementarily to the overall prediction distribution analysis, a bivariate analysis was conducted using the model's predicted attrition results to examine patterns across the most influential features identified for the LightGBM model.

The predicted attrition rates by functional promotion interval (see Annex H.1, pp. 91) indicate that employees who are approximately 13–36 months past their most recent functional promotion face the highest predicted risk of departure, reaching up to 20%. This period may reflect unmet career progression expectations and an increased likelihood of attrition. In contrast, employees who were recently promoted (0–12 months) or those with longer promotion intervals (61+ months) exhibit substantially lower predicted attrition rates (below 8%), indicating greater short-term stability and long-term retention, respectively. Additionally, the distribution of this feature in the test dataset (see Annex I.1, pp. 94) shows that a large proportion of employees fall within the 61–120-month interval, while relatively fewer are concentrated in the 13–36-month range. This pattern aligns with the observed historical attrition trends (see Annex D.7, pp. 85), demonstrating that the model effectively captures the relationship between promotion timing and attrition risk and underscores the significance of promotion cadence in employee retention.

Regarding variable compensation (see Annex H.2, pp. 91), the model predicts that employees receiving moderate incentives (11–15%) have the highest predicted attrition rate (10.6%), consistent with the historical trend (see Annex D.8, pp. 85), where the same range shows 11% of attrition. Smaller increases in predicted attrition (approximately 5%) are also observed in the 6–10% and 16–20% incentive ranges. In contrast, the model predicts 0% attrition in the very low (0–5%) and very high (21% or higher) incentive ranges. This absence of predicted attrition at the extremes should not be interpreted as absence of attrition risk, but rather as a reflection of the limited number of employees in these groups (see Annex I.2, pp. 94): because they contain few observations, there is insufficient evidence to push any individual prediction above that cutoff. As a result, the model defaults to the majority outcome ("no attrition") in these sparsely populated ranges. Overall, this pattern aligns with the historical attrition distribution, reinforcing the role of perceived reward adequacy and incentive expectations in shaping employees' intentions to depart.

Model predictions by organisational unit (see Annex H.3, pp. 92) reveal that Units 10.3 and 23.5 exhibit the highest predicted attrition rates (approximately 11.8% and 9.5%), respectively. Several large units, such as Units 34.6, 41.3, and 23.4, show 0% predicted attrition. Comparison with observed patterns (see Annex D.9, pp. 86) shows that Unit 10.3 and the 'Other' grouping exhibit similarly high actual attrition rates, around 8–9%. The model underestimates attrition in Units 19.2 and 7.1, which have the highest actual rates, approximately 20.6% and 14.5%, but receive low or zero predicted risk. The test-set Unit distribution (see Annex I.3, pp. 94) is highly uneven, with a few units containing much larger headcounts. This imbalance limits the available evidence for some units and contributes to conservative, zero predictions in sparsely represented units.

With respect to loyalty level (see Annex H.4, pp. 92), the results suggest a non-linear risk profile: predicted attrition peaks at Level 3 (11.6%), is moderate at Level 4 (6.1%), lower at Levels 2 and 1 (4.6% and 2.9%), and effectively zero at Level 5. Considering the test-set distribution (see Annex I.4, pp. 95), most employees sit at Level 3 (22%), with Level 1 representing only 3% of cases, so the model has little evidence in the lowest-loyalty group and, with a 0.515 cutoff, tends to default to "stay", which helps explain the subdued prediction at Level 1. Compared with the historical pattern (see Annex D.10, pp. 86), where attrition declines monotonically with higher loyalty and is highest at Level 1 (16.7%) and lowest at Level 5 (0.4%), the model captures the low risk at Level 5 but underestimates risk at Level 1 and overstates it at Level 3. Ultimately, loyalty remains a strong inverse indicator of attrition risk,

though the model slightly overestimates risk at medium loyalty levels and underestimates it at the lowest level due to data imbalance.

Finally, the model's predictions for attrition by salary increase (see Annex H.5, pp. 93) show that employees receiving minimal raises (below 1%) exhibit the highest predicted attrition rate at around 14.5%, while moderate increases between 5% and 11% are associated with modest risk levels (approximately 4–6%). Conversely, employees in the 11–15% band show no predicted attrition, and those with raises above 15% present a smaller predicted risk (7.7%). This pattern aligns with the historical attrition distribution (see Annex D.11, pp. 87), where attrition is highest at both very low and very high salary increase levels, suggesting that insufficient or unusually large salary adjustments may signal underlying dissatisfaction or role instability. The observed distribution of salary increases in the test dataset (see Annex I.5, pp. 95) indicates that most employees received modest raises (below 5%), supporting the model's emphasis on this group as the primary driver of predicted attrition risk.

4.4.2. Employee Attrition Risk Profiles

The LightGBM classifier, being an ensemble of decision trees, combines the predictive contributions of many individual trees to produce robust overall performance. However, it also allows for the isolation and analysis of specific trees or nodes that have the greatest impact on model predictions. This subchapter presents the employee profiles extracted from the most impactful decision tree of the selected classifier model, which was selected based on the highest F1 score achieved during model optimization.

After the model was tested, the internal tree structures were examined to identify the single tree contributing most strongly to the model's predictive performance. Each terminal leaf of that tree represents a distinct group of employees (i.e., an employee *profile*) who share similar characteristics and corresponding predicted attrition risk. By translating the tree's decision rules into interpretable profiles, it becomes possible to link the model's internal logic with meaningful, human-readable insights about employee behaviour.

To facilitate interpretation, a summarized table (see Annex J, pp. 96) was constructed to present the five most representative employee profiles derived from the test dataset, each corresponding to one leaf of the best LightGBM tree. The table includes the following columns:

i. *Profile ID*: The numerical identifier assigned to each leaf or employee segment.
ii. *Support in Test*: The number of employees in the test dataset who belong to that profile.

iii. *Positives*: The number of employees who actually left the company in the corresponding profile, as observed in the test data.

iv. *Observed Attrition Rate (%)*: The proportion of employees who left the company within that profile, calculated empirically.

v. *Risk vs Overall (%)*: A relative measure comparing the profile's attrition rate to the overall test-set attrition rate (8.1%). Values above 1 indicate elevated risk.

vi. *Average Predicted Risk (%)*: The mean attrition probability predicted by the model for employees within that profile.

vii. *Profile Rules*: The set of decision rules (based on feature thresholds) that define the characteristics of the employees belonging to that profile.

While these profiles (see Annex J, pp. 96) were derived from the best-performing tree, it is important to note that the specific features appearing in the profile rules do not necessarily coincide with those ranked highest in the global feature importance analysis (see Figure 7). This difference arises because the global importance chart aggregates the contribution of each feature across all trees in the ensemble, reflecting their overall predictive influence, whereas the rules extracted from a single tree represent the local decision paths most relevant to that specific subset of employees. Consequently, while global feature importances provide a comprehensive view of the model's overall behaviour, the tree-level profiles offer more interpretable, case-specific insights into the main drivers of attrition risk for distinct employee groups.

Profile 3 represents a small but highly critical subset of employees, exhibiting a 57.6% observed attrition rate, which is more than seven times the overall test-set average. Employees in this profile have remained at their current functional level for up to 106 months (approximately 8.8 years) between promotions, report high enablement scores (4 out of 5, indicating a sense of capability and sufficient resources), yet did not receive a salary increase in the respective year of observation. This combination suggests a progression–reward mismatch: although these employees feel capable, prolonged stagnation in both career advancement and compensation may indicate blocked opportunities or unmet expectations. Existing literature and organisational practice associate this pattern with increased voluntary turnover risk, primarily due to perceived lack of recognition or advancement opportunities rather than deficiencies in capability.

Profile 9 demonstrates a moderately elevated attrition rate of 23.5%, approximately 2.9 times higher than the overall rate. Employees in this profile experience extended intervals between functional promotions, exceeding 106 months, and have not participated in any cross-functional projects (*CFP_Not_Applicable* = 1). This profile represents employees with

prolonged tenure in their functional roles and minimal exposure to cross-departmental collaboration, such as Data and AI Transformation, Digital Transformation, Operational Transformation, or other strategic initiatives. This isolation from broader organisational initiatives may diminish engagement and perceived visibility, especially for employees whose career progression has plateaued. As a result, the combination of limited advancement and restricted developmental opportunities contributes to an increased risk of voluntary attrition.

On the other hand, Profile 4 encompasses a larger group of employees (88) with an observed attrition rate of 18.2%, which is 2.2 times the overall test-set average, and an average predicted risk of 31.1%. Employees in this profile have a Functional Promotion Interval of up to 106 months, an Enablement Level of 4, and received a salary increase during the observation period (2022–2024). In contrast to Profile 3, which shares similar promotion and enablement characteristics but did not receive a salary increase, this group shows a significant reduction in both observed and predicted attrition risk. This suggests that even modest compensation growth can partially mitigate turnover risk by reinforcing perceptions of recognition and fairness. However, the attrition rate remains substantially above the dataset baseline, indicating that salary adjustments alone are insufficient to offset prolonged career stagnation. To achieve sustained retention, organisations should integrate financial incentives with clearer promotion pathways and expanded developmental opportunities to address underlying dissatisfaction related to career progression.

Profile 10 comprises a smaller employee segment (n = 12) with characteristics similar to Profile 9, except that these employees did not receive a salary increase (≤ 0%) during the analysed period. Despite the absence of salary progression, Profile 10 exhibits a lower observed attrition rate (16.7%) and predicted risk (6.9%) compared to Profile 9 (23.5% and 18.5%, respectively). This unexpected pattern may result from the small sample size or from interactions identified by other trees in the LightGBM ensemble, where the combined effect of features moderates the impact of salary stagnation. Therefore, this result should be interpreted with caution, as it does not imply that zero salary growth reduces attrition risk, but rather that this specific leaf reflects a localized, model-specific interaction within the broader feature space.

Finally, Profile 2 constitutes the largest employee segment (n = 150) identified within the test dataset, characterized by the lowest observed attrition rate (12.0%) and a moderate predicted risk (21.2%), which is 1.47 times the overall average. Employees in this profile have Functional Promotion Intervals of up to 106 months, Loyalty Levels ≤ 4, and no access to a company car. Although these employees display only moderate loyalty scores, indicating

potential for improvement in organisational commitment, the group remains relatively stable. This stability is likely due to other intrinsic or extrinsic motivators not captured in this specific rule. The absence of a company car, often a proxy for seniority or managerial status, suggests these are mid-level employees for whom the perceived value of stability outweighs the incentive to leave. These findings indicate that, while higher loyalty levels could further reduce attrition, maintaining consistent engagement and perceived fairness among this large and stable group is essential for sustaining overall workforce retention.

4.5.Study Limitations

When interpreting the model performance, it is important to acknowledge the limitations introduced by the aggregated nature of certain features. Because sensitive variables such as age, organisational tenure, annual compensation, salary increase percentage and workplace location, were provided in grouped categories rather than in their raw form, the model was restricted in its ability to capture more granular patterns of employee attrition. This may partially explain why some performance metrics, particularly precision and recall for the minority class (attrition cases), were lower than might otherwise be expected with more detailed data. For instance, the model can detect general trends — such as higher attrition risk among early-career employees or certain job levels — but cannot differentiate between more subtle within-group variations, such as the difference between employees with two versus four years of tenure.

Therefore, the results presented here should be interpreted as conservative estimates of the model's predictive performance metrics (i.e., f1 score, precision, recall, specificity, AUC-ROC). The model demonstrates robust generalization to broad categories, but its ability to capture fine-grained relationships is inherently limited by the level of data aggregation. This trade-off reflects the balance between ensuring data confidentiality and enabling predictive modeling within the constraints of the dataset.

A further limitation of this study concerns the interpretability of specific dataset features, such as, *Board_Member_Oversight*, *Department*, *Unit*, *Job Level*, and (Office) *Location*, which were intentionally in encoded formats to maintain confidentiality. While this approach safeguards employee privacy, it reduces transparency by obscuring the direct real-world meaning behind each encoded category.

Lastly, a significant issue in this study is the pronounced class imbalance in the dataset, as only a small proportion of employees left the company. Because the model was exposed to many more examples of retention than attrition, it learned majority-class patterns more

effectively, which contributed to its observed overfitting behaviour. This effect is reflected in the difference between performance on the training and validation sets (i.e., the generalization gap reported in Chapter IV), where the model achieved higher predictive performance during training than on unseen data. Although measures such as stratified sampling and evaluation using F1 score, recall, precision, AUC-ROC were applied to mitigate this imbalance, the limited representation of attrition cases inherently constrained the model's ability to generalize, particularly when distinguishing employees at higher risk of leaving.

