Using Machine Learning to predict Employee Turnover: a case study of the Willis Towers Watson Lisbon Hub

Luísa Alexandra Sousa Pereira

152022027

Dissertation written under the supervision of
Professor Pedro Afonso Fernandes

Dissertation submitted in partial fulfillment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 03.01.2024.

Using Machine Learning to predict Employee Turnover

Luísa Pereira

Resumo

Atualmente, assiste-se a uma elevada competitividade empresarial, onde cada fator é determinante para garantir vantagem competitiva. Neste sentido, os Recursos Humanos assumem grande importância, sendo a componente essencial para garantir o sucesso de qualquer organização. Deste modo, manter funcionários altamente qualificados é um dos maiores desafios enfrentados na atualidade, sendo o *turnover* dos funcionários um problema caro e destrutivo para as empresas. O Lisbon Regional Delivery Hub (LRDH) da Willis Towers Watson (WTW) não é excessão, assistindo frequentemente à saída dos seus funcionários altamente qualificados.

Esta tese apresenta um duplo sentido: investigar os fatores que influenciam o turnover dos funcionários na Willis Towers Watson e determinar o modelo mais eficaz para prever esse turnover. Com este estudo foi possível concluir que funcionários com mais tempo de empresa, aqueles que foram promovidos, portugueses e indivíduos com cargos mais elevados são mais propensos a permanecer ativos na organização. Para além disso, o estudo também mostrou a eficácia dos modelos de machine learning, onde o modelo de random forest mostrou ser altamente capaz de prever o turnover dos funcionários da WTW.

Estes resultados fornecem insights valiosos para o departamento de recursos humanos da empresa, uma vez que agora é possível verificar quais são os fatores que mais influenciam a saída dos funcionários e, assim, adotar medidas de retenção dos mesmos. Ao abordar esses fatores, o LRDH da WTW, além de reduzir a rotatividade dos funcionários, pode melhorar seu desempenho produtivo e retorno financeiro.

Palavras-chave: *MachineLearning*; Gestão de Recursos Humanos; *Turnover* dos funcionários; Willis Towers Watson (WTW); Modelos de previsão

Using Machine Learning to predict Employee Turnover

Luísa Pereira

Abstract

Nowadays, the business world is characterized by high competitiveness, where every factor is essential for ensuring the success of a company. In this sense, Human Resources take on great importance, being the essential component for ensuring the success of any organization. Thus, retaining highly qualified employees is one of the biggest challenges faced today, with employee turnover being a costly and destructive problem for companies. Thee Lisbon Regional Delivery Hub (LRDH) of Willis Towers Watson (WTW) is no exception, with its highly qualified employees leaving annually.

This thesis has a dual purpose: to investigate the factors that influence employee turnover at Willis Towers Watson and to determine the most effective predictive model for predicting that turnover. This study found that employees with longer tenure, those who have been promoted, Portuguese nationals, and individuals in higher-level positions are more likely to remain active in the organization. Additionally, the study also showed the effectiveness of machine learning models, with the random forest model being highly capable of predicting employee turnover at WTW.

These results provide valuable insights for the company's human resources department, as it is now possible to identify the factors that most influence employee departures and, thus, adopt retention measures. By addressing these factors, the LRDH of WTW can not only reduce employee turnover, but also improve its productive and financial performance.

Keywords: Machine Learning; Human Resources Management; Employee Turnover; Willis Towers Watson (WTW); Predictive models

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to all those who have helped me with my thesis and throughout my academic journey.

First and foremost, I would like to extend my deepest appreciation to my advisor, Professor Pedro Afonso Fernandes, for all his support and guidance throughout the dissertation process. Thank you for your essential insights and for always being available to assist me.

I am also immensely grateful to Alexandra Ramada, Joana Valentim and António Navarro Caeiro, from the Willis Towers Watson Lisbon Hub team for their invaluable contributions to this project from its inception. Thank you for your consistent feedback, collaboration, and support.

I extend my sincere thanks to César Vieira from the Human Resources department for his timely assistance with the data set. Without his promptness, this study would have been impossible.

Special thanks to David Soares from the Continuous Improvement department of the WTW's Lisbon Hub for his unwavering support and guidance from the very beginning of this thesis. Thank you for being there from the start and for all your valuable advice.

Lastly, I would like to express my deepest gratitude to my family, particularly my parents and brother, for all their support throughout my academic career. Thank you for being my support and for always believing in my success. Without you, none of this would have been possible.

Thank you all for being a part of my journey.

Luísa Pereira

Contents

1 Introduction **1**

2 Literature Review **3**

2.1 Introduction to employee turnover 3

2.1.1 Definition of turnover and importance of managing employee turnover 3

2.1.2 Economic and organizational impact 3

2.2 Traditional models of Employee Turnover 4

2.3 Factors affecting employee turnover 5

2.4 Machine Learning in Human Resources: Challenges and Ethical considerations 6

2.5 Relevant research in machine learning for predicting employee turnover . . 7

2.6 Interventions in Human Resources to reduce turnover 9

3 Methodology **11**

3.1 Business understanding 11

3.2 Data Understanding and Exploratory Data Analysis 11

3.3 Data Preparation 16

3.4 Modelling 16

3.4.1 Logistic Regression 17

3.4.2 K-Nearest Neighbor (KNN) 18

3.4.3 Tree-based Models and the Decision Tree 19

3.4.4 Random Forest 20

3.5 Evaluation 21

3.5.1 Accuracy 21

3.5.2 Confusion Matrix 21

3.5.3 Area Under the ROC Curve (AUC-ROC) 23

3.6 Implementation 23

4 Results **25**

5 Discussion **32**

6 Conclusion **36**

Acronyms

HR Human Resources

ET Employee Turnover

WTW Willis Towers Watson

LRDH Lisbon Regional Delivery Hub

ML Machine Learning

LR Logistic Regression

KNN k-Nearest Neighbor

DT Decision Tree

RF Random Forest

SVM Support Vector Machine

List of Figures

1 Gender Distribution by Active/Non-active Employees 13
2 Age Distribution by Gender and Activity (Active/Non-active) 14
3 Active and Non-active Employees by Job Level 14
4 Number of Services (Non-Active/Active) 15
5 Comparison of Employee Promotion (Non-active/Active) 15
6 Confusion matrix 22
7 Feature Importance using random forest 25
8 Feature importance using logistic regression 27
9 Probability of Being Active over Seniority time 29

List of Tables

1 HR data features 12
2 Mean Decrease Gini results using random forest 26
3 Logistic regression coefficients 28
4 Comparison of the Metrics 30

1 Introduction

There are several reasons why individuals are considering leaving their current organizations, making it crucial to anticipate potential workforce attrition. In recent years, there has been a significant increase in research focusing on employee turnover. Currently, companies are highly competitive, and a loss of talent can impede a company's performance. Consequently, organizations increasingly invest in their human resources, providing training and development opportunities to thrive in their respective markets. Moreover, nowadays, the market presents abundant opportunities, and employees are always seeking to advance their professional careers. Thus, retaining highly skilled employees is one of the biggest concerns for contemporary organizations. A high employee turnover rate can lead to substantial cost increases and diminished productivity, compromising organizations' financial health and operational stability (Yigit and Shourabizadeh, 2017; Winne et al., 2019; Carmeli and Weisberg, 2006; Chowdhury et al., 2022).

In this sense, replacing an employee is an arduous, time-consuming process fraught with associated costs. Predicting in advance the factors that lead an individual to leave the organization can prove invaluable, as it allows pre-emptive retention measures to be implemented, averting the loss of highly qualified personnel. Consequently, using current and past employee data becomes a valuable asset to identify the most significant factors contributing to employee turnover. This information can be used in employee retention strategies (Yigit and Shourabizadeh, 2017; Carmeli and Weisberg, 2006; Chow et al., 2007).

In addition to the increasing number of studies on employee turnover, machine learning algorithms are increasingly used in human resources studies. Accurately predicting the factors influencing an individual's intention to leave their current employment can be challenging. However, the fusion of machine learning algorithms with employee data from a specific organization can be pivotal in precisely predicting these factors, facilitating the implementation of measures to counteract this desire (Garg et al., 2022; Andrews and Mohammed, 2020).

The Willis Towers Watson (WTW) Lisbon Regional Delivery Hub (LRDH) experiences a recurrent loss of highly skilled employees yearly. The company heavily invests in nurturing its human capital through constant training programs tailored to its employees' work domains. Consequently, employee turnover results in additional costs and erodes team productivity, requiring continuous readjustment and a clear understanding of human resources requirements for smooth operations. Accurately assessing employee turnover and the variables influencing it poses a challenge for the company's Human Resources (HR) department. Thus, it is essential to scrutinize these factors to enable HR to address turnover or even gain insights into staff requirements proactively.

Willis Towers Watson is a globally recognized consulting firm specializing in risk man-

agement, insurance, benefits, and human resources. The company is recognized for its analytical excellence and insights expertise, which aids organizations in making informed decisions. It offers consulting services and human capital solutions to organizations worldwide, helping them optimize benefits planning, manage risk, and improve talent management. In this sense, WTW's LRDH partners, with consulting teams across Great Britain and Western Europe, aim to deliver high-quality technical work using streamlined and efficient processes. The areas of expertise encompass actuarial valuations, UK-specific pension calculations, data auditing and cleaning, project management and service coordination, calculation automation, and configuration of specialist software. At WTW's Lisbon Hub, employees deliver on the promise of respect, trust, integrity, teamwork, and client focus every day. Given the high expertise of its workforce, retaining talent is imperative, making employee turnover prediction a central concern for the company.

The main objective of this thesis is to understand the factors influencing employee turnover at WTW's LRDH. Initially, supervised machine learning will be employed to predict the determinants of employee turnover. Subsequently, each model's accuracy will be evaluated, ensuring that the information delivered to the Human Resources department is precise and actionable, leading to effective retention strategies. Thus, this thesis focuses on the following research questions:

1. How can machine learning be used to predict employee turnover at Willis Towers Watson Lisbon Hub?

2. What are the key factors affecting employee turnover at the WTW Lisbon Hub?

3. How can machine learning be used to develop tools for employee retention strategies?

In conclusion, this study will focus mainly on WTW's LRDH, using internal organizational data to establish a robust framework for predicting turnover within the company, incorporating supervised machine learning algorithms.

2 Literature Review

2.1 Introduction to employee turnover

2.1.1 Definition of turnover and importance of managing employee turnover

Effective turnover management is a current concern of all competitive organizations, as human capital plays a pivotal role in determining an organization's long-term success (Lee and Mitchell, 1994).

Consequently, employee turnover (ET) is one of the biggest challenges faced by modern companies, is a fundamental concept in the field of human resources management, and has been subject to extensive study over the years (Chowdhury et al., 2022). Therefore, ET can be defined as the rate at which employees depart from an organization and are subsequently replaced by others, being a critical metric to understand the dynamics involved in managing the human capital of an organization (Lyons and Bandura, 2020; Allen and Meyer, 1990).

Turnover can be categorized as either voluntary or involuntary. The first occurs when an employee leaves the organization, while the second arises when the employer decides to terminate the contract with the employee (Ongori, 2007). According to Lyons and Bandura (2020), voluntary turnover can result from several factors, including job dissatisfaction, an absence of career advancement opportunities, ineffective management, low pay, and lack of recognition. Conversely, involuntary turnover can be attributed to factors such as poor performance, violation of company policies, or downsizing (Ongori, 2007).

2.1.2 Economic and organizational impact

Employee turnover carries adverse economic and organizational consequences, and it is crucial to understand the financial and operational implications of this phenomenon for organizations to formulate effective human resource management strategies (Becker and Huselid, 1998).

From an economic perspective, significant costs are incurred when a key and senior employee voluntarily departs from the company. These costs include direct costs of decreased productivity, costs associated with recruitment and training, and indirect costs, including loss of human capital, waste of organizational processes, and time to replace new employees (Wöcke and Heymann, 2012).

