Universidade do Minho
Escola de Engenharia

Diogo Gonçalves Graçoeiro

The Importance of Workforce Planning and Turnover Prediction in Retail: Leveraging People Analytics for Strategic Decision Making

Abril de 2025

Universidade do Minho
Escola de Engenharia

Diogo Gonçalves Graçoeiro

The Importance of Workforce Planning and Turnover Prediction in Retail: Leveraging People Analytics for Strategic Decision Making

Dissertação de Mestrado
Mestrado em Engenharia e Gestão Industrial

Trabalho efetuado sob a orientação da
Professora Maria do Sameiro Carvalho
Professor João Nuno Gonçalves

Abril de 2025

ACKNOWLEDGMENTS

As I reach the culmination of my academic path, I am overwhelmed with a profound sense of accomplishment and gratitude. Completing this thesis means the end of a remarkable chapter in my life and represents the fulfillment of a long-held aspiration. This transformative journey has enriched me with invaluable experiences.

A first word to the University of Minho, where I pursued my degree, for providing me with a comprehensive education and fostering an environment of academic excellence. The knowledge and skills I acquired during my studies have been pivotal in shaping my career path. I would like to express my heartfelt thanks to Professor Maria Carvalho and Professor João Gonçalves, my supervisors at the University. Their unwavering support, valuable guidance and insightful feedback have significantly influenced the direction and quality of my research.

Foremost, I would like to express my deepest gratitude to LTPlabs, the company where I had the privilege to complete my internship. Their unwavering support and guidance throughout my internship were instrumental in shaping my skills and knowledge in the field. I am immensely thankful to all the LTPeers for their warm welcome, collaboration and shared expertise, which greatly contributed to my professional growth.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

A Importância do Planeamento da Força de Trabalho e da Previsão de Rotatividade no Retalho: Alavancar People Analytics para a Tomada de Decisão Estratégica

RESUMO

O presente trabalho propõe uma abordagem inovadora para o planeamento estratégico de recursos humanos no setor do retalho, através do desenvolvimento de modelos preditivos de *turnover* com base em técnicas de People Analytics. A elevada taxa de turnover neste setor, aliada à necessidade de uma gestão mais eficiente da força de trabalho, justificou a criação de duas soluções distintas, ambas baseadas no algoritmo XGBoost: um modelo de previsão a um ano e outro a cinco anos.

O modelo de um ano tem como objetivo estimar a probabilidade de saída voluntária de cada colaborador com base nas suas características individuais. Já o modelo de cinco anos é desenvolvido a partir de dados agregados, considerando combinações de características como função, localização e tipo de loja, bem como perfis-tipo de colaboradores com atributos semelhantes. Esta estrutura permite realizar projeções estratégicas de longo prazo em contextos de maior incerteza. Para assegurar a aplicabilidade prática das previsões, o modelo utiliza apenas variáveis e categorias cujo valor é conhecido ou estimável à partida.

Ambos os modelos demonstraram capacidades preditivas relevantes, destacando-se a importância de variáveis como remuneração fixa, senioridade e idade. No entanto, verificaram-se enviesamentos sistemáticos em alguns segmentos — especialmente nos grupos com menor senioridade — e uma utilização limitada das avaliações de performance e potencial, fatores que foram analisados criticamente. As previsões desenvolvidas nesta dissertação destinam-se a alimentar um modelo de otimização já existente, reforçando o seu potencial para apoiar decisões estratégicas de recursos humanos.

Com base nos resultados obtidos, o estudo apresenta também um conjunto de propostas para investigação e desenvolvimento futuro, que incluem a revisão da forma como os perfis-tipo de colaboradores são definidos, a incorporação da mobilidade interna e o reforço da interpretabilidade local do modelo. Este projeto representa um contributo prático e relevante para a adoção de abordagens *data-driven* na gestão de talento, aproximando os recursos humanos da estratégia organizacional.

PALAVRAS-CHAVE

People Analytics; Retalho; Turnover; Machine Learning; Planeamento Estratégico de Recursos Humanos

ABSTRACT

This dissertation proposes an innovative approach to strategic workforce planning in the retail sector through the development of predictive turnover models based on People Analytics techniques. The high turnover rate in this sector, combined with the need for more efficient workforce management, justified the creation of two distinct solutions, both using the XGBoost algorithm: a one-year prediction model and a five-year forecast model.

The one-year model aims to estimate the probability of voluntary employee exit based on individual characteristics. The five-year model, on the other hand, is built from aggregated data, considering combinations of features such as role, location and store type, as well as employee-type profiles with similar attributes. This structure enables long-term strategic projections in contexts of higher uncertainty. To ensure practical applicability, the model relies only on variables and categories whose future values are known or can be reliably estimated in advance.