Chapter V - Conclusion and Recommendations

This study investigated the application of machine learning predictive models to support human resource management decisions, focusing specifically on employee attrition within a Portuguese technology company. The organisation identified substantial direct and indirect costs resulting from voluntary employee departures. Direct costs include recruitment, advertising, training expenses, and salary differentials between departing and newly hired staff. Indirect costs, which are often more significant, include reduced productivity during onboarding, diminished team cohesion, loss of organisational knowledge, and decreased engagement among remaining employees. The company's recognition of these effects reinforces the strategic importance of implementing data-driven retention mechanisms to address both financial and cultural consequences of attrition.

Several classification algorithms, including ensemble and tree-based methods, were evaluated to predict voluntary employee attrition and to identify the most influential factors contributing to turnover. The LightGBM classifier achieved the highest overall performance, demonstrating superior predictive precision, interpretability, and generalization as measured by the F1 score on the test dataset (0.578). The results demonstrate that predictive models are effective tools for anticipating voluntary attrition. By identifying employees at elevated risk of departure, these models enable organisations to implement preventive and targeted interventions before resignations occur. The company reported that the primary value of the model lies in its ability to identify the variables that most significantly influence attrition, thus supporting the development of strategic, organisation-wide retention initiatives rather than solely individualized actions. The predictive framework therefore functions both as a forecasting instrument and as a decision-support system for human capital management.

The study also identified that variables such as Functional Promotion Interval, Variable Compensation, Loyalty Level, and Salary Increase are particularly influential in predicting attrition in this organisation. These results highlight the significance of career progression opportunities, equitable compensation structures, and engagement-focused policies. In addition, the analysis of representative employee profiles, constructed from the LightGBM model's most influential decision rules, offered interpretable insights into retention and departure risk patterns, facilitating the translation of technical predictions into actionable organisational understanding.

From a managerial perspective, the outputs of the model have several practical implications for human resources operations and strategic planning. The capacity to identify employees with

a high probability of leaving supports proactive workforce planning, a process that the company has explicitly prioritised as it transitions toward a skills-based organisational structure. Predictive insights inform targeted retention strategies for critical roles or high-potential employees, enabling early intervention in cases of stagnation, dissatisfaction, or disengagement.

Additionally, based on the model's results, the company identified several actionable policy adjustments guided by the variables found to be most decisive for attrition. For example:

i. Reducing the functional promotion interval or creating micro-promotion opportunities could enhance career progression and perceived fairness.
ii. Revising the variable compensation structure, by adjusting its timing or weighting, could strengthen employee motivation and perceived equity.
iii. Reinforcing programs that foster loyalty and organisational culture, such as wellness initiatives, feedback systems, and inclusion strategies, could improve engagement.
iv. Reviewing the salary increase policy to ensure consistency and transparency could mitigate attrition driven by perceived inequity.

Data visualization techniques play a key role for communicating these insights effectively. In this research, visual tools such as charts, tables, and heatmaps were employed to facilitate the interpretation of model outcomes, highlight variable importance, and reveal patterns of attrition across employee segments.

On the other hand, successful adoption of predictive models requires continuous monitoring mechanisms. To maintain their reliability and organisational relevance over time, model performance should be periodically validated by systematically comparing predicted attrition probabilities with actual employee departure events. As a natural extension of this work, the company's HR analytics team could further enhance these monitoring efforts by developing interactive dashboards for real-time visualization of attrition risk, integrating predictive analytics into daily HR decision-making and promoting continuous learning.

In conclusion, this research demonstrates that machine learning models, when properly designed, validated, and interpreted, can deliver significant value to organisations aiming to understand and reduce voluntary employee attrition. Beyond prediction, these models facilitate strategic foresight, support evidence-based decision-making, and enhance alignment between human capital strategies and organisational objectives. Integrating predictive analytics into human resources practice is therefore a crucial step toward proactive talent management and long-term workforce sustainability.

References

Abii, F. E., Ogula, D. C. N., & Rose, J. M. (2013). Effects of Individual and Organizational Factors on the Turnover Intentions of Information Technology Professionals. International Journal of Management, 30(2), 740-756. https://www.proquest.com/scholarly-journals/effects-individual-organizational-factors-on/docview/1367069024/se-2

Aggarwal, A., Jaisinghani, D., & Nobi, K. (2022). Effect of organizational justice and support on organizational commitment and employee turnover intentions: the mediating role of employee engagement. International Journal of Quality and Service Sciences, 14(4), 525–554. https://doi.org/10.1108/ijqss-08-2021-0112

Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

Ang, S., & Slaughter, S. (2004). Turnover of information technology professionals. ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 35(3), 11–27. https://doi.org/10.1145/1017114.1017118

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: why HR is set to fail the big data challenge. Human Resource Management Journal, 26(1), 1–11. Portico. https://doi.org/10.1111/1748-8583.12090

Baldi, F., & Trigeorgis, L. (2020). Valuing human capital career development: a real options approach. Journal of Intellectual Capital, 21(5), 781–807. https://doi.org/10.1108/jic-06-2019-0134

Baptista, J., Stein, M.-K., Klein, S., Watson-Manheim, M. B., & Lee, J. (2020). Digital work and organisational transformation: Emergent Digital/Human work configurations in modern organisations. The Journal of Strategic Information Systems, 29(2), 101618. https://doi.org/10.1016/j.jsis.2020.101618

Baron, J. N., Hannan, M. T., & Burton, M. D. (2001). Labor Pains: Change in Organizational Models and Employee Turnover in Young, High-Tech Firms. American Journal of Sociology, 106(4), 960–1012. https://doi.org/10.1086/320296

Bhupathi, P., Prabu, S., & Goh, A. P. I. (2023). ARTIFICIAL INTELLIGENCE-ENABLED KNOWLEDGE MANAGEMENT USING A MULTIDIMENSIONAL ANALYTICAL FRAMEWORK OF VISUALIZATIONS. International Journal of Cognitive Computing in Engineering, 4, 240–247. https://doi.org/10.1016/j.ijcce.2023.06.003

Bhuvanaiah, T., & Raya, R. P. (2016). Predicting Employee Work Engagement Levels, Determinants and Performance Outcome: Empirical Validation in the Context of an Information Technology Organization. Global Business Review, 17(4), 934–951. https://doi.org/10.1177/0972150916645696

Castanias, R. P., & Helfat, C. E. (2001). The managerial rents model: Theory and empirical analysis. Journal of Management, 27(6), 661–678. https://doi.org/10.1177/014920630102700604

Chalutz Ben-Gal, H. (2019). An ROI-based review of HR analytics: practical implementation tools. Personnel Review, 48(6), 1429–1448. https://doi.org/10.1108/pr-11-2017-0362

Chen, D., Fraiberger, S. P., Moakler, R., & Provost, F. (2017). Enhancing Transparency and Control When Drawing Data-Driven Inferences About Individuals. Big Data, 5(3), 197–212. https://doi.org/10.1089/big.2017.0074

Cnor, L.L. (2004). The ROI of Human Capital: Measuring the Economic Value of Employee Performance. AORN Journal. https://doi.org/10.5860/choice.38-1029

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. International Journal of Organizational Analysis, 28(6), 1273–1291. https://doi.org/10.1108/ijoa-10-2019-1903

Felstead, A., & Henseke, G. (2017). Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. New Technology, Work and Employment, 32(3), 195–212. Portico. https://doi.org/10.1111/ntwe.12097

Field, A. (2013). Discovering statistics using IBM SPSS Statistics (4th ed.). SAGE Publications.

Gupta, M., Jain, A., Nayak, N. B., Kumar, A., & Behl, A. (2021). Developing a Framework for Electronic Engagement at Work. Journal of Global Information Management, 30(5), 1–16. https://doi.org/10.4018/jgim.292063

Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. (2011). Meta-Analytic

Review of Employee Turnover as a Predictor of Firm Performance. Journal of Management, 39(3), 573–603. https://doi.org/10.1177/0149206311424943

HILL, E. J., MILLER, B. C., WEINER, S. P., & COLIHAN, J. (1998). INFLUENCES OF THE VIRTUAL OFFICE ON ASPECTS OF WORK AND WORK/LIFE BALANCE. Personnel Psychology, 51(3), 667–683. Portico. https://doi.org/10.1111/j.1744-6570.1998.tb00256.x

Hom, P. W., Allen, D. G., & Griffeth, R. W. (2019). Employee Retention and Turnover: Why employees stay or leave. Routledge. https://doi.org/10.4324/9781315145587

Jain, P. K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees' attrition: a machine learning approach. SN Applied Sciences, 2(4). https://doi.org/10.1007/s42452-020-2519-4

Johnson, T. (2018, June 29). The real problem with tech professionals: High turnover. Forbes. Retrieved November 10, 2024, from https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2018/06/29/the-real-problem-with-tech-professionals-high-turnover/

Kabalina, V., & Makarova, A. (2022). Employee engagement: systematization of approaches to the definition and measurement. Организационная Психология, 12(3), 110–137. https://doi.org/10.17323/2312-5942-2022-12-3-110-137

Kamalanabhan, T. J., Sai, L. P., & Mayuri, D. (2009). Employee Engagement and Job Satisfaction in the Information Technology Industry. Psychological Reports, 105(3), 759–770. https://doi.org/10.2466/pr0.105.3.759-770

Kapoor, B., & Sherif, J. (2012). Human resources in an enriched environment of business intelligence. Kybernetes, 41(10), 1625–1637. https://doi.org/10.1108/03684921211276792

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khera, S. N., & Divya. (2019). Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. Vision: The Journal of Business Perspective, 23(1), 12–21. https://doi.org/10.1177/0972262918821221

Khodakarami, N., Dirani, K., & Rezaei, F. (2018). Employee engagement: finding a generally accepted measurement scale. Industrial and Commercial Training, 50(6), 305–311. https://doi.org/10.1108/ict-11-2017-0090

King, K. G. (2016). Data Analytics in Human Resources: A Case Study and Critical Review. Human Resource Development Review, 15(4), 487-495. https://doi.org/10.1177/1534484316675818

Krishna, S., & Borah, D. (2024). Data-Driven Strategies for Sustainable Employee Retention: Leveraging Machine Learning to Predict and Prevent Attrition. South Asian Journal of Management, 30(5), 11–38. https://doi.org/10.62206/sajm.30.5.2024.11-38

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, S. (2024, January 12). What is label encoding? Application of label encoder in machine learning and deep learning models. Medium. https://medium.com/@sunnykumar1516/what-is-label-encoding-application-of-label-encoder-in-machine-learning-and-deep-learning-models-c593669483ed

Kuriakose, A. (2021). HR ANALYTICS – EMPOWERING ORGANIZATIONS. Distance Education in Ukraine: Innovative, Normative-Legal, Pedagogical Aspects, 1, 35–45. https://doi.org/10.18372/2786-5495.1.15737

Lewis, G., & Soroñgon, J. (2022, August 11). Industries with the highest (and lowest) turnover rates. LinkedIn. Retrieved November 10, 2024, from https://www.linkedin.com/business/talent/blog/talent-strategy/industries-with-the-highest-turnover-rates

Mackey, A., Molloy, J. C., & Morris, S. S. (2013). Scarce Human Capital in Managerial Labor Markets. Journal of Management, 40(2), 399–421. https://doi.org/10.1177/0149206313517265

Mansi, & Goswami, T. G. (2024). How HR Analytics Can Help Improve Decision-Making, HR Practices and Firm Performance: A Systematic Review. IUP Journal of Organizational Behavior, 23(1), 87-111. https://www.proquest.com/scholarly-journals/how-hr-analytics-can-help-improve-decision-making/docview/2973218176/se-2

Marín Díaz, G., Galán Hernández, J. J., & Galdón Salvador, J. L. (2023). Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. Mathematics, 11(22), 4677. https://doi.org/10.3390/math11224677

Martin, K. (2023). Predatory predictions and the ethics of predictive analytics. Journal of the Association for Information Science and Technology, 74(5), 531–545. Portico. https://doi.org/10.1002/asi.24743

McCartney, S., & Fu, N. (2022). Bridging the gap: why, how and when HR analytics can impact organizational performance. Management Decision, 60(13), 25–47. https://doi.org/10.1108/md-12-2020-1581

ML Contests. (2024). State of machine learning competitions 2024. Retrieved from https://mlcontests.com/state-of-machine-learning-competitions-2024/#time-series--tabular-data

Mueller, F. (1996). HUMAN RESOURCES AS STRATEGIC ASSETS: AN EVOLUTIONARY RESOURCE-BASED THEORY*. Journal of Management Studies, 33(6), 757–785. Portico. https://doi.org/10.1111/j.1467-6486.1996.tb00171.x

Pandita, D., & Ray, S. (2018). Talent management and employee engagement – a meta-analysis of their impact on talent retention. Industrial and Commercial Training, 50(4), 185–199. https://doi.org/10.1108/ict-09-2017-0073

Pourkhodabakhsh, N., Mamoudan, M. M., & Bozorgi-Amiri, A. (2022). Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover. Applied Intelligence, 53(12), 16309–16331. https://doi.org/10.1007/s10489-022-04294-6

Pramoditha, R. (2021, December 16). Encoding categorical variables: One-hot vs dummy encoding. Towards Data Science. https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db/