In organizational terms, employee turnover can exert repercussions that extend beyond the financial aspects (Becker and Huselid, 1998). High turnover rates can engender instability within the work environment, negatively affecting the organizational culture. This phenomenon can lead to a weaker employee connection to the company and reduced motivation to perform at their best (Lazzari et al., 2022). Furthermore, excessive employee turnover can hinder operational continuity, as remaining employees must frequently adapt

to new colleagues and processes, potentially impacting the quality of the work (Brenyah and Tetteh, 2016).

2.2 Traditional models of Employee Turnover

Over the years, research on employee turnover has significantly increased, resulting in a substantial number of articles and critical studies focused on ET. Some of the prominent studies in this area include those by Mobley (1977), Lee (1988), Lee and Mitchell (1994), Morrell et al. (2008), and Wöcke and Heymann (2012).

Mobley's model was particularly significant as it introduced the concept of the relationship between job dissatisfaction and the decision to leave a company. This model argues that job dissatisfaction is the primary determinant of voluntary turnover. When employees become dissatisfied with their work, they explore alternative opportunities, evaluate job offers, and assess the associated costs of leaving their current job (Lee and Rwigema, 2005; Wöcke and Heymann, 2012).

Lee (1988) model expanded upon Mobley's (1977) model, replacing the concept of job satisfaction with job commitment and involvement. This model concluded that these factors play a pivotal role in voluntary turnover (Lee, 1988).

The Unfolding Model, introduced by Lee and Mitchell (1994), offers an alternative perspective to understand the concept of voluntary turnover. This model suggests that the turnover process follows four distinct decision paths, influenced by shocks or natural processes that trigger change. These shocks can alter an individual's perception of their current employment, prompting them to reconsider their commitment. The first path involves a system shock that compels the worker to evaluate their situation based on personal characteristics and past experiences. The second path occurs when a system shock occurs but without prior work alternatives or past experiences. The third path arises when a system shock happens, but there is no recollection of past experiences. However, alternative job options are available. The fourth path occurs when there is no system shock, and the employee merely reevaluates their current situation and desire to remain in the company (Lee and Mitchell, 1994).

Lastly, the model proposed by Morrell et al. (2008), titled 'Mapping the Decision to Quit: A Refinement and Test of the Unfolding Model of Voluntary Turnover,' aimed to test the models presented by Lee and Mitchell (1994) and Lee at. Al. (1999). This model includes criticisms and suggests that the decision to leave the current job unfolds over time, with the employee going through a series of steps before making a final decision. However, this model proved to be more accurate than the previous ones (Morrell et al., 2008).

2.3 Factors affecting employee turnover

Before an individual formally departs from the organization they are employed by, there is an intention to do so, referred to as turnover intention. Consequently, before the actual departure from the company, an employee nurtures the desire to leave, which can occur due to several factors. These factors can be categorized into individual, organizational, and environmental factors (Lo, 2013).

Lo initiates by dividing the individual factors into two primary groups, namely job-related factors and individual characteristics, which are inherent traits of each individual. Job-related factors are the most studied factors, which include attributes such as autonomy, job feedback, job/task significance, skill variety, and task identity. Furthermore, organizational factors encompass the work environment, compensation and incentives, fairness procedures and distribution, and human resource strategies. Lastly, environmental factors pertain to all social factors and the prevailing labor market conditions (Lo, 2013).

Before the occurrence of turnover, the employee often experiences the intention to leave their current organization, referred to as "turnover intention," which significantly influences their behavior (Lo, 2013). Job satisfaction is a critical factor impacting turnover. Contented employees tend to have no intention to seek alternative employment and remain loyal to their current company (Bender and Heywood, 2006). Demographic variables are social categories for individuals and also play a substantial role in influencing job satisfaction (Sibiya et al., 2014). Among these demographic variables, age, gender, and education are particularly significant factors contributing to employee turnover (Wöcke and Heymann, 2012). However, it is crucial to exercise caution when utilizing human resources data for studies, as misusing algorithms can lead to uninformed decision-making (Chowdhury et al., 2022).

Gender is an influential factor in employee turnover, with studies indicating that male employees are more likely to leave their organization (Bender and Heywood, 2006; Lo, 2013; Wöcke and Heymann, 2012). Conversely, women responsible for caregiving in their families tend to find greater workplace satisfaction when flexible in balancing their family commitments. Consequently, married individuals with children are more interested in job stability, resulting in lower turnover (Bender and Heywood, 2006).

Another influential factor is employee age (Finegold et al., 2002; Nagadevara et al., 2008; Wöcke and Heymann, 2012). Finegold et al. (2002) suggest that senior workers are more committed to their organizations. The author argues that young technical employees, in the early stages of their professional careers, prioritize professional development and career advancement. In contrast, older workers prioritize job security and work-life balance (Finegold et al., 2002). Consequently, turnover is more prevalent among young, qualified workers than senior workers (Nagadevara et al., 2008; Wöcke and Heymann,

2012).

Education also plays a role in predicting an employee's intention to leave their current company (Bender and Heywood, 2006; Sibiya et al., 2014). Lo (2013) argues that education exhibits a negative relationship with job and career satisfaction, indicating that highly educated individuals are less satisfied with their current jobs and more desirous of career advancement (Lo, 2013; Pattie et al., 2006). Additionally, if their roles match their educational backgrounds, they are likelier to stay with the company (Pattie et al., 2006).

Tenure corresponds to the length of an employee's service in a specific company and role, which is another influential factor in employee turnover (Carmeli, 2003). According to Lo (2013), younger, single, highly educated individuals with shorter organizational tenure are more prone to turnover. Consequently, older individuals with longer years of service are more likely to continue working in the same company (Nagadevara et al., 2008). (Lee and Rwigema, 2005) propose that this phenomenon may be attributed to the challenges that older employees face in adapting to the new work environment.

Employees' job level is another factor affecting their commitment to the company. Consequently, individuals in higher positions with higher salaries tend to have less intention of leaving their current job (Bender and Heywood, 2006). Moreover, when companies prioritize employee benefits such as promotions and training, employees tend to exhibit an increased commitment to the organization, leading to reduced turnover (Chow et al., 2007; Lo, 2013).

Another factor that is linked to employee turnover is employee burnout. Employees experiencing high levels of burnout due to an increased workload are more likely to contemplate leaving their current job. Hence, the human resources team must take proactive measures to prevent employee burnout (Koo et al., 2020). Conversely, when employees feel that they matter to the organization and their opinions are highly valued, they are less inclined to apply for new jobs (Atouba, 2018).

2.4 Machine Learning in Human Resources: Challenges and Ethical considerations

Companies like Google and P&G are increasingly using sophisticated data collection technology and analysis to maximize the potential of their available talent pool. The key to thriving in talent analytics lies in having access to high-quality data (Davenport et al., 2010).

Machine Learning (ML) finds multiple applications in the realm of human resources, encompassing recruitment, selection, employee engagement, training and development, performance management, employee turnover prediction and team dynamics, and human resource allocation (Garg et al., 2022). Many organizations have transitioned from traditional human resource management models to analytical models that enable real-time

deployment adjustments based on current requirements in an ever-evolving market landscape (Davenport et al., 2010).

According to Garg et al. (2022), the use of machine learning in managing human capital within companies is still in its nascent stages but has garnered significant attention from technology researchers. The integration of ML in HR ensures swifter, more cost-effective, and more intelligent human resource management, enabling data-driven decision-making and enhancing efficiency.

In summary, human resource management is a labor-intensive and costly process. However, with the implementation of machine learning, many operations can be automated and streamlined to facilitate the optimization of available talent within an organization. It is important to emphasize that the goal is to improve decision-making, not to change the people involved in HR data processes (Garg et al., 2022).

In the contemporary landscape, many HR decisions are driven by algorithms, to reduce costs and enhance operational efficiency (Köchling and Wehner, 2020). Falletta and Combs (2021) define HR analytics as a *'proactive and systematic process for ethically gathering, analyzing, communicating and using evidence-based HR research and analytical insights to help organizations achieve their strategic objectives'.*

Leveraging human resources data can pose challenges, particularly concerning issues of privacy and compliance during data collecting and processing. Consequently, it is essential to recognize that HR professionals bear ethical responsibility for promoting fairness and justice for all employees within an organization (Falletta and Combs, 2021).

In conclusion, although organizations often rely on policies to reduce bias, it is essential to recognize that these policies are not neutral or unbiased (Köchling and Wehner, 2020). Therefore, using HR data must be approached with diligence, and companies must contemplate how they acquire and manage the data from their employees. This should be done based on evidence-supported practices and ethical principles prioritizing transparency and establishing trust (Falletta and Combs, 2021).

2.5 Relevant research in machine learning for predicting employee turnover

Despite the traditional methods mentioned above, there has been a notable increase in studies using Machine Learning to predict and mitigate employee turnover. Numerous studies have explored predictive analyses of employee turnover within specific companies or sectors. However, many of the published articles focus on variables like job satisfaction, career opportunities, benefits, and organizational commitment, among others (Lazzari et al., 2022; Kae et al., 2021; Sikaroudi et al., 2015; Punnoose and Ajit, 2016). These variables were typically measured through questionnaires and surveys, which will not be employed in this study because certain variables used in other studies are inaccessible for

inclusion in the presented model. Nevertheless, some variables will be examined, such as marital status, gender, education, and time worked in the organization, among others.

One of the recent studies conducted by Lazzari et al. (2022) employed both qualitative interviews and quantitative methods. In this study, the authors apply a questionnaire of 123 questions that inquires about contextual information, items related to several HR themes, and a target question, used to assess the employee's intention to leave the organization. The questionnaire was distributed to 18322 employees of a multinational company operating across 56 countries. The study employed various machine learning approaches, including Logistic Regressions (LR), K-Nearest Neighbor (KNN), Decision Trees (DT), Random Forests (RF), and Boosting. The findings indicated that logistic regressions and Boosting classifiers exhibited the highest predictive power for turnover intention (Lazzari et al., 2022).

Another similar study conducted by Sikaroudi et al. (2015), focused on a case study in Arak, province of Iran, involving a supplier of automobile manufacturing and products automotive parts. This study analyzed similar variables to the previous one but also evaluated technical skills, compatibility of the body with the job, working experience, and the number of job changes of employees, among others. Furthermore, it utilized two different databases, the first with data from employees who are in the company and the second with data from employees who left the company. The first group corresponds to about 80% of the database. Machine Learning models were used, such as decision trees, random forests, Support Vector Machine, and logistic regression. The study analyzed the feature importance and concluded that age and marital status are less important (25.8% and 25.4%, respectively). Conversely, it showed that factors such as a number of previous jobs and knowledge of working conditions and laws exhibited 100% importance in employees' intention to leave (Sikaroudi et al., 2015).

In 2021, Kae et al. (2021) conducted a study using a dataset that could contain the information referred to through different methods, such as surveys, interviews, web scraping tools, or datasets from online. In this study, machine learning models were used similar to the previous ones, such as logistic regression, Support Vector Machines, K-nearest neighbor, decision tree, Naive Bayes, and random forest. This study analyzed the variables affecting turnover and identified a strong correlation between overtime worked and turnover intention, indicating employee aversion to overtime. Conversely, it was also possible to conclude that employees with a higher tenure were less likely to leave the organization, showing a negative correlation between turnover intention and Total Working Years (Kae et al., 2021).

A more specific study was carried out by Punnoose and Ajit (2016) where they used an advanced Machine Learning algorithm called Extreme Gradient Boosting, comparing it with six other supervised classifiers (logistic regression, Naïve Bayesian, random forest, Support Vector Machine, Linear Discriminant Analysis, K-Nearest Neighbor). The

study's primary goal was to predict employee turnover in a particular level of store leadership team of a global retailer over 18 months. The findings indicated that the Boosting classifier outperformed the others in terms of accuracy, relatively low runtime, and efficient memory utilization (Punnoose and Ajit, 2016).