Ravalier, J. M., Dandil, Y., & Limehouse, H. (2015). Employee engagement and management standards: a concurrent evaluation. Occupational Medicine, 65(6), 496–498. https://doi.org/10.1093/occmed/kqv071

Raj, A. (2023). Deciphering the augmentation of classification models in predicting employee attrition [Master's thesis, National College of Ireland]. NORMA eResearch Repository. https://norma.ncirl.ie/6267/1/adityaraj.pdf

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 172, 1–102. https://op.europa.eu/en/publication-detail/-/publication/d79f3e5d-41bc-11f0-b9f2-01aa75ed71a1/language-en

Sandoval-Reyes, J., Idrovo-Carlier, S., & Duque-Oliva, E. J. (2021). Remote Work, Work Stress, and Work–Life during Pandemic Times: A Latin America Situation. International Journal of Environmental Research and Public Health, 18(13), 7069. https://doi.org/10.3390/ijerph18137069

Saratun, M. (2016). Performance management to enhance employee engagement for corporate sustainability. Asia-Pacific Journal of Business Administration, 8(1), 84–102. https://doi.org/10.1108/apjba-07-2015-0064

Saradha, H., & Patrick, H.A. (2010). Employee Engagement in Relation to Organizational Citizenship Behavior in Information Technology Organizations. Journal of Marketing Management, 2, 74. https://www.semanticscholar.org/paper/Employee-Engagement-in-Relation-to-Organizational-Saradha-Patrick/e047a8a62cd3f313b662df25814b4788055e2d36

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Sekhar, C., Patwardhan, M., & Vyas, V. (2017). Linking Work Engagement to Job Performance Through Flexible Human Resource Management. Advances in Developing Human Resources, 20(1), 72–87. https://doi.org/10.1177/1523422317743250

Sharma, P. K., & Kumra, R. (2020). Relationship between workplace spirituality, organizational justice and mental health: mediation role of employee engagement. Journal of Advances in Management Research, 17(5), 627–650. https://doi.org/10.1108/jamr-01-2020-0007

Shaw, J. D. (2011). Turnover rates and organizational performance. Organizational Psychology Review, 1(3), 187–213. https://doi.org/10.1177/2041386610382152

Shrotryia, V. K., & Dhanda, U. (2020). Development of employee engagement measure: experiences from best companies to work for in India. Measuring Business Excellence, 24(3), 319–343. https://doi.org/10.1108/mbe-07-2019-0071

Shuck, B., & Reio, T. G. (2011). The Employee Engagement Landscape and HRD. Advances in Developing Human Resources, 13(4), 419–428. https://doi.org/10.1177/1523422311431153

Shuck, B., Twyford, D., Reio, T. G., & Shuck, A. (2014). Human Resource Development Practices and Employee Engagement: Examining the Connection With Employee Turnover Intentions. Human Resource Development Quarterly, 25(2), 239–270. Portico. https://doi.org/10.1002/hrdq.21190

Singh, R., Sharma, P., Foropon, C., & Belal, H. M. (2022). The role of big data and predictive analytics in the employee retention: a resource-based view. International Journal of Manpower, 43(2), 411–447. https://doi.org/10.1108/ijm-03-2021-0197

Song, H. D., Hong, A. J., & Jo, Y. (2021). Development and validation of the Korean Employee Engagement Scale. Human Resource Development International, 24(5), 533–557. https://doi.org/10.1080/13678868.2021.1959831

Statistics Portugal (INE). (2025, February 7). Unemployment rate stood at 6.7% in Q4 2024. https://www.ine.pt/xportal/xmain?DESTAQUESdest_boui=645506474&DESTAQUESmodo=2&xpgid=ine_destaques&xpid=INE

Statistics Portugal (INE). (2025, July 10). Consumer Price Index (CPI) – June 2025. https://www.ine.pt/xportal/xmain?DESTAQUESdest_boui=707084164&DESTAQUESmodo=2&xpgid=ine_destaques&xpid=INE

Tejero, L. M. S., Seva, R. R., & Fadrilan-Camacho, V. F. F. (2021). Factors Associated With Work-Life Balance and Productivity Before and During Work From Home. Journal of Occupational & Environmental Medicine, 63(12), 1065–1072. https://doi.org/10.1097/jom.0000000000002377

Thakral, P., Srivastava, P. R., Dash, S. S., Jasimuddin, S. M., & Zhang, Z. (Justin). (2023). Trends in the thematic landscape of HR analytics research: a structural topic modeling approach. Management Decision, 61(12), 3665–3690. https://doi.org/10.1108/md-01-2023-0080

Tiwari, B., & Lenka, U. (2020). Employee engagement: A study of survivors in Indian IT/ITES sector. IIMB Management Review, 32(3), 249–266. https://doi.org/10.1016/j.iimb.2019.10.003

Trisna Wijayanthi, I. A., Subagio, M., Suhendra, A., Hartini, H., & Amin Darmawan, A. (2024). Employee Retention Implementation Prevents Employee Turnover. Siber Journal of Advanced Multidisciplinary, 2(2), 288–301. https://doi.org/10.38035/sjam.v2i2.211

Ulrich, D., & Smallwood, N. (2005). HR's new ROI: Return on intangibles. Human Resource Management, 44(2), 137–142. Portico. https://doi.org/10.1002/hrm.20055

Vasujoseph, V. (2023, August 22). Understanding feature relationships through correlation analysis: A guide for data enthusiasts. Medium. https://medium.com/@vvasujoseph/understanding-feature-relationships-through-correlation-analysis-a-guide-for-data-enthusiasts-f139095e8228

Wehrstein, L., & Bachmann, B. (2020, December 19). CRISP-DM ready for machine learning projects. Towards Data Science. https://towardsdatascience.com/crisp-dm-ready-for-machine-learning-projects-2aad9172056a/

Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. IEEE Access, 9, 60447–60458. https://doi.org/10.1109/access.2021.3074559

Zieglmeier, V., & Pretschner, A. (2023). Rethinking People Analytics With Inverse Transparency by Design. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1–29. https://doi.org/10.1145/3610083

Annexes

A.Employee Data Dictionary 70

B.Distribution of Numerical Features (Histograms) 79

C.Distribution of Categorical Features (Bar Charts) 80

D.Bivariate Attrition Analysis 82

D.1 - Attrition Rate by Generation 82

D.2 - Attrition Rate by Job Level 82

D.3 - Attrition Rate by Department (Top 10 by Headcount) 83

D.4 - Attrition Rate by Functional Area 83

D.5 - Attrition Rate by Work Model 84

D.6 - Attrition Rate by Performance Rating 84

D.7 - Attrition Rate by Functional Promotion Interval 85

D.8 - Attrition Rate by Variable Compensation 85

D.9 - Attrition Rate by Unit 86

D.10 - Attrition Rate by Loyalty Level 86

D.11 - Attrition Rate by Salary Increase 87

E.Filtered Correlation Heatmap 88

F.Percentage of Missing Values per Feature 89

G.Summary of Classifier Experiments 90

H.Bivariate Predicted Attrition Analysis 91

H.1 - Predicted Attrition Rate by Functional Promotion Interval 91

H.2 - Predicted Attrition Rate by Variable Compensation 91

H.3 - Predicted Attrition Rate by Unit 92

H.4 - Predicted Attrition Rate by Loyalty Level 92

H.5 - Predicted Attrition Rate by Salary Increase 93

I.Distribution of Most Influential Features for the LightGBM Model (Test Dataset) 94

I.1 – Distribution of Functional Promotion Interval in the Test Set 94

I.2 - Distribution of Variable Compensation in the Test Set 94

I.3 - Distribution of Unit in the Test Dataset 94

I.4 – Distribution of Loyalty Level in the Test Dataset 95

I.5 – Distribution of Salary Increase in the Test Dataset 95

J.Representative Employee Profiles (Best-performing LightGBM Tree) 96

A. Employee Data Dictionary

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
Employee Demographics	Year	It refers to the year of analysis. The last three years were considered.	INT	2022 2023 2024	HRM System
	Employee_Number	Unique identification key for each employee.	INT	-	HRM System
	Gender	Indicates the employee's gender ('F' for Female and 'M' for Male).	VARCHAR	F M	HRM System
	Generation	It refers to a group of people who were born and lived around the same time and share common experiences, cultural milestones, and historical events that shape their worldview.	VARCHAR	Baby boomer Generation X Millennial Generation Z	HRM System
	Children	Indicates whether the employee has children, with 0 for No and 1 for Yes.	INT	0 1	HRM System
	Tenure	Represents the tenure of an employee within the company, categorized into different levels based on the length of time that an employee has been working in the organisation	VARCHAR	New Joiners Pathfinders Growing Talent Core Contributors Legacy Builders Experienced Veterans	HRM System
Professional & Organisational Information	Location	Workplace or office building where an employee is assigned, categorized numerically (0, 1, 2, 3) to indicate different company locations.	VARCHAR	A B C D	HRM System

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
	Functional_Area	Represents functional clusters that group departments with related functions and expertise within the organisation. Categorization is based on the nature of the work performed, required skills, and overall functional objectives.	VARCHAR	Business of Entertainment Commercial Others Product and Marketing Information Systems Corporate Support Business Support Technology	HRM System
	Board_Member_Oversight	Aggregates departments into groups, with responsibilities assigned to each member of the executive board.	VARCHAR	A, B, C, (…) H	HRM System
	Department	Represents the specific divisions within the organisation that are responsible for certain tasks and functions. A unique numerical identifier was assigned to each department.	INT	1, 2, 3, (…), 49	HRM System
	Unit	Represents the specialized subgroup within a department that focuses on a specific function, task, or project. A unique numerical identifier was assigned to each unit.	FLOAT	1, 2, 2.1, (…), 49.7	HRM System
	Job_Level	Represents the employee's hierarchical level within the organisation, structured based on responsibilities, experience, and tenure. It categorizes employees into different levels of seniority, reflecting their position in the corporate hierarchy.	VARCHAR	EL ML MLL PL AL SL ALL SLL	HRM System
	Leadership_Role	Indicates whether the employee holds a managerial position within the organisation, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
Employee Satisfaction	Work_Model	Indicates the employee's work model, specifying whether they work in a hybrid, on-site, or remote format.	VARCHAR	Hybrid On Site On Site T Remote	HRM System
	Engagement_Level	Represents the engagement level of the employee, indicating their overall commitment, motivation, and involvement within the organisation. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Enablement_Level	Represents the degree to which an employee feels they have the necessary resources, tools, and support to perform their job effectively. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Commitment_Level	Measures an employee's dedication and emotional attachment to their organisation, including their willingness to stay and contribute to its success. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Loyalty_Level	Assesses how likely an employee is to remain with the company, based on factors like job satisfaction, engagement, and long-term career alignment. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Effectiveness_Category	Categorizes employees based on their perceived impact and performance in the workplace, ranging from least effective to most effective. Rated on a scale from 1 to 4, where 1 represents 'Least Effective' and 4	VARCHAR	Least Effective Detached Frustrated Most Effective No data available	Employee Sentiment Analysis

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
		represents the 'Most Effective'.			
	Satisfaction_with_Autonomy	Measures an employee's satisfaction with the level of independence and control they have over their work, decision-making, and daily tasks. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Collaboration	Evaluates how satisfied an employee is with teamwork, communication, and cooperation among colleagues and departments. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Compensation	Assesses an employee's satisfaction with their salary, benefits, and overall financial rewards provided by the organisation. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Leadership_Trust	Reflects the level of trust and confidence an employee has in the organisation's leadership, including their integrity, transparency, and decision-making. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Training	Measures an employee’s level of satisfaction with the training programs provided by the organisation, including relevance, effectiveness, and accessibility. Rated	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
		on a scale from 1 to 5, where 5 represents the highest level of satisfaction.			
	Satisfaction_with_Ways_of_Working	Evaluates how satisfied an employee is with the company's work processes, policies, flexibility, and overall work environment. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Performance_Management	Assesses an employee's perception of the performance evaluation process, including fairness, feedback quality, and alignment with career growth. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Development_Opportunities	Reflects an employee's satisfaction with career growth opportunities, including promotions, mentorship, skill-building, and professional development programs. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Resources	Measures an employee's satisfaction with the availability and quality of resources provided by the organisation, including tools, technology, and support needed to perform their job effectively. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
	Satisfaction_with_Respect_and_Recognition	Assesses how valued employees feel in the workplace, considering both the level of respect from colleagues and management as well as recognition for their contributions and achievements. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
	Satisfaction_with_Clear_Vision	Evaluates an employee's perception of the organisation's strategic direction, including how well the company's mission, vision, and goals are communicated and understood. Rated on a scale from 1 to 5, where 5 represents the highest level of satisfaction.	VARCHAR	1 2 3 4 5 No Data Available	Employee Sentiment Analysis
Career development	Functional_Promotion	Indicates whether the employee has transitioned to a higher Job Level with greater responsibility, in the year in analysis, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System
	Functional_Promotion_Interval	Represents the time interval, in months, between each functional promotion of an employee, indicating a transition to a different Job Level.	INT	0,1,2, (…) 131	HRM System
	Promotion_in_the_Last_3_Years	Indicates whether the employee has received a promotion within the last 3 years, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System
	Promotion_in_the_Last_5_Years	Indicates whether the employee has received a promotion within the last 5 years, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System

Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
	Performance_Rating	Annual performance rating on a scale of 1 to 5, where 1 is insufficient and 5 is excellent. 'No Data Available' means that the employee was ineligible for evaluation in the year in analysis (e.g., due to prolonged leave, seniority, etc.).	VARCHAR	1 2 3 4 5 No Data Available	HRM System
	Internal_Mobility	Indicates whether the employee has undergone internal mobility, transitioning between departments or areas within the organisation, usually through lateral movements (0 represents 'No' and 1 represents 'Yes')	INT	0 1	HRM System
	Career_Opportunity	Indicates whether the employee has had a career opportunity, determined by the combination of 'Internal Mobility' and 'Functional Promotion' attributes. If the employee has experienced either (internal) mobility or a functional promotion, the value is 1 ('Yes'); otherwise, it is 0 ('No'). *Example: If 'Internal_Mobility' = 1 and 'Functional_Promotion' = 0, then 'Career_Opportunity' = 1.*	INT	0 1	HRM System
	Type_of_Cross-Functional_Project	Specifies the type of cross-functional project the employee participated within the organisation, such as Data & AI Transformation, Digital Transformation, Operational Transformation or Others. If no project participation is recorded, the value is 'NA'.	VARCHAR	Data & AI Transformation Digital Transformation Operational Transformation Not Applicable Other No Data Available	HRM System

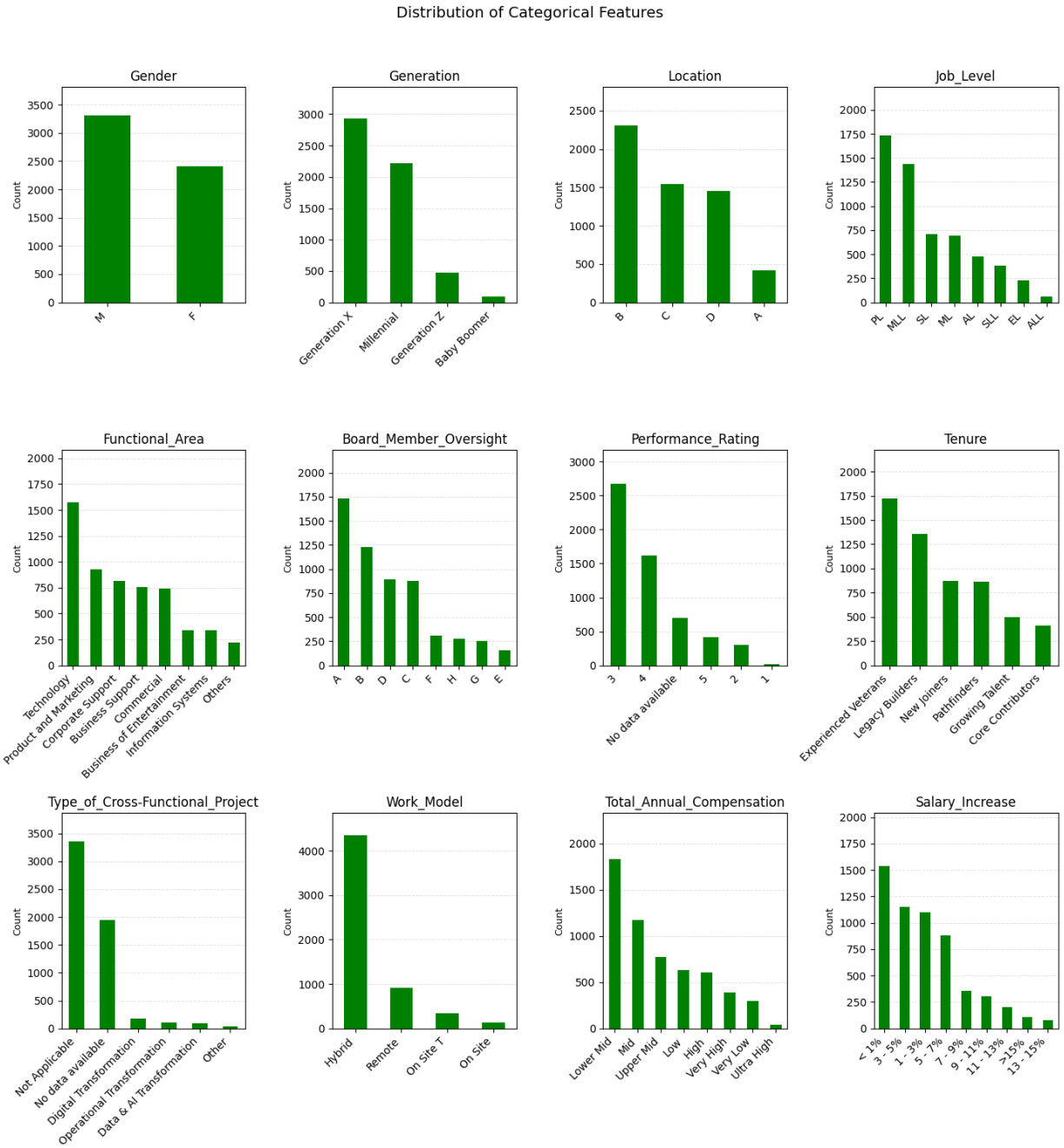
Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
Compensation & Benefits	Total_Annual_Compensation	Represents the employee's total annual remuneration level, categorized into eight tiers: very low, low, lower mid, mid, upper mid, high, very high, and ultra-high.	VARCHAR	Very Low Low Lower Mid Mid Upper Mid High Very High Ultra High	HRM System
	Variable_Compensation	Indicates the percentage of variable remuneration within the employee's total annual salary, representing the adjustable portion of their pay.	FLOAT	0%, 1%, 2%, …, 47%	HRM System
	Company_Car	Indicates whether the employee has been given a company car, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System
	Salary_Increase	Represents the percentage increase in the employee's salary in the year in analysis, expressed as a percentage interval.	VARCHAR	< 1% --> 0 1 - 3% --> 1 3 - 5% --> 2 5 - 7% --> 3 7 - 9% --> 4 9 - 11% --> 5 11 - 13% --> 6 13 - 15% --> 7 > 15% --> 8 No data available	HRM System
Direct Leadership	Change_of_Direct_Supervisor	Indicates whether the employee has experienced a change in their direct manager within the year, where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System
	Change_of_Direct_Supervisor_due_to_Mobility	Indicates whether the employee's change in direct manager was due to internal mobility, determined by the combination of the variables 'Internal Mobility' and 'Change of Supervisor', where 0 represents 'No' and 1 represents 'Yes'.	INT	0 1	HRM System

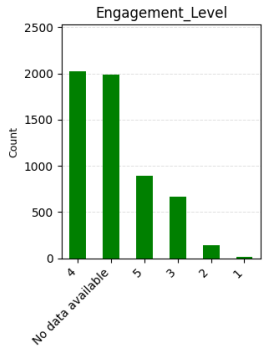
Subject	Attribute Name	Attribute Description	Data Type	Categories	Data Source
		Example: If 'Change_of_Direct_Supervisor' = 1 and 'Internal_Mobility' = 0, then 'Change_of_Direct_Supervisor_due_to_Mobility' = 0.			
External factors	Annual_Unemployment_Rate	Percentage of the labor force that was unemployed in Portugal during the year in analysis.	FLOAT		
	Consumer_Price_Index	A measure of inflation in Portugal, expressed as an index reflecting the average change in prices of a representative basket of goods and services consumed by resident households.	INT		
Exit Information	Attrition	Employee Attrition. Indicates whether the employee has voluntarily left the company, where 1 represents 'Yes' (the employee has exited) and 0 represents 'No' (the employee is still employed).	INT	0 1	

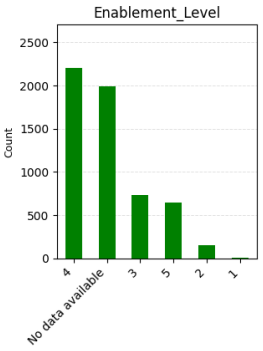
B. Distribution of Numerical Features (Histograms)

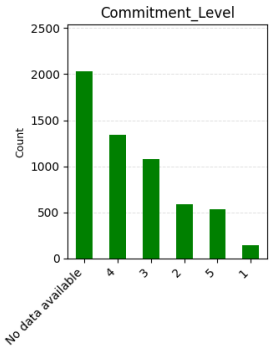


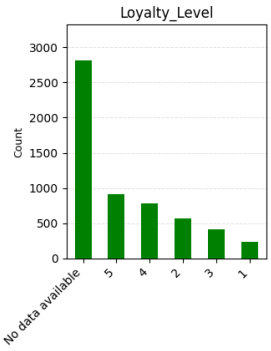
C. Distribution of Categorical Features (Bar Charts)

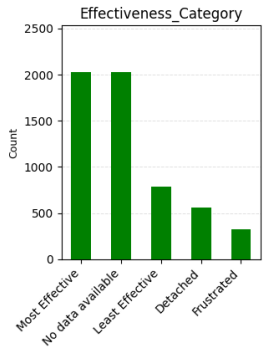


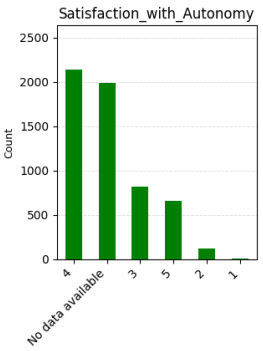

Engagement_Level
Count
No data available

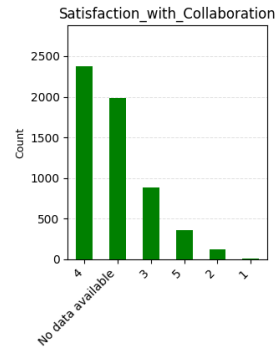

Enablement_Level
Count
No data available

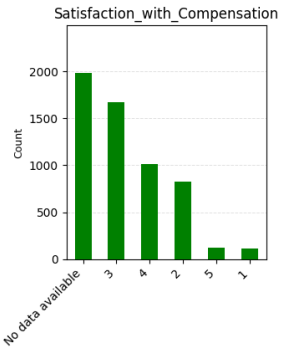

Commitment_Level
Count
No data available

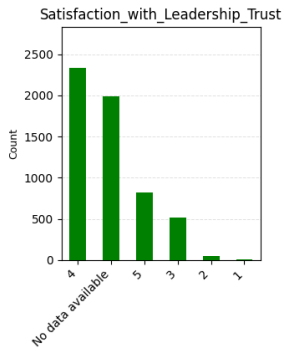

Loyalty_Level
Count
No data available

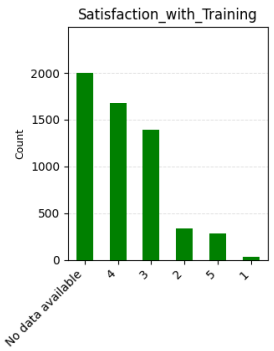

Effectiveness_Category
Count
Most Effective
No data available
Least Effective
Detached
Frustrated