Lastly, the study proposed by Yigit and Shourabizadeh (2017) analyzed 30 company-related variables using various machine learning models, including logistic regression, random forest, decision tree, KNN, Naïve Bayes, and support vector machine (SVM). The study evaluated the performance of these models using classification metrics such as accuracy, precision, and recall. SVM outperformed the other models in terms of both accuracy and precision, with values of 0.914 and 0.51, respectively. However, logistic regression and random forest also demonstrated high accuracy, with values of 0.855 and 0.879, respectively. While logistic regression and random forest had slightly lower precision scores than SVM (0.35 and 0.45, respectively), the decision tree achieved the highest recall score of 0.43, outperforming SVM.

These studies underscore the potential of Machine Learning as a powerful tool in predicting employee turnover and understanding the contributing variables. Employing different algorithms enables organizations to predict the turnover of individuals in a given organization or sector. Nevertheless, as in traditional models, the use of Machine Learning for turnover prediction presents its own set of limitations and challenges, as outlined in these studies. These limitations may include difficulty in interpreting the results or the need for standardized and representative data. Thus, while using machine learning to predict turnover offers numerous advantages, it is essential to be aware of the existing limitations and actively work to address them (Lazzari et al., 2022; Kae et al., 2021; Sikaroudi et al., 2015; Punnoose and Ajit, 2016).

2.6 Interventions in Human Resources to reduce turnover

Employee turnover can result in several consequences for a company, some of which may be challenging to replace or rebuild. Therefore, it becomes crucial to accurately predict who intends to leave the organization in advance. This perspective ability empowers the human resources team to develop necessary strategies to mitigate this intention (Wang and Zhi, 2021). Consequently, reducing turnover is essential to minimize costs and enhance overall performance, necessitating the development of strategies aimed at reducing employee turnover (Andrews and Mohammed, 2020).

By adopting models capable of identifying the pivotal factors that are determinant in the intention of employees to leave the company, human resources can take measures that contradict this will. Such measures may include instituting rewards and recognition programs for employees aligned with the organization's objectives, offering increased opportunities for training and development, providing competitive salaries and benefits,

and fostering an environment where employees feel valued and heard, allowing them to voice their ideas (Andrews and Mohammed, 2020; Abdalkrim, 2012). When an employee perceives that their current company offers advantages over others, such as opportunities for career growth opportunities, their intention to leave the organization diminishes, and their commitment to the company strengthens (Abdalkrim, 2012).

For instance, Finegold et al. (2002) contends that younger technical workers highly value career opportunities. To retain such talents, companies must formulate appropriate strategies that offer career development prospects tailored to these employees' aspirations. Hence, it is imperative for the Human Resources Team to monitor variables that affect job satisfaction and employees' emotional commitment. Armed with this insight, HR team can take the necessary steps to reduce turnover, increasing the productivity of the employees (Carmeli and Weisberg, 2006; Chow et al., 2007; Winne et al., 2019).

3 Methodology

3.1 Business understanding

Every year, Willis Towers Watson invests substantially in employee training and has an active team to provide all the training sessions. Each employee who joins the organization undergoes an extensive 2-3 months of training sessions, incurring both time and considerable costs for the company. This study aims to identify the factors influencing turnover so that the Human Resources team can implement preventive strategies to retain qualified employees. In addition to the direct costs associated with training, there are additional expenses related to recruitment processes and the consequential loss of productivity resulting from the departure of highly skilled employees, among other contributing factors.

This study is conducted using a data sample from individuals employed at Willis Towers Watson's Lisbon Hub. The main department at the company's hub is the Retirement department, where a significant number of employees depart every year. It is crucial to understand the factors causing employees to leave so that these issues can be addressed proactively. Therefore, the dataset used in this study comes from a sample of employees who work or have previously worked in this department.

Thus, the main objective of this study is to use Machine Learning models to predict employee turnover. This approach enables swift action to prevent the loss of qualified employees and mitigate additional costs associated with managing human capital.

3.2 Data Understanding and Exploratory Data Analysis

As previously mentioned, the data used in this research were provided by the HR department of Willis Towers Watson's Lisbon Hub and covers the period between 2018 and 2023. The dataset contains 199 observations, 91 of which correspond to non-active employees, while 108 observations are from active employees. The features used in the study are presented in Table 1, along with their descriptions.

It is essential to note that there are five different branches for the 'Education Field' variable: 'Mathematics or Actuarial Science', 'Management, Economics or Administration', 'Engineering', 'Science, Physics and Chemistry' and 'Humanities and Languages'. Additionally, the Job Levels presented in the dataset are 48, 53, 58, 63 and 68. It is important to note that this variable appears in reverse order, indicating that a higher value corresponds to a lower employee Job Level. Other variables that are important for this study are Total Services and Total Projects. It is essential to understand that WTW employees may be involved in multiple Billing Groups (Services), and within these Billing Groups, there are several Billing Units (Projects), which represent various projects that an employee may have participated in within the organization between the years of 2018

and 2023.

Table 1: HR data features

No	Feature	Description	Data type
1	Employee code	Code that identifies the employee	Categorical
2	Active	1 if is an active member, 0 otherwise	Binary
3	Gender	1 if female, 0 if male	Binary
4	Age	Age of each employee	Numeric
5	Nationality	1 if Portuguese, 0 otherwise	Binary
6	Education field	Education field of the employee	Categorical
7	Job level	Level of the employee within WTW	Categorical
8	Marital status	1 if married, 0 otherwise	Binary
9	Children	1 if has children, 0 otherwise	Binary
10	Seniority	Tenure at WTW	Numeric
11	Promotion	1 if had a promotion, 0 otherwise	Binary
12	Distance to office	Number of kilometers between home and the Lisbon Hub of WTW	Numeric
13	Overtime	Average of the overtime worked, per year	Numeric
14	Training hours	Average of the training hours, per year	Numeric
15	Total Services	The maximum number of services worked	Numeric
16	Total Projects	The maximum number of projects within the services worked	Numeric

All the information in the dataset comes from private data provided by WTW's HR department, using an 'Employee Code' to guarantee the confidentiality of the information shared. In general, the dataset has no missing values or suspicious outliers.

Of the 199 observations in the dataset relating to WTW Lisbon Hub employees, approximately 45.7% are female, while the remaining 54.3% are male employees. The graph below illustrates that the difference between men and women is more pronounced among the employees who have already left the company (non-active employees).

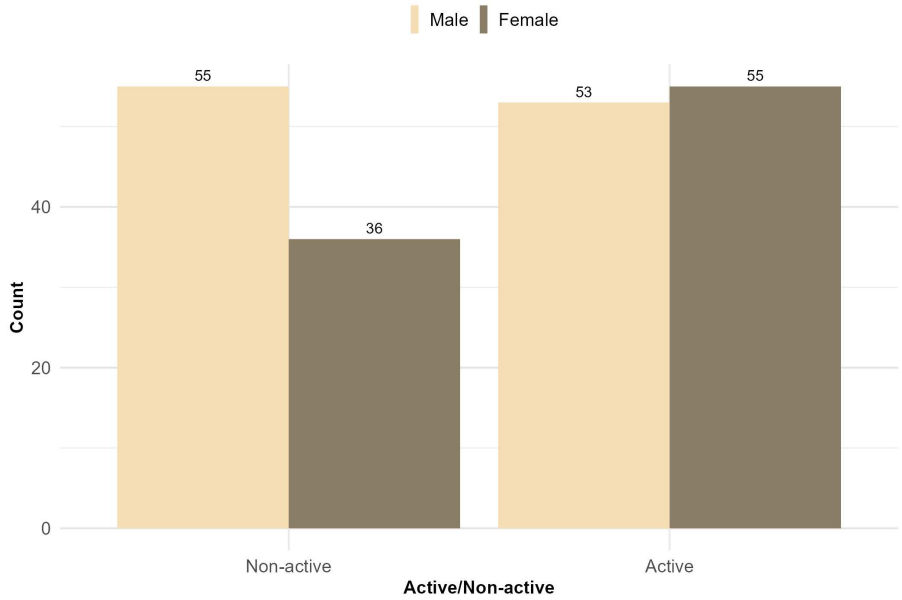

Figure 1: Gender Distribution by Active/Non-active Employees

Furthermore, by examining Figure 2, it is possible to analyze the age distribution by gender and employment status. At first glance, it is evident that most individuals are concentrated at younger ages, and the average age of inactive employees is higher than that of their active counterparts. This is normal, as employees typically leave the company at an older age than when they initially joined, resulting in an overall increase in age due to their tenure in the organization. Moreover, the average age of female employees in the company is also lower than that of males, suggesting that women tend to join earlier. The density of active women between the ages of 30 and 40 is higher than that of men, indicating that women seek more excellent professional stability and don't risk as many new career opportunities at this stage. Conversely, the concentration of men in the inactive group during this stage is notably higher, suggesting that they take more professional risks at this stage. Another common trend observed is the higher concentration of active men after age 40, potentially indicating that many women in this age group may opt to step back from their professional careers to the detriment of increased family responsibility.

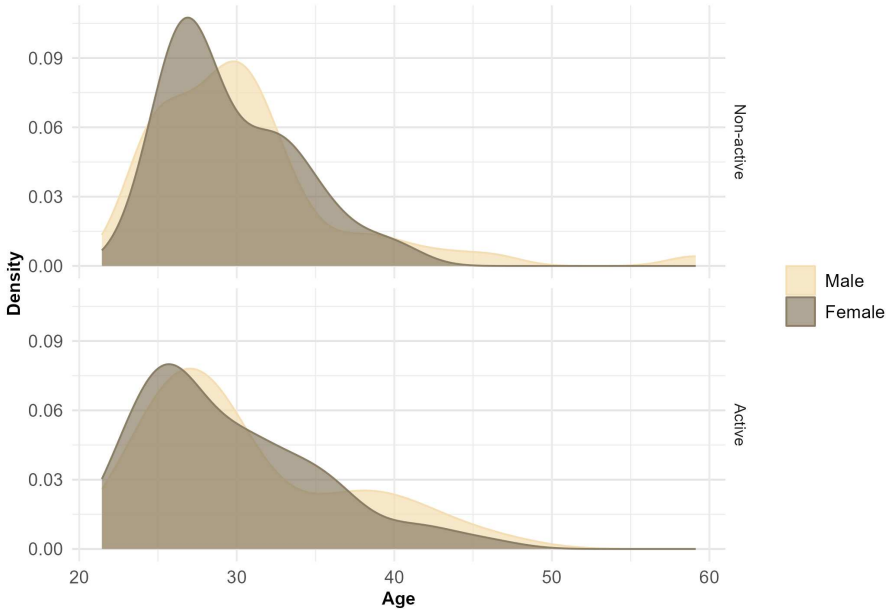

Figure 2: Age Distribution by Gender and Activity (Active/Non-active)

Analyzing the Job Level, the plot below indicates the concentration of cases regarding employees with a lower level in the company, specifically at level 68. This asymmetry results from the limited data available for employees with more tenure in the company and higher job levels. However, in the three intermediate levels, there are more active employees than non-active ones. Focusing on level 53, out of the 15 sampled employees, 14 are currently active, which may be a lower employee attrition rate at this organizational level.

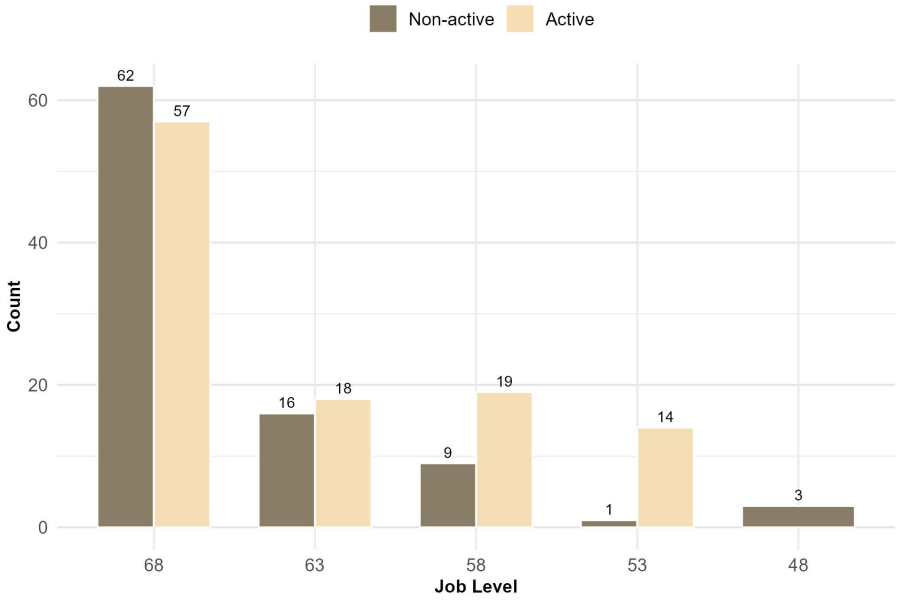

Figure 3: Active and Non-active Employees by Job Level

Another crucial variable in this study is 'Total Services' since it is essential to understand the impact of working in different Billing Groups on employees. Analyzing the figure below, there is little difference between active and inactive employees across different service levels. However, the most significant difference arises for a single service, possibly because there is a lot of information about newcomers to the company who have not yet had the opportunity to work and explore various Billing Groups.

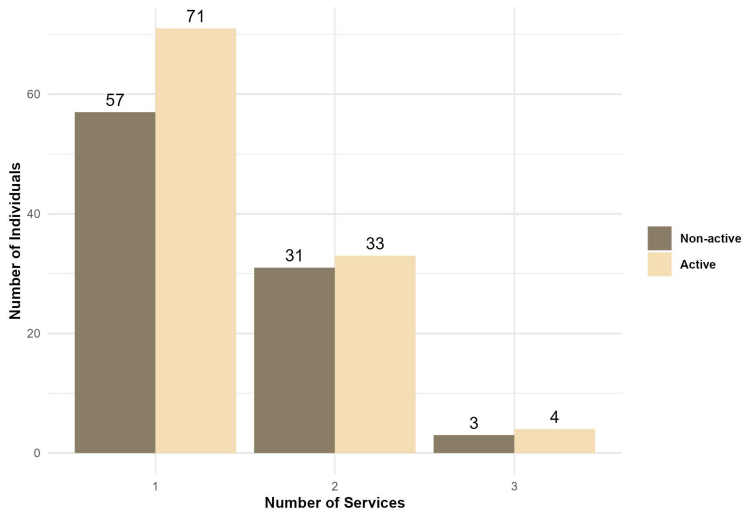

Figure 4: Number of Services (Non-Active/Active)

Finally, promotion is another crucial variable, as it is essential to understand whether promotion positively affects the intention to stay in the company. Analyzing the figure below, it is possible to see that for 'No promoted', there isn't much difference between Actives and Inactives. However, those who have been promoted are more active than non-active employees, suggesting that if an individual is promoted, they are more likely to stay in the company.

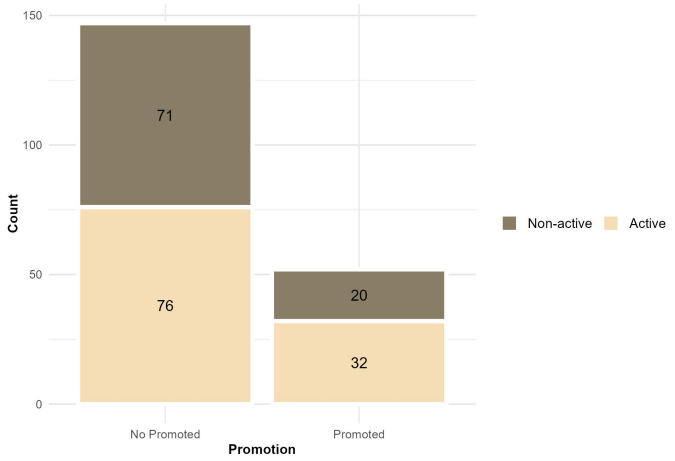

Figure 5: Comparison of Employee Promotion (Non-active/Active)

3.3 Data Preparation

For this study, the employee dataset underwent a data-cleaning process to ensure the reliability and quality of the models. In general, the analysis focused on missing values and possible outliers. This analysis revealed no missing values or suspicious outliers, and no changes or deletions were necessary.

Initially, the HR department provided the employees' postcodes to calculate the 'Distance to office' variable. For each postcode, coordinates were obtained for the approximate distance from the office to each employee's home. Additionally, the Education Field variable had to be converted into three separate columns: 'Education_field_Mathematics', 'Education_field_Management' and 'Education_field_Other_Distinct_Area'. Each of these dummy variables is assigned a value of 1 if an individual belongs to that field and 0 otherwise. After creating these three columns, the initial column representing each individual's Education Field was deleted, as it was no longer needed. The exact process was applied to the Job Level column, which was split into five columns based on the values in the original column. Similarly, the initial column was also eliminated. However, the 'Education_Field_Other_Distinct_Area' and 'job_level_68' columns were removed to address multicollinearity issues. In addition, the Employee Code column was also eliminated, as it only served to identify the employee's information and was no longer needed for the study.

To conduct predictive analytics studies, it is necessary to proceed with the preparation of the dataset to facilitate the training and evaluation of machine learning models. The changes made served to transform categorical variables into dummy variables, ensuring a more suitable format for statistical analysis or machine learning algorithms.

After implementing all the necessary changes to make the dataset operational for the study, it was essential to split it into train and test sets. To guarantee the reproducibility of the result, a random seed of 123 was set, ensuring that subsequent runs of the analysis would produce consistent results. Subsequently, the dataset was randomly divided into training and test sets, using the *'createDataPartition'* function from the *'caret'* library. It is worth noting that the variable used as the target variable for the predictive modelling was the 'Active' variable, which identifies whether the employee is active or inactive in the company. Furthermore, the dataset was split in a 70/30 ratio, indicating that 70% of the data was allocated to the training set and the remaining 30% to the test set.

3.4 Modelling

This thesis aims to predict employee turnover at Willis Towers Watson's Lisbon Hub using labelled data, so it makes sense to delve into the area of Machine Learning (ML) that addresses this type of prediction. In this context, supervised learning is the branch of Machine Learning that instructs machines through examples. During training, systems

are exposed to large amounts of labelled data, allowing them to learn to recognize clusters and patterns in the dataset, after being provided with enough examples (Heath, 2020). This type of machine learning uses labelled training data and a collection of training examples to infer a function (Sarker, 2021).

In a study, when the result is quantitative or numerical, the research is based on a regression problem. Conversely, when the result variable is categorical or binary, the research is based on a classification problem (James et al., 2013). Thus, classification is a supervised machine learning concept that categorizes data into classes, predicting a class label for a given example. Mathematically, it maps a function (f) of input variables (X) to output variables (Y) as a target, label, or category. The most common classification problems can be binary classification, multiclass classification, or multi-label classification, and there are several methods widely used in various application areas, such as Naïve Bayes, support vector machine, decision tree, random forest, among others (Sarker, 2021). In this research, four supervised machine learning models were used: Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF).

3.4.1 Logistic Regression

One of the most widely used models for classification problems is logistic regression, represented in the equation below (Wang and Zhi, 2021; Taddy, 2019). This supervised machine learning algorithm models binary response variables, taking values of 0/1 and yes/no. Even when the response of the interest is not binary, there are instances where decision-relevant information is binary. Examples of this include profit versus loss, and it is often simplest to think in binary terms (Bewick et al., 2005; Taddy, 2019). Thus, LR is a linear model for log odds, and the odds of an event are the probability that it happens over the probability that it doesn't (Taddy, 2019). In summary, with LR, the occurrence of a given event is predicted by fitting the data to logit function or log-odds. Therefore, this algorithm can also be used to predict employee turnover (Kae et al., 2021).

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

The fundamental logistic value, represented by the Euler number (e), transforms the values so that they always fall between 0 and 1. The coefficients of the logistic function are estimated through maximum likelihood, optimizing, and maximizing the probability of an observation being assigned to one of the two classes. Consequently, the graphical representation of the logistic function generates an S-shaped curve that never exceeds the limits of 0 and 1 (James et al., 2013). In logistic regression, the p-value is used to determine whether the relationship between the independent variables and the dependent variable occurs by chance. It's crucial to recognize that interpreting logit model coefficients is not straightforward, as it deviates from the linear framework. Unlike linear regression,

where coefficients represent the average change in the dependent variable for a one-unit shift in the independent variable, in logistic regression, the interpretation is not direct. To obtain more precise insights, the $logitmfx()$ function is employed to calculate partial derivatives of the probabilities relative to the independent variables. Therefore, even when the p-value is less than 0.05 (indicating statistical significance), it is essential to recognize the inherent complexity in directly interpreting logistic regression coefficients and seek additional methods to understand the relationship between the variables under study (Bevans, 2023).

3.4.2 K-Nearest Neighbor (KNN)

Linear regression is an example of a parametric approach since it assumes a linear functional form for *f(x)*. These types of methods are easy to adjust because only a small number of coefficients need to be estimated. In the case of linear regression, the coefficients have simple interpretations, and statistical significance tests can be efficiently conducted. However, if the functional form *f(x)* is far from the truth and prediction accuracy is the main objective, this method may perform poorly. In contrast, non-parametric methods do not explicitly assume a parametric form for *f(x)*, providing a more flexible approach to regression (James et al., 2013).

This flexibility gives rise to the K-Nearest Neighbor model, one of the simplest and most widely used non-parametric methods (James et al., 2013). The KNN model can be employed for both classification and regression. The model operates on the principle that similar input values produce similar output values. KNN selects a specified number *(K)* of neighboring data points and averages their values to make a prediction (Sarker, 2021; Chandra, 2023). This algorithm makes predictions based on the closest distance of the new example with the training dataset, earning it the classification of a lazy learner as it does not require learning from the data (Sisodia et al., 2017).

The KNN regression method is closely related to the KNN classifier. In KNN regression, when a value K and a prediction point x_0 are specified, the process begins by identifying the K training observations closest to x_0, which are represented by N_0. It then estimates $f(x_0)$ by calculating the average of all the responses from the training data found in N_0, as shown in the formula below. Thus, the predicted value of the target variable for the new data point is equal to the average of the target values for the K nearest neighbors, where this average process reflects the idea that similar data points tend to have identical target values (James et al., 2013).

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i \tag{2}$$

Sisodia et al. (2017) argues that KNN exhibits good performance metrics, such as high accuracy or precision, even though it may not be the ideal model compared to other

algorithms.

3.4.3 Tree-based Models and the Decision Tree

Tree-based models are another example of popular ML algorithm models that use one or more decision trees, representing another non-parametric method. Data scientists are increasingly using this model to visually and explicitly depict decisions and decision-making. This algorithm presents the representation of an inverted tree, as the root is positioned at the top. In this representation, each internal node signifies a decision based on a specific variable, while the edges connect the nodes, illustrating the decision flow, and the leaf nodes represent the final outcomes of the decision (Ross, 2021; Sisodia et al., 2017).

When the decision tree is successful, it begins from the root, converting the entire space of predictor variables and establishing conditions that observations must fulfill for the selected input variable, referred to as 'decision nodes.' It subsequently divides the observations into two subsets. This division process continues with each input variable, resembling branches called 'internal nodes', until reaching the final result, referred to as 'leaf nodes,' which meet specific anticipated criteria of the target variable. Each node represents the average of that particular subgroup (James et al., 2013; Taddy, 2019).