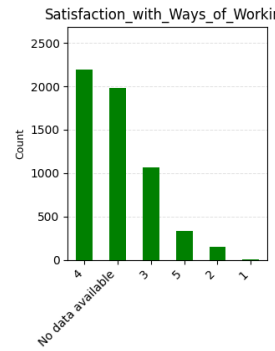

Satisfaction_with_Autonomy
Count
No data available

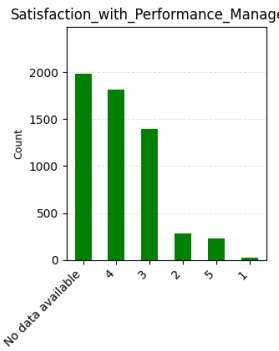

Satisfaction_with_Collaboration
Count
No data available

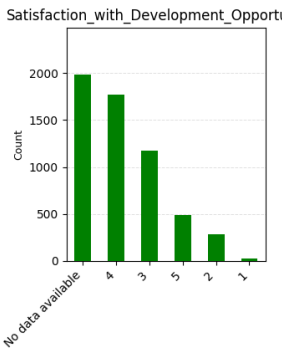

Satisfaction_with_Compensation
Count
No data available

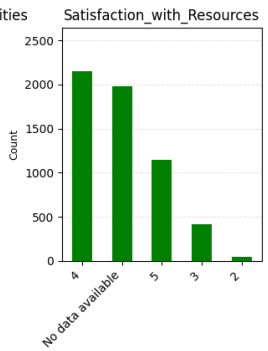

Satisfaction_with_Leadership_Trust
Count
No data available

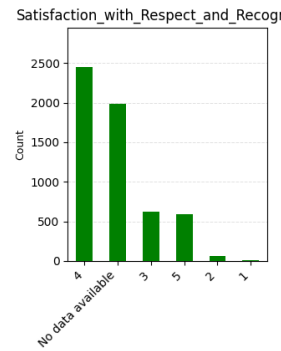

Satisfaction_with_Training
Count
No data available

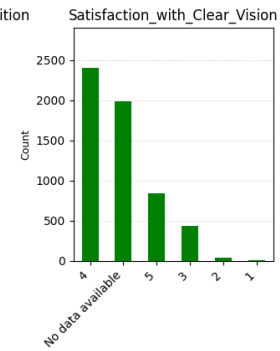

Satisfaction_with_Ways_of_Working
Count
No data available


Satisfaction_with_Performance_Management
Count
No data available


Satisfaction_with_Development_Opportunities
Count
No data available

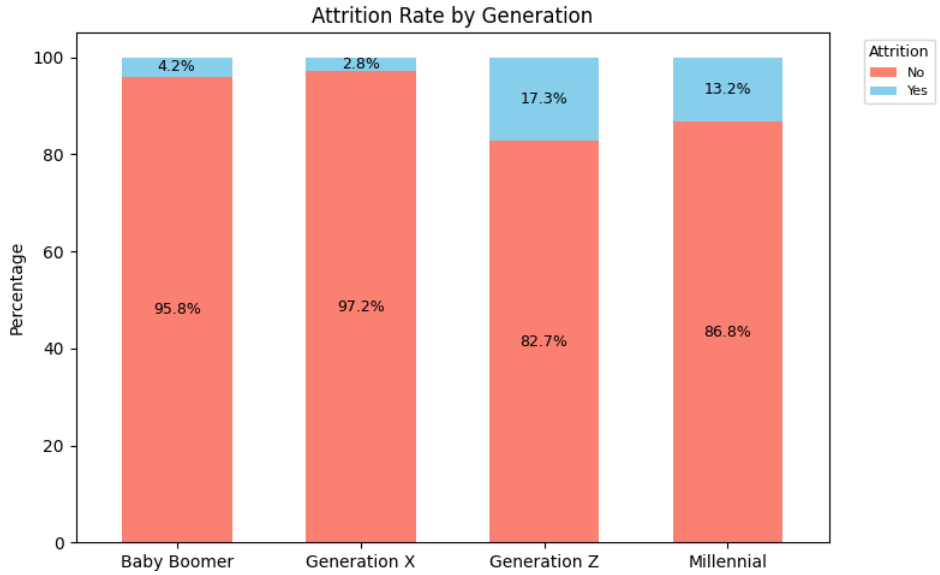

Satisfaction_with_Resources
Count
No data available


Satisfaction_with_Respect_and_Recognition
Count
No data available

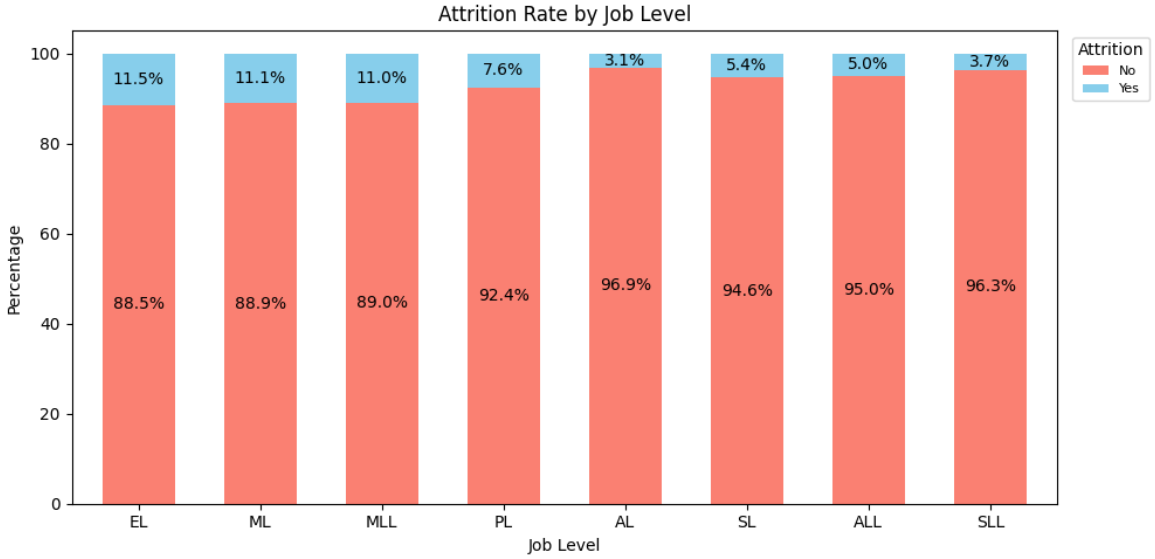

Satisfaction_with_Clear_Vision
Count
No data available

D. Bivariate Attrition Analysis

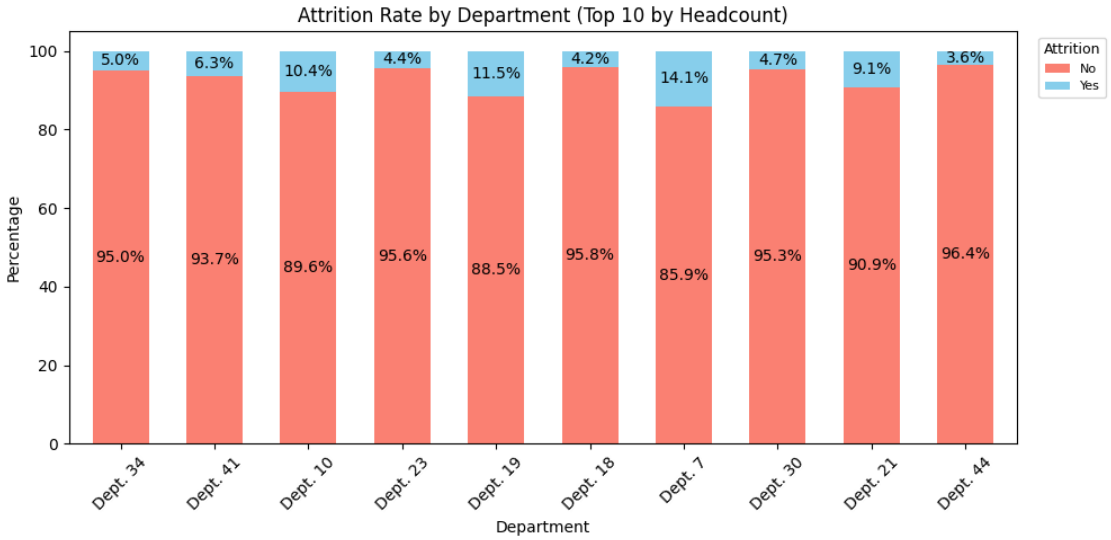
D.1 - Attrition Rate by Generation



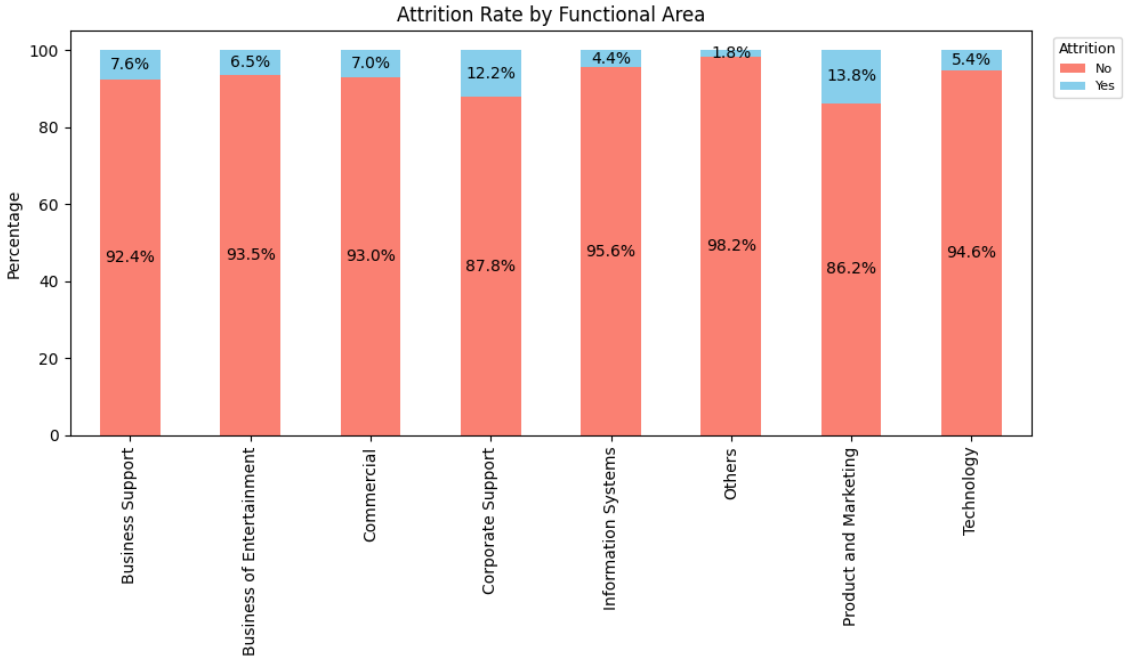
D.2 - Attrition Rate by Job Level



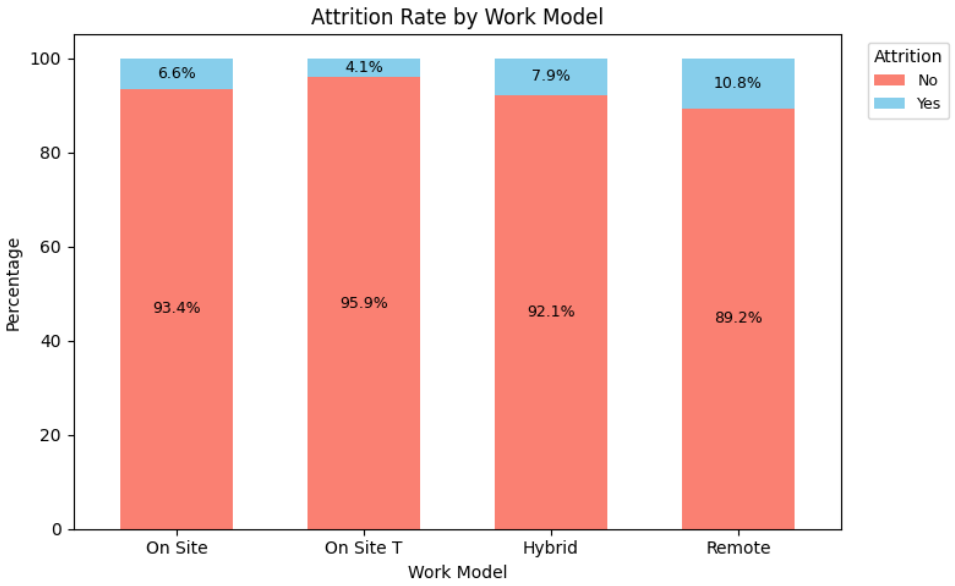
D.3 - Attrition Rate by Department (Top 10 by Headcount)