Ross (2021) argues that one of the primary advantages of using decision trees is their output's readability and interpretability without requiring statistical knowledge, making them valuable for non-technical audiences. Additionally, this model can process numerical, binary, and categorical predictor variables and can easily handle the presence of missing data. DTs can be analyzed quite easily, even by non-experts, especially if they are small (Friedman, 2006; James et al., 2013).

On the other hand, using decision trees has some disadvantages since they are unstable. This implies that a simple change in the observations can alter the tree's structure, causing a change in the ongoing predictive study. Another disadvantage arises as the data is partitioned, as decision trees tend to make predictions at terminal nodes with minimal observations. This disadvantage can increase when decision trees become excessively complex, with a lot of nodes, leading to over-fitting, where the model performs well on the training data but poorly on unseen data. This may occur because the tree may start to learn patterns that are unique to the training data, which may not be present in the real world. However, researchers can control this disadvantage by using more extensive datasets and implementing appropriate adjustments, such as pruning techniques, where unnecessary branches are removed to reduce complexity. In general, DTs are another example of a non-parametric method and do not achieve the same level of predictive accuracy as other classification models discussed in this study (Friedman, 2006; James et al., 2013; Taddy, 2019).

3.4.4 Random Forest

Random forest is another algorithm frequently employed for supervised machine learning problems, consisting of multiple decision trees rather than a single decision tree for classification. Sisodia et al. (2017) This method addresses the tendency of DTs to exhibit high variance in their prediction by using a more sophisticated bootstrapping method. This method involves randomly selecting a sample from a population and then re-sampling this sample. During this process, observations are replaced with unseen observations randomly selected from the population, or observations in the sample may even be duplicated (Taddy, 2019; Khan et al., 2018; Shin, 2020).

James et al. (2013) argue that the practice of creating several training sets from bootstrapping is called bootstrap aggregation, or bagging. The central idea of bagging is based on the assumption that each set of n independent observations (Z_1, Z_2, Z_3, ..., Z_n) has variance σ^2. Thus, the average variance of these $\overline{Z}$ observations can be expressed as σ^2/n, implying that averaging the observations reduces the total variance (James et al., 2013).

When making a prediction using random forest, there are two main concepts. Firstly, RF is created by random sampling of training data points, and, secondly, each decision tree in a forest considers a random subset of features. The prediction using RF is calculated as the average of all the individual decision trees, providing more accurate results compared to an individual DT. However, the more diverse the random forest, the more robust the overall prediction (Koehrsen, 2017).

In this manner, random forests improve performance compared to bagged trees by reducing the correlation between trees. Instead of using all the predictors in each split, RFs randomly select a subset of predictors, mitigating the high correlation between trees and enhancing the robustness of predictions. Typically, the number of predictors (m) used in each split is limited for each decision tree and is usually defined as $m \approx \sqrt{p}$, where m is approximately equal to the square root of the total number of available predictors(p). The primary goal of this approach is to prevent all decision trees in the set from becoming too similar, as that could result in selecting the same strong predictor variables for the initial splits.

Furthermore, in random forests, trees are pruned to make them more compact and effective. This, combined with the random selection of the m predictor variables, prevents model over-fitting (James et al., 2013). In the Sisodia et al. (2017) study, it was found that the use of RF provides greater accuracy and precision in predicting employee attrition compared to the other models used.

To evaluate feature importance in tree-based models, such as decision trees and random forests, the Mean Decrease Gini (MDG) index is used. The MDG index quantifies the average impurity decrease in all the tree nodes where a given feature is used. In this

context, the higher the value of this measure, the more influential the variable is for the model. In summary, the MDG index measures whether a given feature contributes to reducing impurity in the data and, consequently, enhances the model's accuracy (Menze et al., 2009).

3.5 Evaluation

Models can be evaluated using various metrics. However, it is necessary to understand which metric is the best for evaluating the model, as it depends on the problem being solved. Evaluation metrics are quantitative measures used to assess the performance and effectiveness of a statistical or machine learning model. These metrics provide insights into how well the model performs and aid in comparing different models or algorithms (Ghosh, 2023).

3.5.1 Accuracy

Classification accuracy is one of the most used evaluation metrics in research, and the term 'classifier accuracy' is commonly used to describe any general measure of classifier performance. Accuracy measures the number of correct decisions divided by the total number of decisions made by the classification model, being equal to $1 - \textit{error rate}$ (Provost and Fawcett, 2013).

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}} \tag{3}$$

Accuracy is widely used in data mining studies as an evaluation metric because it condenses classifier performance into a single number and is simple to measure. However, this measure is overly simplistic for applications of data mining to real business problems and entails some measurement challenges. To address these issues, it is necessary to disaggregate and account for the different types of correct and incorrect decisions made by a classifier, which can be achieved through the use of a confusion matrix (Provost and Fawcett, 2013).

3.5.2 Confusion Matrix

In classification model predictions, the confusion matrix can be constructed to illustrate the number of test cases correctly and incorrectly classified. This matrix adopts a format based on the number of outcome classes, n, resulting in an n x n table, where the actual classes are represented in the columns, while the model's predictions are represented in the rows, as depicted in the figure below (Ghosh, 2023).

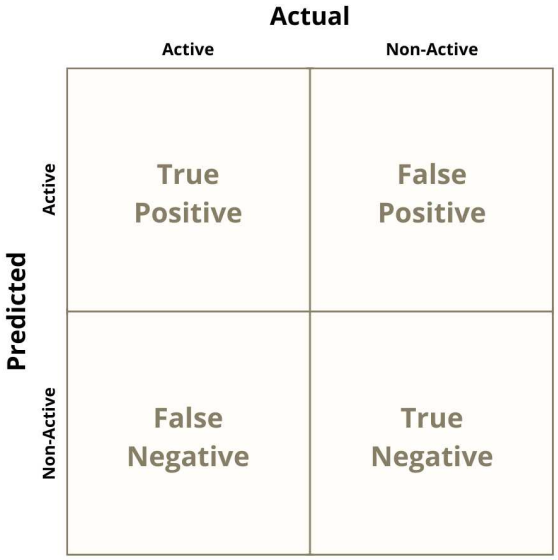

Figure 6: Confusion matrix

From this diagram, since n = 2 (Active and Non-active), the confusion matrix has dimensions two by two, where the true classes are represented across the top row of the diagram, while the predicted classes are listed down the side. Each class can be classified as either 'Active' or 'Non-active'. To summarise, the cell in the top-left corner is called True Positive (TP), representing the number of observations that have been correctly classified - meaning all the observations in the sample are predicted to be positive and are indeed positive. On the other hand, the False Positives (FP) are in the top-right corner, representing the number of observations that were incorrectly classified as belonging to the positive class by the model when, in reality, they belonged to the negative class. In the bottom-left corner are the False Negatives (FN), defined as the number of observations that were incorrectly classified as belonging to the negative class by the model. Finally, in the bottom-right corner are the True Negatives (TN), representing the number of observations that were correctly predicted as belonging to the negative class by the model when, in reality, they belonged to the positive class (Provost and Fawcett, 2013).

Precision: Precision measures the proportion of positive predictions that were correctly made by the model in relation to the total number of positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

This metric is crucial for avoiding false positives and ensuring confidence in the accuracy of positive predictions.

Recall or sensitivity: Recall measures the proportion of positive observations that were correctly predicted by the model in relation to the total number of observations that

are actually positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

This metric is important for identifying the majority of positive observations, even if this results in some false positives.

F1-Score: Finally, the F1-Score is useful when seeking a balance between precision and recall, providing a single metric that considers both false positives and false negatives.

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{6}$$

3.5.3 Area Under the ROC Curve (AUC-ROC)

Finally, the last evaluation metric to be analyzed in this study is the Area Under the ROC Curve (AUC-ROC). The ROC curve is a plot of the true positive rate (recall) against the false positive rate, illustrating the trade-off between the number of accurate positive classifications and the increasing number of false positives allowed. This metric is calculated by plotting the ROC curve and then measuring the area under the curve. The AUC-ROC can take values between 0 and 1, where the higher this value, the better the model's performance, i.e., the model has a high ability to distinguish positive and negative classes (Suresh, 2020; James et al., 2013; Ghosh, 2023).

In conclusion, the ROC curve represents a complete curve, not limited to a single number, providing refined details about the classifier's performance. An ideal ROC curve will closely follow the top-left corner, which means that the larger the AUC, the more effective the classifier. The higher the AUC-ROC value, the better the model is at predicting and correctly assigning samples to the appropriate classes (Suresh, 2020; Ghosh, 2023).

3.6 Implementation

One of the main objectives of this study is to understand the variables that influence turnover and their impact on the likelihood of an individual leaving WTW. To investigate the feature's importance and its effects, an initial analysis is conducted using two different models: logistic regression and random forest. These models were chosen because they have been used in previous studies to predict employee turnover, making them well-suited for evaluating the relationship between employee characteristics and the intention to leave the company.

Using LR, it is possible to observe the feature importance, along with the coefficients and p-values for each variable. This analysis enables the observation of both the significance of the variables and the relationship between the independent variables and the

probability of a particular employee remaining active. Additionally, the p-values allow an analysis of whether the variables are statistically significant for predicting turnover, while the coefficients provide valuable information on the direction and impact of a given variable. With random forest, it is also possible to determine the feature importance of the variables, as well as the Mean Decrease Gini index. The MDG index identifies the variables that significantly contribute to the reduction of impurity in tree nodes, indicating their predictive importance.

After identifying the most explanatory and important variables for predicting the turnover of WTW employees, the least significant variables were eliminated to develop the final predictive models. In a second analysis, the four different models proposed in this study were employed: logistic regression, K-Nearest Neighbor, decision tree, and random forest. Each model undergoes an evaluation process, where the confusion matrix and all the associated metrics are determined. Finally, an analysis is conducted to understand which is the best model for predicting employee turnover at Willis Towers Watson's Lisbon Hub.

4 Results

One of the main aims of this study is to examine the factors influencing employee turnover. This objective arises from the need to furnish accurate information to the Human Resources team at the Willis Towers Watson Lisbon Hub. Thus, the key goal is identifying the factors driving employees to depart the company and adopting essential strategies to mitigate this tendency. To scrutinize these variables, two different models were employed - random forest and logistic regression, which were used to extract valuable insights regarding the response variable, 'Active'. These models can handle different data types and offer unique advantages in modelling and establishing relationships between the variables under study.

In the tables and figures below is possible to see the results of the analysis. Beyond showcasing the results of the feature importance in RF, which can effectively handle many variables and observations, Table 2 presents the values of the 'Mean Decrease Gini' index, a metric derived from this model. The random forest analysis reveals that 'Seniority', 'Distance_to_office', 'Training_hours', 'Age', 'Overtime', 'Promotion', 'job_level_53', 'Total_Projects' and 'Gender' are the nine most influential variables in predicting turnover.

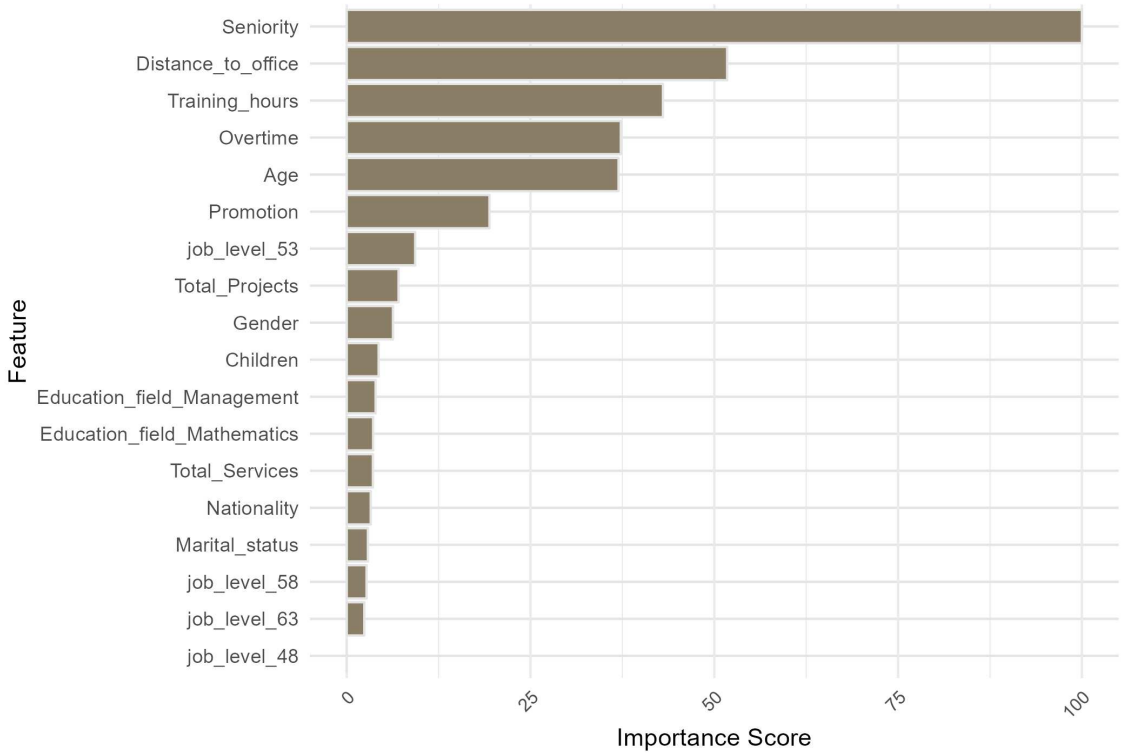

Figure 7: Feature Importance using random forest

Table 2: Mean Decrease Gini results using random forest

No	Feature	MeanDecreaseGini
1	Gender	2.366
2	Age	11.093
3	Nationality	1.775
4	Marital_status	1.253
5	Children	1.355
6	Seniority	23.189
7	Promotion	4.040
8	Distance_to_office	13.976
9	Overtime	10.663
10	Training_hours	12.617
11	Total_Services	1.672
12	Total_Projects	2.782
13	Education_field_Mathematics	1.749
14	Education_field_Management	2.304
15	job_level_48	0.277
16	job_level_53	2.620
17	job_level_58	1.476
18	job_level_63	1.338

In addition to the insights provided by the random forest model, a logistic regression was conducted to complement the results. Similar to the random forest model, the feature importance of the variables under study in logistic regression is illustrated in Figure 8. This model is useful for understanding the relationship between the independent variables and the probability of a given instance belonging to a specific class. Table 3 provides an analysis of the coefficients and p-values, offering a more detailed understanding of the contribution of each variable in predicting the response variable. Compared to the results of the feature importance score using RF, the logistic regression results highlight 'Seniority', 'job_level_53', 'job_level_58', 'Promotion', 'job_level_63', 'Nationality', 'Overtime' and 'Education_Field_Management' as the eight most crucial variables for turnover prediction.

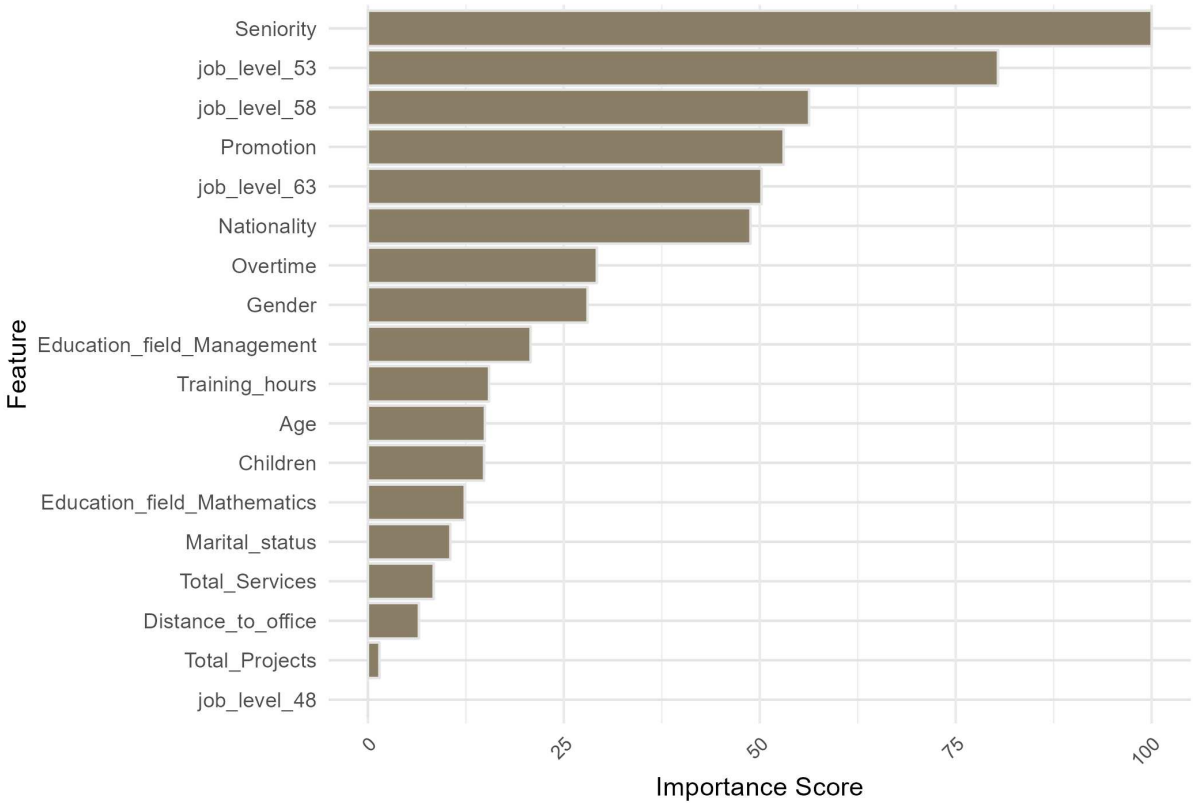

Figure 8: Feature importance using logistic regression

Table 3 shows the coefficients derived from the logistic regression using the $logitmfx()$ function, offering fundamental details on the direction and magnitude of the effect of each variable on the probability of turnover. This table provides the estimated results and corresponding p-values. Thus, the coefficients in the logistic regression results represent the estimated change in the likelihood of an employee remaining active in WTW for an increase of one unit in the predictor variable, keeping all others constant. The results show that the variables 'Seniority', 'Promotion', 'job_level_53', 'job_level_58' and 'job_level_63' have significantly low p-values (below 0.05), indicating that these variables have statistically significant effect on the probability of an employee to remain active in WTW. Conversely, variables with high p-values, such as 'Total_Services', 'Total_Projects', 'Education_field_Mathematics', and 'Education_field_Management', may indicate a lack of significant contribution to explaining variability in the response variable. Therefore, interpreting the p-values together with the coefficients reinforces the statistical validity of the model, and variables with significant p-values are essential for understanding the impact of employee characteristics on the intention of turnover.

Table 3: Logistic regression coefficients

No	Feature	Estimate	p-value
1	(Intercept)	0.738	0.683
2	Gender	0.586	0.138
3	Age	-0.047	0.427
4	Nationality	1.316	0.010
5	Marital_status	0.437	0.574
6	Children	0.868	0.430
7	Seniority	-0.981	0.0000001
8	Promotion	1.303	0.005
9	Distance_to_office	0.001	0.726
10	Overtime	0.018	0.122
11	Training_hours	0.0012	0.411
12	Total_Services	0.243	0.653
13	Total_Projects	0.036	0.932
14	Education_field_Mathematics	-0.364	0.510
15	Education_field_Management	-0.535	0.271
16	job_level_48	-10.604	0.994
17	job_level_53	9.166	0.00002
18	job_level_58	2.589	0.003
19	job_level_63	1.845	0.008

Focusing on the two models, a comparison of the different approaches reveals a consistency in some variables such as 'Seniority', 'Promotion', and 'job_level_53', both of which are important, but it is crucial to emphasize that each model offers a unique perspective. While random forest highlights the relative importance of the variables, logistic regression provides insights into the direction and magnitude of the effect of independent variables on the response variable. Using logistic regression, it becomes possible not only to identify which variables are important but also to understand the direction of the impact of each variable on the probability of the event of interest occurring - whether it is an employee staying or leaving the organization.

From the obtained results, it can be concluded that 'Seniority' is the variable with the most explanatory power in the model, and it is also a highly important variable in the study being carried out. In this sense, it is important to focus on it and understand the likelihood of WTW employees remaining active throughout their time at the company. An initial analysis shows that the variable's coefficient is negative in the logistic regression model. However, this does not generalize the total effect of the variable over the years, it only shows the effect in the first years of seniority. Thus, as the relationship between

the variables 'Seniority' and 'Active' is not strictly linear, the effect of seniority on the probability of being active can vary non-uniformly over time. It is, therefore, necessary to add the variable $Seniority^2$ to the logistic regression model, which will allow the model to capture non-linear patterns in the association between 'Seniority' and the probability of being active.

Through this, it is possible to verify the quadratic effect, where, initially, an increase in seniority reduces the probability of being active, but this effect decreases as the years of permanence in the company continue to increase. After adjusting the model, it was possible to calculate the predicted probabilities by using the coefficients from the model, which allowed controlling the process and understanding how each coefficient influenced the prediction. It is important to emphasize that when the predicted probabilities were calculated, their value was limited to the interval [0, 1]. Carrying out this analysis made it possible to analyze the results in a richer and more detailed way since the 'Seniority' variable is complex, and obtaining it through automated models can lead to errors and misinterpretation of its effect. With Plot 9, it is possible to understand the effect of this variable on the likelihood of a WTW employee remaining active.

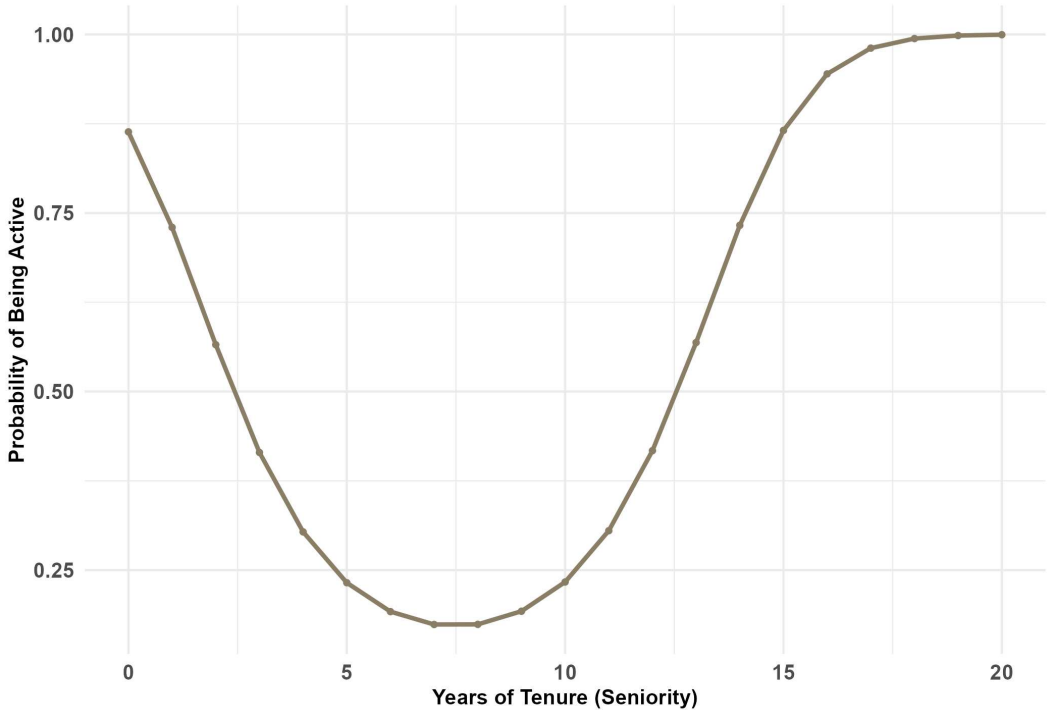

Figure 9: Probability of Being Active over Seniority time

In the plot, the x-axis represents the years of seniority, while the y-axis represents the probability of a particular WTW employee remaining active in the company. Analyzing the curve, it is possible to conclude that at the beginning of their career in the company, there is a tendency for WTW employees to leave the organization and that it

is between the 7th and 8th year of seniority that the probability of being active is lower. This means that between these years, the Human Resources department must adopt retention strategies to keep the employees in the organization. It is also possible to notice that an employee with 18 years of seniority onwards is less likely to leave the organization as they become committed to the organization and work environment. Analyzing the other variables, it can be seen that 'Age', 'Distance_to_office', 'Total_Projects', 'Education_field_Mathematics', 'Education_field_Management' and 'job_level_48' have a negative relationship with the probability of remaining active.