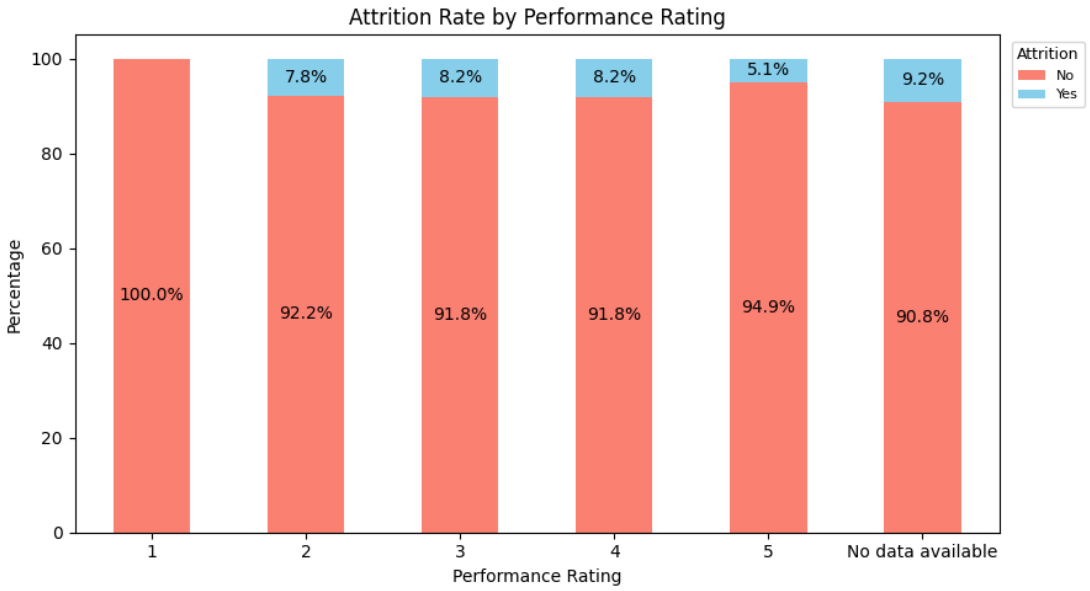
D.4 - Attrition Rate by Functional Area



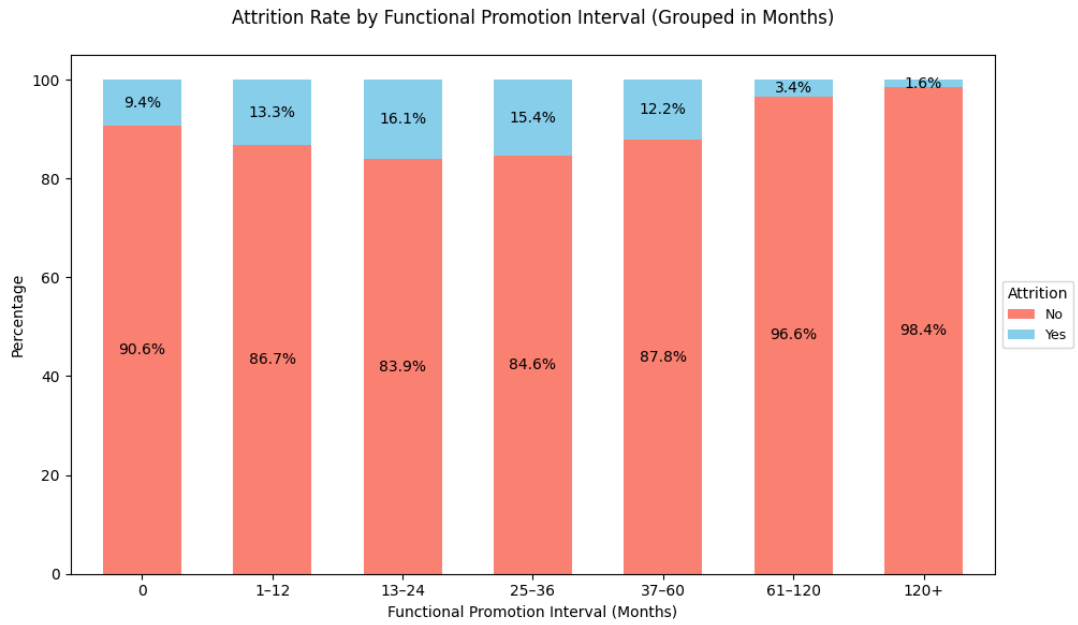
D.5 - Attrition Rate by Work Model



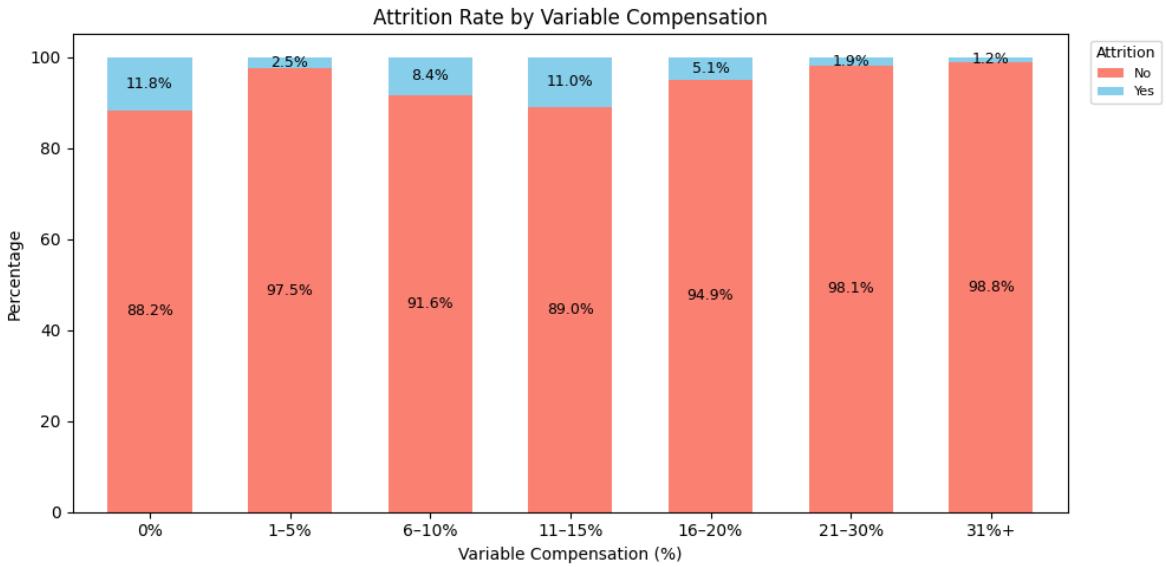
D.6 - Attrition Rate by Performance Rating



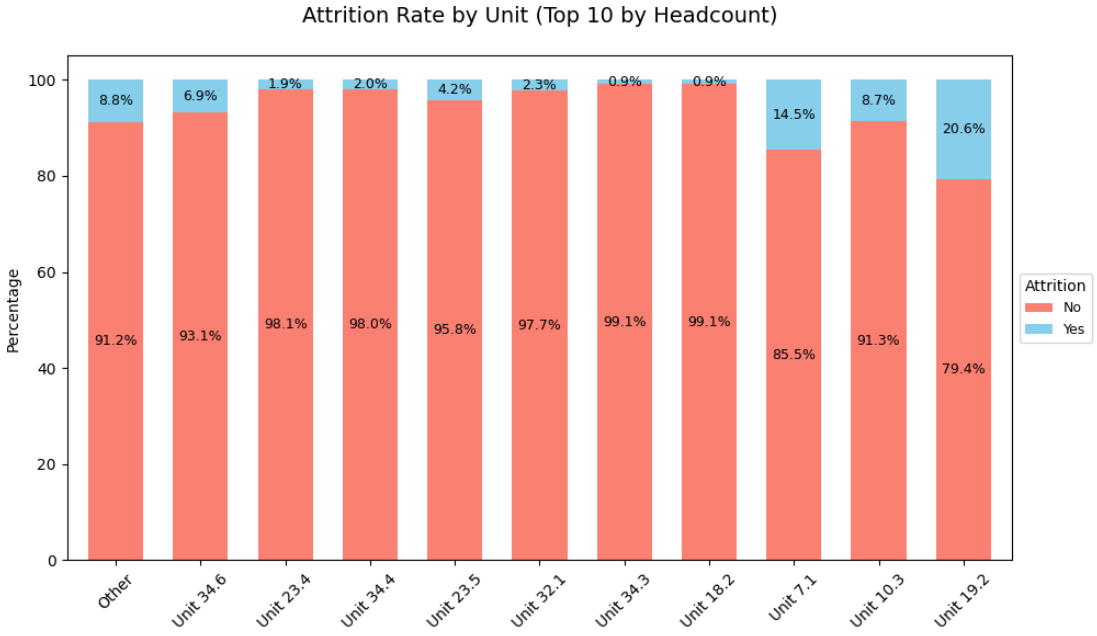
D.7 - Attrition Rate by Functional Promotion Interval



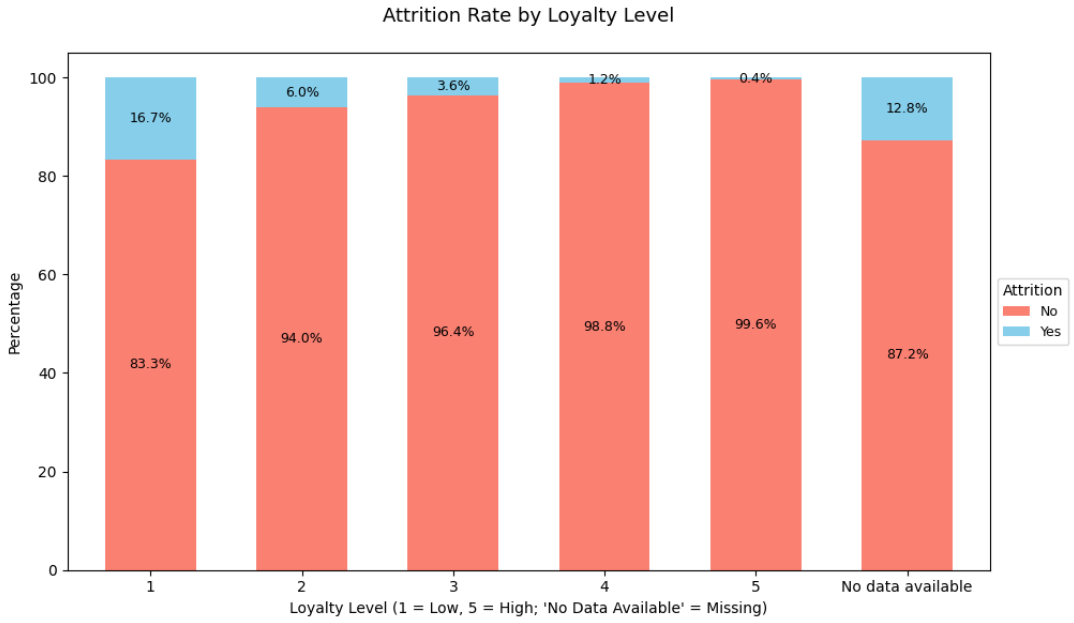
D.8 - Attrition Rate by Variable Compensation



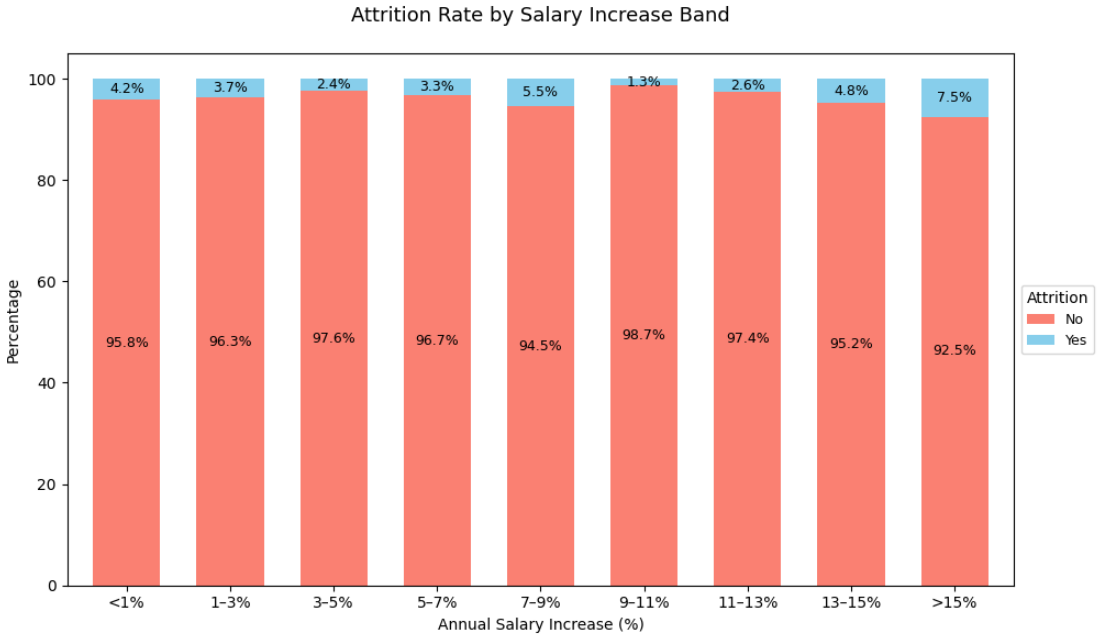
D.9 - Attrition Rate by Unit



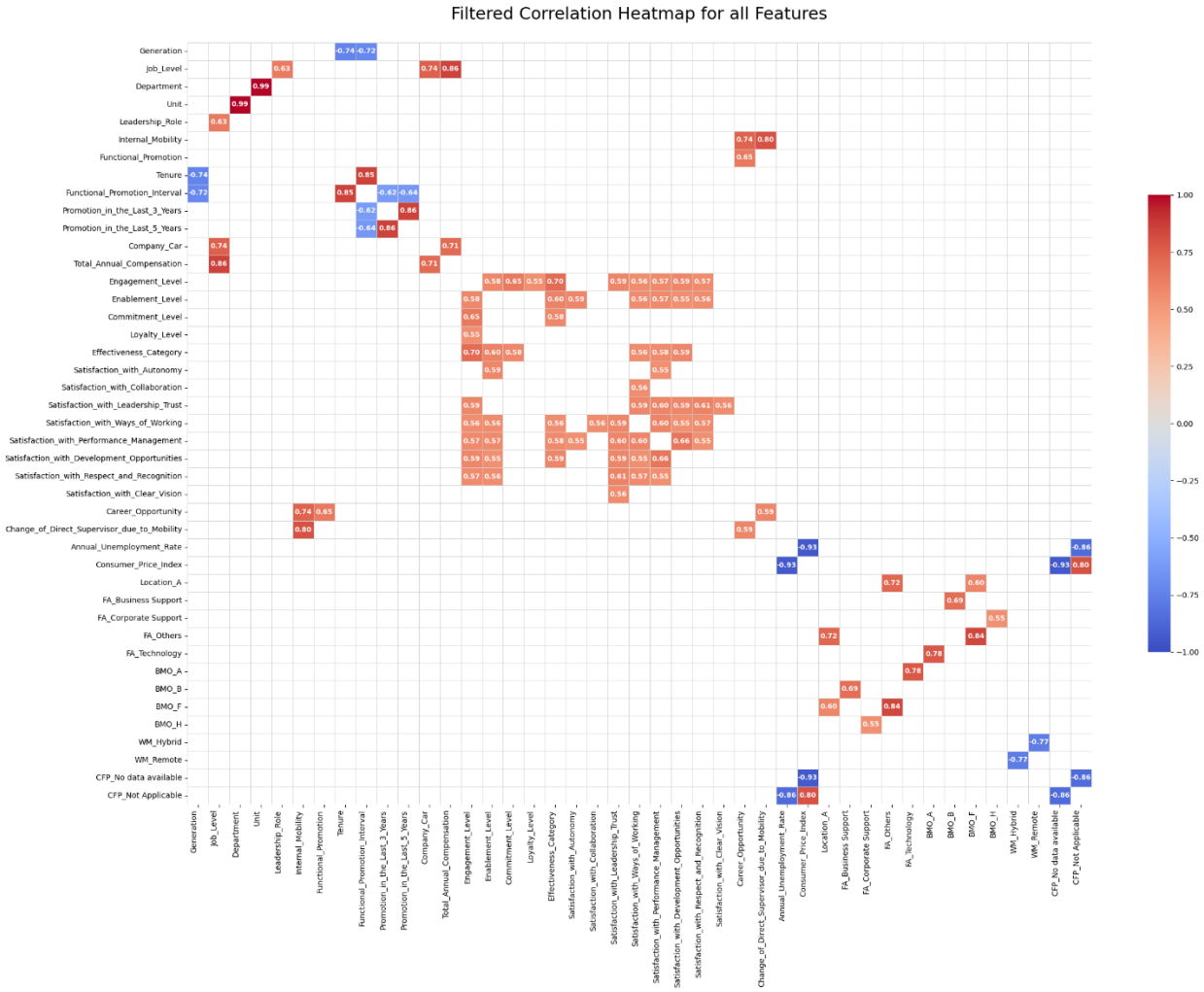
D.10 - Attrition Rate by Loyalty Level



D.11 - Attrition Rate by Salary Increase



E. Filtered Correlation Heatmap



F. Percentage of Missing Values per Feature

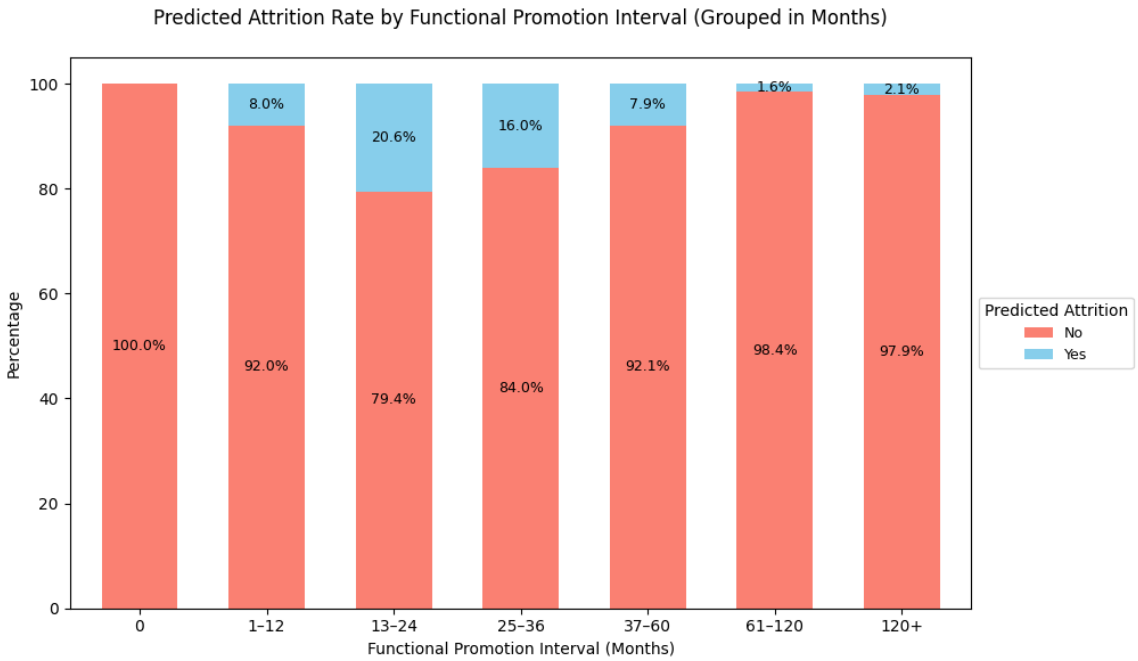
#	Feature	Missing Data (%)
1	Loyalty_Level	49.2
2	Commitment_Level	35.5
3	Effectiveness_Category	35.4
4	Satisfaction_with_Training	34.8
5	Satisfaction_with_Autonomy	34.8
6	Satisfaction_with_Leadership_Trust	34.8
7	Satisfaction_with_Clear_Vision	34.7
8	Satisfaction_with_Respect_and_Recognition	34.7
9	Satisfaction_with_Development_Opportunities	34.7
10	Satisfaction_with_Performance_Management	34.7
11	Satisfaction_with_Compensation	34.7
12	Satisfaction_with_Collaboration	34.7
13	Enablement_Level	34.7
14	Satisfaction_with_Ways_of_Working	34.6
15	Satisfaction_with_Resources	34.6
16	Engagement_Level	34.6

G. Summary of Classifier Experiments

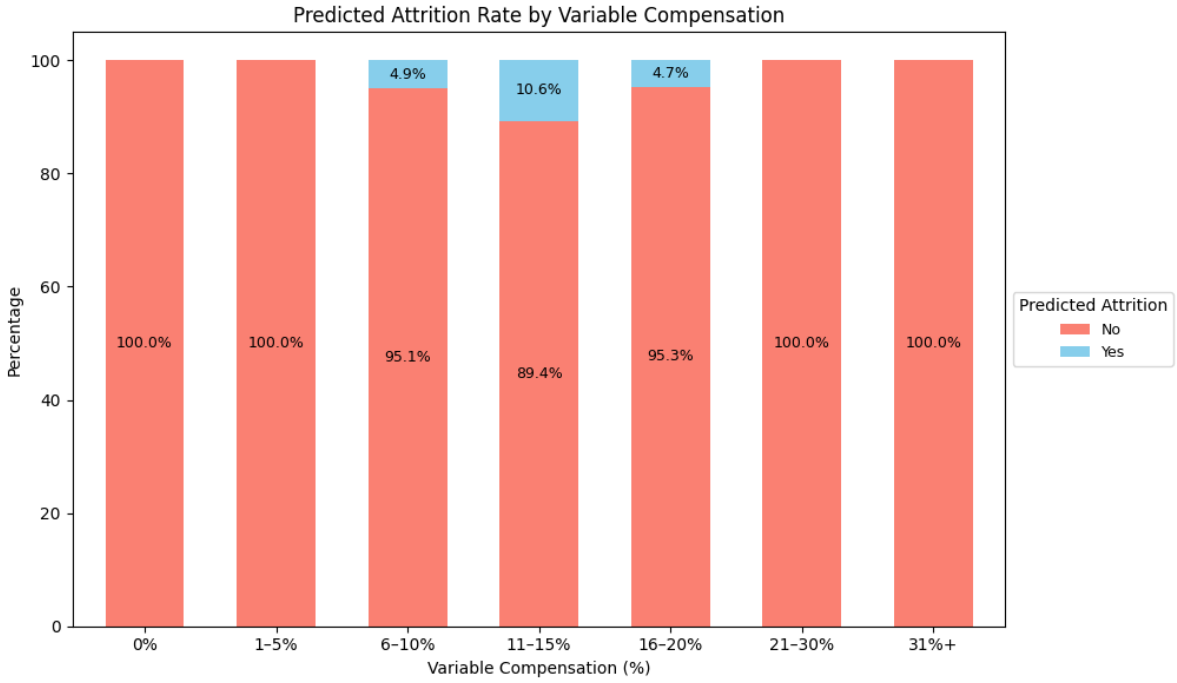
Classifier Model	Hyperparameter Tuning Technique	Decision Threshold Tuning Technique
DecisionTreeClassifier	• RandomizedSearchCV • GridSearchCV • Manual Hyperparameter Tuning	—
BaggingClassifier	• RandomizedSearchCV • GridSearchCV	Manual validation-based threshold
GradientBoostingClassifier	• RandomizedSearchCV • GridSearchCV	Manual validation-based threshold
XGBClassifier	• RandomizedSearchCV • GridSearchCV	• Manual validation-based threshold • OOF tuned threshold • CalibratedClassifierCV • TunedThresholdClassifierCV
CatBoostClassifier	• RandomizedSearchCV • GridSearchCV	• Manual validation-based threshold • OOF tuned threshold • CalibratedClassifierCV • TunedThresholdClassifierCV
LGBMClassifier	• RandomizedSearchCV • GridSearchCV	• Manual validation-based threshold • OOF tuned threshold • CalibratedClassifierCV • TunedThresholdClassifierCV
RandomForestClassifier	• RandomizedSearchCV • GridSearchCV	TunedThresholdClassifierCV

H. Bivariate Predicted Attrition Analysis

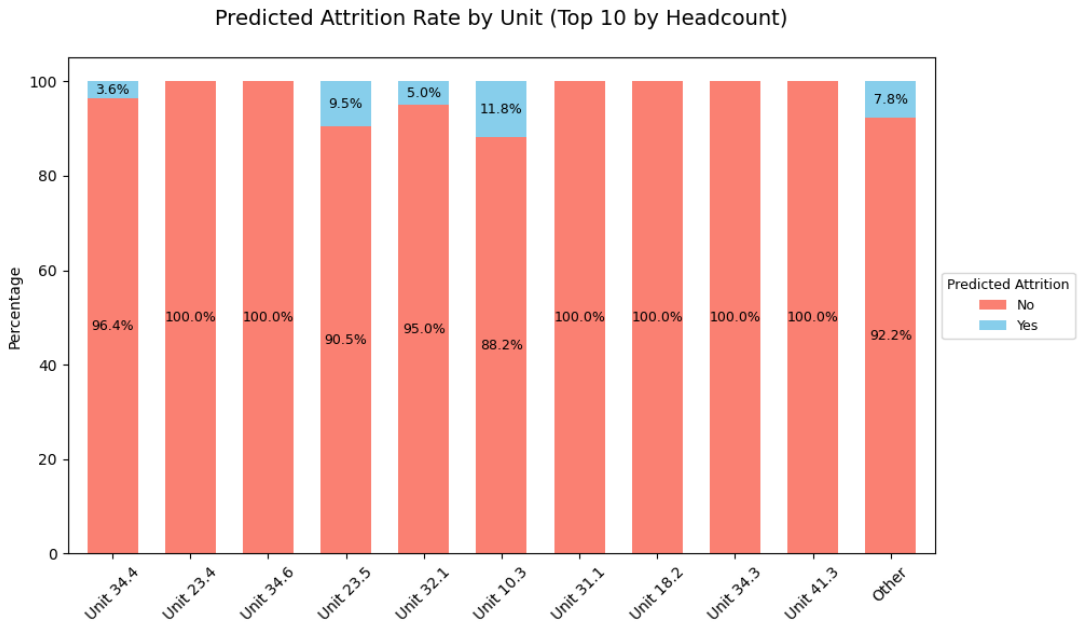
H.1 - Predicted Attrition Rate by Functional Promotion Interval