Furthermore, to develop a good predictive model, it is necessary to understand which variables to include and exclude to obtain accurate results with minimal redundancy, avoiding unnecessary complexity and the risk of over-fitting. Having said this, by carefully examining the results of the feature importance in the models and the p-values obtained, the following variables were incorporated into the prediction models: 'Seniority', 'Promotion', 'Nationality', 'job_level_53', 'job_level_58' and 'job_level_63'.

To do this, four different models were tested and trained: LR, KNN, DT, and RF. When analyzing the different models for predicting employee turnover, some differences were observed in the performance of each one. To this end, various metrics were analyzed to provide a unique perspective on the performance of the models. The results obtained are summarised in the table below.

Table 4: Comparison of the Metrics

Models	Accuracy	Precision	Recall	F1Score	AUC-ROC
Logistic Regression	0.746	0.742	0.767	0.754	0.745
K-Nearest Neighbor	0.763	0.742	0.793	0.767	0.763
Decision Tree	0.644	0.655	0.633	0.644	0.644
Random Forest	0.780	0.774	0.800	0.787	0.779

Analyzing the accuracy, it is evident that random forest outperforms other models, achieving an approximate accuracy of 78%, followed by KNN with 76.3% and logistic regression with 74.6%. This suggested that RF has a superior ability to make accurate predictions. In terms of precision, which represents the model's ability to correctly identify positive cases, RF also excels, yielding 77.4%, followed by KNN and LR with 74.2%. This good performance indicates their comparable ability to minimize false positives. Recall/sensitivity measures the ability to capture all positive instances, again favoring RF with 80%, closely followed by KNN with 79.3%. The F1-score, which balances precision and recall, highlights random forest's superior performance at 78.7%, showcasing a well-rounded approach to handling both false positives and false negatives. Finally, analyzing the AUC-ROC metric, which measures the model's ability to distinguish between classes, random forest once again takes the lead with a value of 77.9%, indicating strong

discriminatory power.

In conclusion, considering all the metrics, the random forest model emerges as the preferred choice, displaying robustness in accuracy, precision, recall, and a balanced approach to false positives and false negatives. Consequently, LR and KNN also exhibit competitive performance across several metrics. Therefore, RF is the model that offers comprehensive performance in predicting turnover, meaning that it can more accurately predict which employees are most likely to leave the company. Conversely, the DT model displays the poorest overall performance with low accuracy, recall, F1-score, and AUC-ROC, which can be potentially attributed to its lower ability to capture the complexity of the data. Thus, the poor performance of the decision tree model may be due to over-fitting caused by its complexity.

5 Discussion

The results obtained in this study provide valuable insights into the factors that influence employee turnover at Willis Towers Watson's Lisbon Hub. Logistic regression and random forest were studied to understand the significance of these variables. Alongside presenting feature importance figures for these models, the 'Mean Decrease Gini' index metric was applied for the random forest, while p-values were obtained from the logistic regression. The first metric facilitated an evaluation of the relative importance of each variable in the model's decision-making process, providing valuable insights into which factors significantly contribute to reducing impurity in the tree's nodes. On the other hand, the second metric offered a detailed perspective on the relationship between independent variables and the probability of an instance belonging to a specific class, aiding in understanding each variable's influence on predicting the response variable.

Analyzing the variables in detail, it was evident that variables such as seniority, promotion, nationality, and job level are statistically significant. A closer examination of the seniority variable, which represents the employee's tenure at the company, the logistic regression results showed that the coefficient of this variable is negative. This implies that keeping all the other variables constant, with an increase of one year of seniority, the probability of a particular member remaining active at Willis Towers Watson decreases. However, as previously analyzed, the relationship between this variable and the 'Active' variable is not strictly linear. Therefore, a more detailed analysis was necessary to understand the influence of this variable on the likelihood of remaining active over the years. The result of this study showed that between the 7^{th} and 8^{th} year of work at WTW, the probability of remaining active begins to grow, and until the 7^{th} year of work, the probability of leaving the company is higher. This finding aligns with studies by Nagadevara et al. (2008), where they discovered that older individuals with more time working for a particular company had a greater tendency to remain there. Other authors have also reached similar conclusions, stating that the tenure variable negatively affects an employee's intention to leave the company (Carmeli, 2003; Lo, 2013; Nagadevara et al., 2008; Lee and Rwigema, 2005).

Other variables that demonstrated to be statistically significant were promotion and job_level variables, which correspond to whether an employee has been promoted or not and the organizational grade they occupy, respectively. It is important to note that a lower numerical value of Job Level corresponds to a higher organizational grade for employees. The logistic regression results showed that the coefficient of the promotion variable is positive, implying that employees who have been promoted are more likely to remain active. Analyzing the job level results, it was possible to conclude that individuals holding job levels 53, 58, and 63 exert a significant positive impact on the probability of remaining active. This suggests that individuals at these job levels have a substantially higher

likelihood of remaining active compared to individuals at other levels. The outcomes of this study align with the findings of Bender and Heywood (2006), Chow et al. (2007), and Lo (2013), indicating that being promoted and occupying higher job levels contribute to employees feeling more valued and having a reduced tendency to leave the company. Consequently, employees who have been promoted are more likely to feel valued by the company and have more opportunities for internal professional growth. Similarly, individuals with higher job levels typically have higher salaries and better benefits and may face more challenges and responsibilities, making them more satisfied with their work.

Another noteworthy variable is Nationality. The results of the logistic regression indicated that being Portuguese has a positive impact on an employee's likelihood to remain at WTW's Lisbon Hub. Consequently, foreign employees are more prone to leaving the organization. This could be attributed to various factors, including language. Despite the organization's use of English as the primary language, the prevalence of Portuguese speakers may create communication challenges for foreign employees, hindering their integration. In addition, local employees may have greater residential stability, positively influencing their commitment to staying with the company.

While other variables under study did not exhibit statistical significance, and the logistic regression model did not find sufficient evidence to establish a relationship between these variables and the likelihood of staying in the organization, they may still be relevant to WTW. The study concluded that male employees have a greater tendency to leave the organization compared to female employees. This aligns with the findings of studies by Bender and Heywood (2006), Lo (2013), and Wöcke and Heymann (2012), suggesting that women, typically bearing greater family responsibilities, are less inclined to take professional risks.

Additionally, to analyze the 'Age' variable, it was necessary to make the same interpretation as was done for the 'Seniority' variable. Analyzing this variable, it was possible to conclude that employees aged 35 or less have a greater tendency to leave the organization. From the age of 35, WTW employees tend to remain active. The results obtained align with the studies proposed by Finegold et al. (2002), Nagadevara et al. (2008), and Wöcke and Heymann (2012), which conclude that turnover is lower for older employees. Analyzing the results, it was evident that married employees and those with children are more likely to remain active. This was also found in the studies of Bender and Heywood (2006), who concluded that married individuals with children tend to prioritize professional stability.

It was also found that a higher number of services is associated with a higher probability of an employee staying active. This suggests that the more integrated they are in the organization, and the more challenges they have, the more the tendency for the employee to feel valued and useful in the company increases. This correlation is also supported by Atouba (2018), who concluded that individuals tend to stay longer in companies when

they perceive their opinions and work are valued.

Additionally, it was observed that overtime worked, and training hours positively affect the likelihood of staying with the organization. Therefore, individuals who work longer and have more training hours are more likely to remain active. The first observation contradicts the study proposed by Koo et al. (2020), while the second aligns with the findings of studies proposed by Chow et al. (2007) and Lo (2013). The discrepancy in the first case may suggest that the WTW's employees do not experience excessive workloads, diminishing the significance of overtime worked in influencing the intention to leave or stay with the organization.

Finally, the distance to the office variable measures the distance from an employee's home to the company office. This variable is also not significant, but by analyzing the results of the logistic regression model, it can be concluded that a higher distance to the office is associated with a lower probability of being active. However, it's essential to note that this variable may not be particularly relevant to the study, as employees work on a hybrid basis, so distance doesn't have much influence on their decision to stay or leave the organization.

As mentioned earlier, one of the main objectives of this study was to understand the variables that affect turnover with the aim of providing insights to the human resources department to mitigate this trend. Consequently, reducing turnover is fundamental to ensure the productivity of the teamwork and reduce the costs associated with recruitment Andrews and Mohammed (2020). By analyzing the proposals for intervention measures to reduce turnover put forward by Carmeli and Weisberg (2006), Chow et al. (2007), Winne et al. (2019), and looking at the results obtained, it can be concluded that the human resources department should adopt the following measures:

- As the seniority variable is one of the most important features and considering that the likelihood of leaving the organization is higher between the 7^{th} and 8^{th} years of employment, the Human Resources team should give special attention to these individuals. In this way, implementing benefits for employees with six years in the company could be considered to enhance their satisfaction within the organization. Additionally, a professional development program could be introduced for employees with tenures between 6 and 13 years. This program aims to help employees acquire new skills to keep up to date with the latest market trends, fostering motivation and a sense of belonging within the company, and not desire to change to another company.

- Considering that promoted employees are more likely to stay with the company, the HR department may enhance internal promotion practices, fostering a sense of fairness and offering professional growth opportunities to all employees.

- As job level proved to be another explanatory variable, the HR department should review the company's salary structure to ensure equitable compensation for employees at different levels. This approach not only promotes fairness but also diminishes the intention of junior or senior employees to leave the company.

- Nationality has also emerged as a significant variable in predicting turnover, where foreign employees are more likely to leave the organization. In response, the HR department can implement new integration programs specifically designed for foreign employees. These programs can help them adapt to the company culture, fostering a sense of integration.

Finally, by analyzing the proposed predicted models, it becomes evident that the obtained results align with existing studies on predicting employee turnover. The study proposed by Yigit and Shourabizadeh (2017) compared the classification methods and revealed that in terms of accuracy, random forest led with a score of 0.88, followed by K-Nearest Neighbors with 0.87, and logistic regression with 0.86. Regarding precision, RF again excelled with a precision of 0.45, followed by KNN with 0.38 and LR with 0.35. For recall, LR demonstrated a score of 0.31, while KNN and random forest showed 0.23 and 0.22, respectively. Overall, RF exhibited strong performance, while DT had poor performance in most metrics, as was concluded in this study.

Concerning the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), the study by Lazzari et al. (2022) concluded that logistic regression was among the best models. In this study, it was possible to conclude that RF achieved a value of approximately 0.78, followed by KNN with 0.76 and LR with 0.75. Another study by Punnoose and Ajit (2016) reported an AUC-ROC of 0.79 for RF, while KNN and LR scored 0.52 and 0.66, respectively, reiterating that RF performs the best. Finally, in study 37, it was found that for small datasets, random forest and logistic regression exhibited the best performance in terms of AUC-ROC (0.96 and 0.93, respectively).