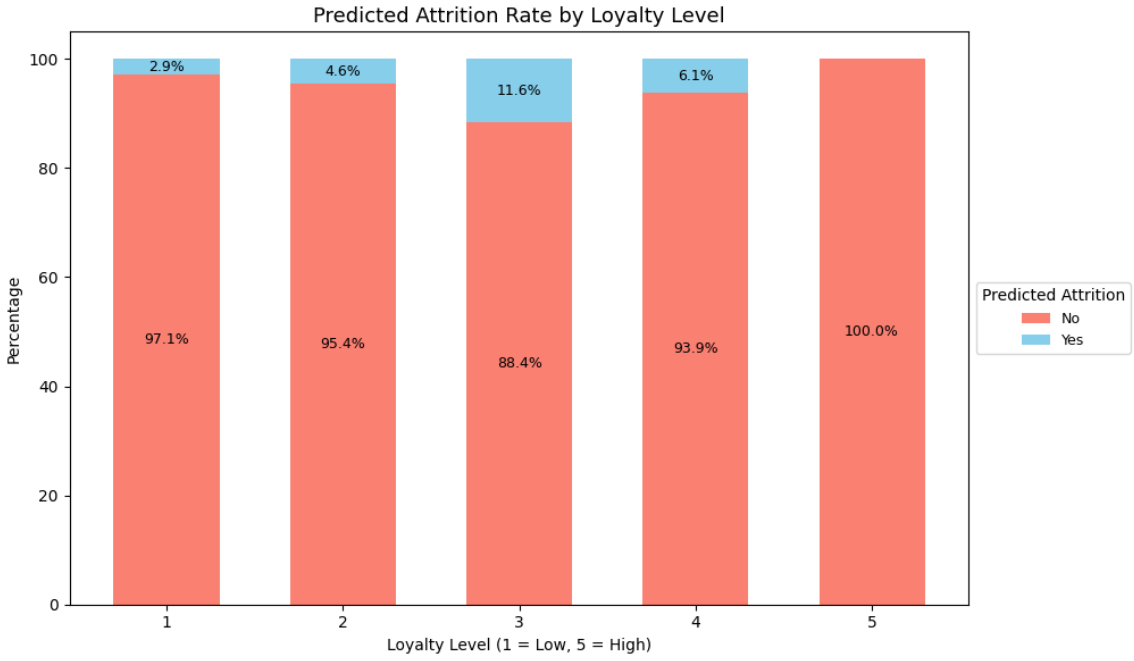
H.2 - Predicted Attrition Rate by Variable Compensation



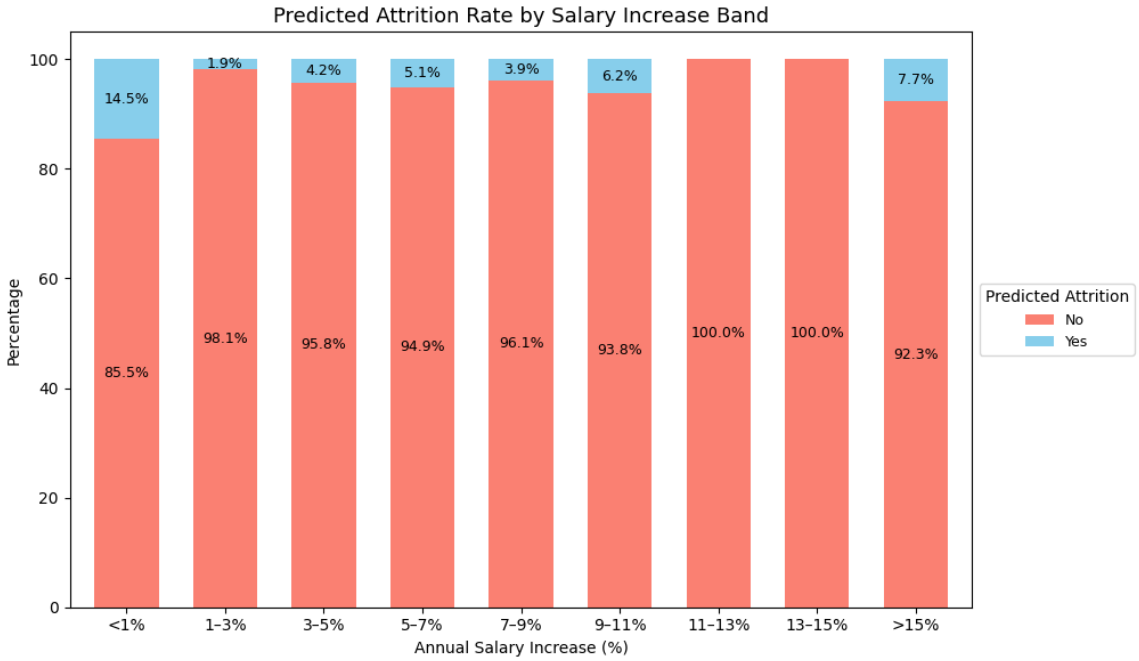
H.3 - Predicted Attrition Rate by Unit



H.4 - Predicted Attrition Rate by Loyalty Level

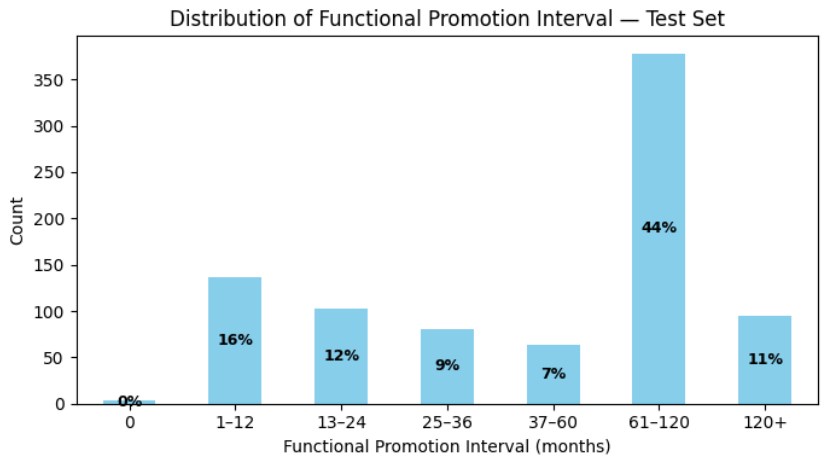


H.5 - Predicted Attrition Rate by Salary Increase

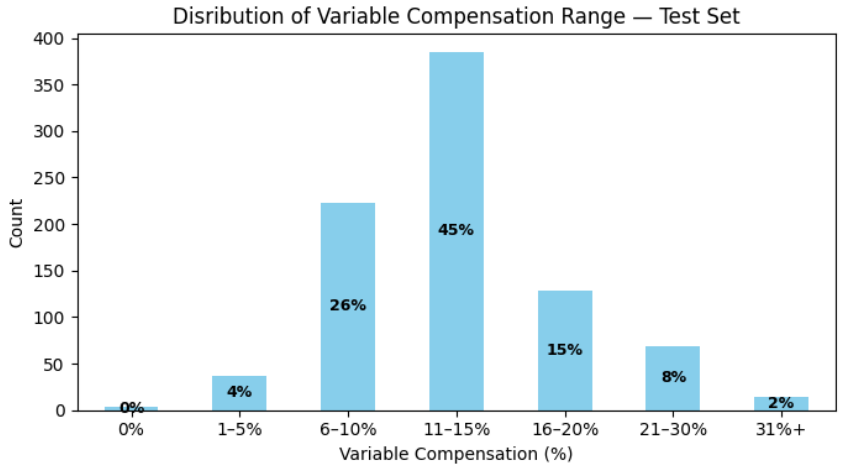


I. Distribution of Most Influential Features for the LightGBM Model (Test Dataset)

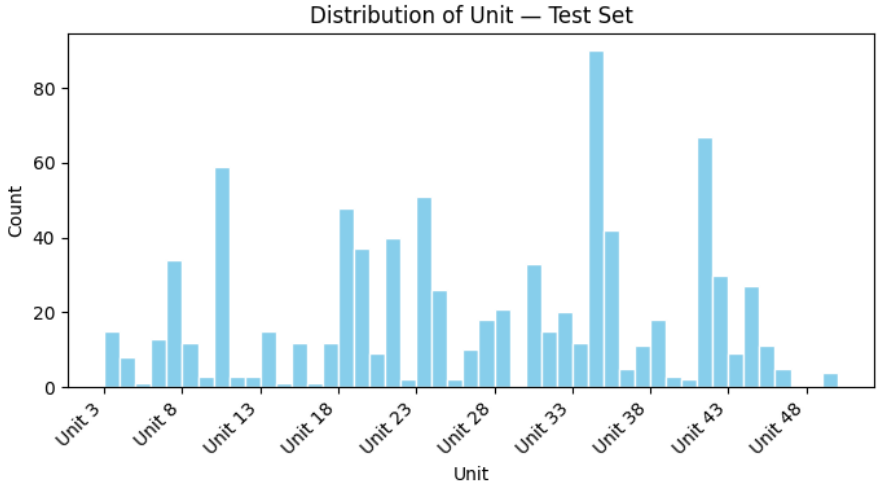
I.1 – Distribution of Functional Promotion Interval in the Test Set



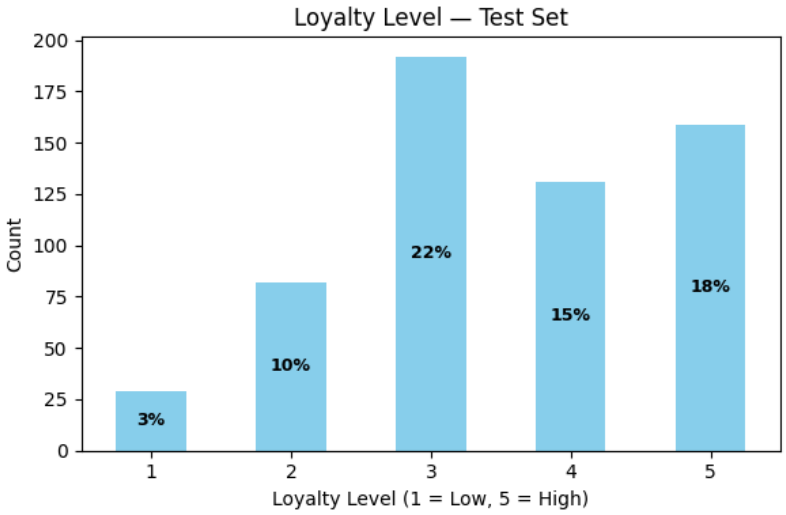
I.2 - Distribution of Variable Compensation in the Test Set



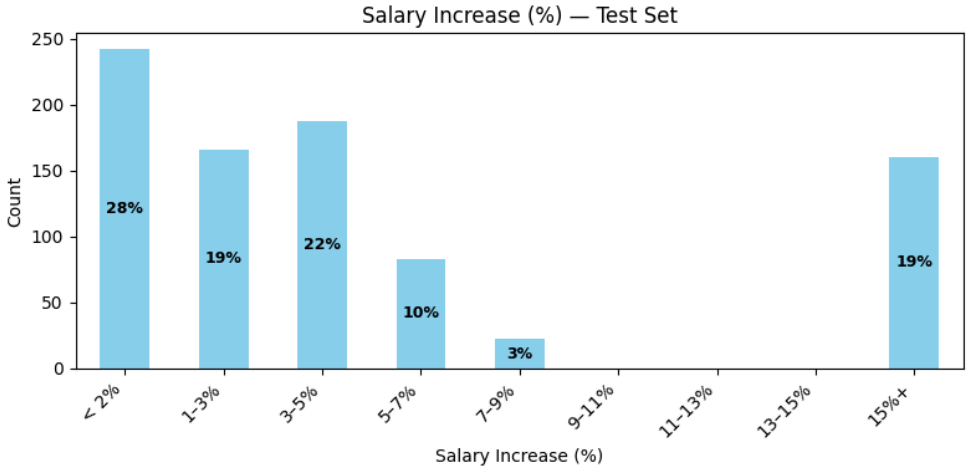
I.3 - Distribution of Unit in the Test Dataset



I.4 – Distribution of Loyalty Level in the Test Dataset



I.5 – Distribution of Salary Increase in the Test Dataset



J. Representative Employee Profiles (Best-performing LightGBM Tree)

Profile Id	Support In Test	Positives	Observed Attrition Rate (%)	Risk Vs Overall (%)	Average Predicted Risk (%)	Profile Rules
3	33	19	57.6	7.07	64.8	Functional_Promotion_Interval <= 106 AND Enablement_Level <= 4 AND NOT(Enablement_Level <= 3) AND Salary_Increase <= 0
9	17	4	23.5	2.89	18.5	NOT(Functional_Promotion_Interval <= 106) AND NOT(CFP_Not_Applicable <= 0) AND Functional_Promotion_Interval <= 118
4	88	16	18.2	2.23	31.1	Functional_Promotion_Interval <= 106 AND Enablement_Level <= 4 AND NOT(Enablement_Level <= 3) AND NOT(Salary_Increase <= 0)
10	12	2	16.7	2.05	6.9	NOT(Functional_Promotion_Interval <= 106) AND NOT(CFP_Not_Applicable <= 0) AND NOT(Functional_Promotion_Interval <= 118) AND Salary_Increase <= 0
2	150	18	12.0	1.47	21.2	Functional_Promotion_Interval <= 106 AND NOT(Enablement_Level <= 4) AND Loyalty_Level <= 4 AND Company_Car <= 0