In conclusion, this study reinforced the effectiveness of machine learning models as powerful tools for predicting employee turnover. However, there are still several limitations, including the absence of standardized and representative data, emphasizing the need to consider these constraints in the decision-making process (Lazzari et al., 2022; Kae et al., 2021; Sikaroudi et al., 2015; Punnoose and Ajit, 2016).

6 Conclusion

The aim of this study was twofold: to investigate the factors influencing employee turnover at Willis Towers Watson's Lisbon Hub and to determine the most effective predictive model for forecasting employee turnover.

As a first step, all the variables under study were analyzed using two different machine learning models - random forest and logistic regression. These models provided valuable insights into the key factors driving employee departures and the potential effectiveness of machine learning in predicting turnover. The analysis revealed that seniority, promotion, nationality, and job level were the most significant variables influencing WTW's employee turnover. Additionally, the study identified some measures that could be implemented by WTW's HR department to counteract the tendency for certain groups of individuals to leave the company.

The analysis revealed that employees with longer tenure, those who had been promoted, Portuguese nationals, and individuals holding higher job levels were more likely to remain active in the organization. Conversely, age, distance to the office, and training hours did not exhibit significant statistical relationships with turnover.

In a second step, after identifying the key factors, the four predictive models proposed in the study - LR, KNN, DT, and RF - were analyzed using only the variables that proved to be explanatory in the previous analysis. This selection enabled the development of more coherent and explanatory models for predicting employee turnover at WTW. This analysis also demonstrated the effectiveness of machine learning models in predicting employee turnover. Among the models tested, random forest exhibited the best performance, reaching remarkable scores of up to 77% across the various metrics - accuracy, precision, recall, F1-score, and AUC-ROC.

In conclusion, the study not only identified the variables that most influence employee turnover at WTW's Lisbon Hub but also demonstrated the value of machine learning models in predicting turnover. However, the study has some limitations, including the relatively small sample. Additionally, as the company is in a growth phase characterized by the hiring of recent graduates and master's degree holders, it resulted in a sample heavily skewed toward this age group. Furthermore, another limitation arises from the impossibility of obtaining survey responses from employees who have already left the company, which precluded the consideration of job satisfaction in the study. Lastly, the limited number of explanatory variables served as another limitation.

Future research directions include expanding the data sample, incorporating additional variables, exploring employee job satisfaction through questionnaire responses, and evaluating more predictive models. By addressing these limitations, a more comprehensive understanding of employee turnover at WTW's Lisbon Hub can be achieved.

References

G. M. Abdalkrim. The impact of human resource management practices on organizational performance in saudi banking sector. European Journal of Business and Management, 4:188–196, 2012.

N. J. Allen and J. P. Meyer. The measurement and antecedents of affective, continuance and normative commitment to the organization. Journal of Occupational Psychology, 63:1–18, 3 1990.

K. S. Andrews and T. Mohammed. Strategies for reducing employee turnover in small- and medium-sized enterprises. Westcliff International Journal of Applied Research, 4: 57–71, 11 2020. ISSN 2572-7176. doi: 10.47670/wuwijar202041KATM.

Y. C. Atouba. Tackling the turnover challenge among it workers: Examining the role of internal communication adequacy, employee work participation, and organizational identification. Communication Reports, 31:174–187, 9 2018. ISSN 0893-4215. doi: 10.1080/08934215.2018.1497180.

B. E. . Becker and M. A. . Huselid. High performance work systems and firm performance: A synthesis of research and managerial implications. Research in Personnel and Human Resources Management, 16:53–101, 1 1998.

K. A. Bender and J. S. Heywood. Job satisfaction of the highly educated: The role of gender, academic tenure, and earnings. Scottish Journal of Political Economy, 53: 253–279, 5 2006. ISSN 0036-9292. doi: 10.1111/j.1467-9485.2006.00379.x.

R. Bevans. Understanding p-values — definition and examples, 6 2023.

V. Bewick, L. Cheek, and J. Ball. Statistics review 14: Logistic regression. Critical Care, 9:112, 2005. ISSN 13648535. doi: 10.1186/cc3045.

R. S. Brenyah and E. N. Tetteh. Organisational culture and its impact on employee retention: Evidence from the private tertiary education sector of ghana. European Journal of Business and Management, 8:47–52, 2016.

A. Carmeli. The relationship between emotional intelligence and work attitudes, behavior and outcomes. Journal of Managerial Psychology, 18:788–813, 12 2003. ISSN 0268-3946. doi: 10.1108/02683940310511881.

A. Carmeli and J. Weisberg. Exploring turnover intentions among three professional groups of employees. Human Resource Development International, 9:191–206, 6 2006. ISSN 1367-8868. doi: 10.1080/13678860600616305.

R. Chandra. Python knn: Dominando a regressão de vizinhos mais próximos com sklearn, 8 2023.

C. W. Chow, K. Haddad, and G. Singh. Human resource management, job satisfaction, morale, optimism, and turnover. International Journal of Hospitality Tourism Administration, 8:73–88, 4 2007. ISSN 1525-6480. doi: 10.1300/J149v08n02_04.

S. Chowdhury, S. Joel-Edgar, P. K. Dey, S. Bhattacharya, and A. Kharlamov. Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover. The International Journal of Human Resource Management, 34:2732–2764, 2022.

T. H. Davenport, J. Harris, and J. Shapiro. Competing on talent analytics. Harvard Business Review, 88, 10 2010.

S. V. Falletta and W. L. Combs. The hr analytics cycle: a seven-step process for building evidence-based and ethical hr analytics capabilities. Journal of Work-Applied Management, 13:51–68, 4 2021. ISSN 2205-2062. doi: 10.1108/JWAM-03-2020-0020.

D. Finegold, S. Mohrman, and G. M. Spreitzer. Age effects on the predictors of technical workers' commitment and willingness to turnover. Journal of Organizational Behavior, 23:655–674, 8 2002. ISSN 0894-3796. doi: 10.1002/job.159.

J. H. Friedman. Recent advances in predictive (machine) learning. Journal of Classification, 23:175–197, 9 2006. ISSN 0176-4268. doi: 10.1007/s00357-006-0012-4.

S. Garg, S. Sinha, A. K. Kar, and M. Mani. A review of machine learning applications in human resource management. International Journal of Productivity and Performance Management, 71:1590–1610, 5 2022. ISSN 1741-0401. doi: 10.1108/IJPPM-08-2020-0427.

S. Ghosh. The ultimate guide to evaluation and selection of models in machine learning, 9 2023.

N. Heath. What is machine learning? everything you need to know, 12 2020.

G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Second edition, 6 2013.

A. C. Kae, C. XinYing, J. O. Victor, and K. K. Wah. Employee turnover prediction by machine learning technique. Journal of Telecommunication, Electronic and Computer Engineering, 13:49–56, 2021.

S. S. Khan, A. Ahmad, and A. Mihailidis. Bootstrapping and multiple imputation ensemble approaches for missing data. 2 2018.

W. Koehrsen. Random forest simple explanation, 12 2017.

B. Koo, J. Yu, B.-L. Chua, S. Lee, and H. Han. Relationships among emotional and material rewards, job satisfaction, burnout, affective commitment, job performance, and turnover intention in the hotel industry. Journal of Quality Assurance in Hospitality Tourism, 21:371–401, 7 2020. ISSN 1528-008X. doi: 10.1080/1528008X.2019.1663572.

A. Köchling and M. C. Wehner. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. Business Research, 13:795–848, 11 2020. ISSN 2198-3402. doi: 10.1007/s40685-020-00134-w.

M. Lazzari, J. M. Alvarez, and S. Ruggieri. Predicting and explaining employee turnover intention. International Journal of Data Science and Analytics, 14:279–292, 9 2022. ISSN 2364-415X. doi: 10.1007/s41060-022-00329-w.

G. J. Lee and H. Rwigema. Mobley revisited: dynamism in the process of employee turnover. The International Journal of Human Resource Management, 16:1671–1690, 9 2005. ISSN 0958-5192. doi: 10.1080/09585190500239333.

T. Lee and T. Mitchell. An alternative approach: The unfolding model of voluntary employee turnover. Academy of Management Review, 19:51–89, 1 1994.

T. W. Lee. How job dissatisfaction leads to employee turnover. Journal of Business and Psychology, 2:263–271, 1988. ISSN 0889-3268. doi: 10.1007/BF01014043.

J. Lo. The information technology workforce: A review and assessment of voluntary turnover research. Information Systems Frontiers, 17:387–411, 4 2013.

P. Lyons and R. Bandura. Employee turnover: features and perspectives. Development and Learning in Organizations,, 34:1–4, 2020.

B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10:213, 12 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-213.

K. Morrell, J. Loan-Clarke, J. Arnold, and A. Wilkinson. Mapping the decision to quit: A refinement and test of the unfolding model of voluntary turnover. Applied Psychology, 57:128–150, 1 2008. ISSN 0269-994X. doi: 10.1111/j.1464-0597.2007.00286.x.

V. Nagadevara, V. Srinivasan, and R. Valk. Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques. Research and Practice in Human Resource Management, 16:1–27, 12 2008.

H. Ongori. A review of the literature on employee turnover. African Journal of Business Management, pages 049–054, 6 2007.

M. Pattie, G. S. Benson, and Y. Baruch. Tuition reimbursement, perceived organizational support, and turnover intention among graduate business school students. Human Resource Development Quarterly, 17:423–442, 2006. ISSN 10448004. doi: 10.1002/hrdq.1184.

F. Provost and T. Fawcett. Data Science for Business. 7 2013.

R. Punnoose and P. Ajit. Prediction of employee turnover in organizations using machine learning algorithms: A case for extreme gradient boosting. International Journal of Advanced Research in Artificial Intelligence, 5:22–26, 2016.

M. Ross. Decision tree visualization: A quick tutorial using python for beginners (like me) to construct a decision tree and visualize it, 9 2021.

I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2:160, 5 2021. ISSN 2662-995X. doi: 10.1007/s42979-021-00592-x.

T. Shin. What is bootstrap sampling in machine learning and why is it important?, 7 2020.

M. Sibiya, J. H. Buitendach, H. Kanengoni, and S. Bobat. The prediction of turnover intention by means of employee engagement and demographic variables in a telecommunications organisation. Journal of Psychology in Africa, 24:131–143, 3 2014. ISSN 1433-0237. doi: 10.1080/14330237.2014.903078.

A. M. E. Sikaroudi, RouzbehGhousi, and A. EsmaieeliSikaroudi. A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). Journal of Industrial and Systems Engineering, 8:106–121, 2015.

D. S. Sisodia, S. Vishwakarma, and A. Pujahari. Evaluation of machine learning models for employee churn prediction. pages 1016–1020. IEEE, 11 2017. ISBN 978-1-5386-4031-9. doi: 10.1109/ICICI.2017.8365293.

A. Suresh. What is the auc — roc curve?, 11 2020.

M. Taddy. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions. 1st edition, 8 2019.

X. Wang and J. Zhi. A machine learning-based analytical framework for employee turnover prediction. Journal of Management Analytics, 8:351–370, 7 2021. ISSN 2327-0012. doi: 10.1080/23270012.2021.1961318.

S. D. Winne, E. Marescaux, L. Sels, I. V. Beveren, and S. Vanormelingen. The impact of employee turnover and turnover volatility on labor productivity: a flexible non-linear approach. The International Journal of Human Resource Management, 30:3049–3079, 11 2019. ISSN 0958-5192. doi: 10.1080/09585192.2018.1449129.

A. Wöcke and M. Heymann. Impact of demographic variables on voluntary labour turnover in south africa. The International Journal of Human Resource Management, 23:3479–3494, 1 2012.

I. O. Yigit and H. Shourabizadeh. An approach for predicting employee churn by using data mining. pages 1–4. IEEE, 9 2017. ISBN 978-1-5386-1880-6. doi: 10.1109/IDAP.2017.8090324.