Both models demonstrated relevant predictive capabilities, with fixed compensation, seniority and age standing out as key drivers. However, systematic biases were identified in some segments — particularly among early-career groups — and performance and potential evaluations showed limited influence, which were critically assessed. These predictions are designed to serve as input for an existing workforce optimization model, supporting more informed hiring and retention planning over time.

Based on the results, the study also presents a set of proposals for future research and development, including a revision of how employee groups are defined, the incorporation of internal mobility and enhanced local interpretability of the model. This project provides a practical and relevant contribution to the adoption of data-driven approaches in talent management, bridging human resources and organizational strategy.

KEYWORDS

People Analytics; Retail; Turnover; Machine Learning; Strategic Workforce Planning

CONTENTS

LIST OF FIGURES

List of Tables

Acronyms

AI – Artificial Intelligence.

AUC – Area Under the Curve.

AUC-ROC – Area Under the Receiver Operating Characteristic Curve.

FTE – Full-Time Equivalent.

GBM – Gradient Boosting Machine.

GBT – Gradient Boosting Tree.

GLM – Generalized Linear Model.

HR – Human Resources.

HRIS – Human Resource Information Systems.

HRM – Human Resource Management.

KNN – K-Nearest Neighbors.

LDA – Linear Discriminant Analysis.

LR – Logistic Regression.

MAE – Mean Absolute Error.

ML – Machine Learning.

NN – Neural Network.

RF – Random Forest.

RNN – Recurrent Neural Network.

RQ – Research Question.

SHAP - SHapley Additive exPlanations.

SMOTE – Synthetic Minority Oversampling Technique.

SVM – Support Vector Machine.

VBM – Value Based Management.

XGBoost – Extreme Gradient Boosting.

1. Introduction

The present dissertation is within the realm of the Master in Industrial Engineering and Management at the University of Minho. The thesis was developed during an internship at LTPLabs, an analytical management consultancy company, as part of a project with one of its clients. The intent of this chapter is to give a general overview of the dissertation's subject and extent, problems that are handled, project's objectives and structure of the entire document.

1.1 Motivation

Employee turnover remains a persistent challenge in the retail sector, affecting operational efficiency, customer service quality and overall business performance. High turnover rates lead to increased recruitment and training costs while also disrupting store operations, particularly in frontline roles where experience and continuity are critical for maintaining service standards (Olubiyi et al., 2019). Despite advancements in Human Resources (HR) technologies — such as workforce analytics tools and talent management platforms — workforce planning in many organizations continues to rely heavily on empirical knowledge rather than data-driven decision-making, limiting the ability to anticipate workforce needs effectively (Majam & Jarbandhan, 2022).

Traditional workforce planning is often reactive, adjusting staffing levels only after turnover has occurred. In contrast, People Analytics enables a proactive approach, leveraging historical workforce data — such as tenure, role transitions and demographic indicators — to develop predictive models capable of forecasting turnover patterns (Polzer, 2022). By integrating predictive analytics into workforce planning, companies can improve long-term staffing strategies, optimize hiring processes and ensure business continuity (Majam & Jarbandhan, 2022).

Beyond workforce planning, organizational culture also influences turnover dynamics. Olubiyi et al. (2019) emphasize that Person-Organization Fit significantly impacts job satisfaction and employee commitment, which in turn affects retention. Employees who align with company values and perceive a strong sense of belonging are more likely to remain, particularly in customer-facing environments where engagement directly influences service quality. Additionally, their study highlights that turnover is driven not only by financial incentives but also by factors such as leadership quality, career development opportunities and workplace culture.

While existing research explores various factors influencing turnover, fewer studies have examined how predictive analytics can be systematically applied to workforce planning. Machine learning models are increasingly being adopted in HR decision-making, offering greater accuracy in identifying workforce trends and informing strategic planning (Heidemann et al., 2024). However, the ethical implications of Artificial Intelligence (AI)-driven HR analytics must also be considered to ensure that predictive models are implemented in a responsible and transparent manner that supports both organizational and employee interests (Giermindl et al., 2022).

This dissertation addresses this gap by developing a predictive turnover model tailored to the retail sector. Rather than focusing on individual retention strategies, the study aims to support strategic workforce planning by forecasting employee departures using machine learning techniques. The predictive model provides annual turnover estimates, allowing for a more data-driven approach in staffing decisions. By identifying key workforce variables and analyzing turnover trends, this research offers a structured methodology for integrating predictive analytics into HR decision-making. The findings will help HR professionals anticipate workforce gaps, optimize hiring needs and improve cost efficiency in workforce management.

1.2 Project Background

This dissertation was developed during a consulting project for a major Portuguese company operating in the retail sector, specifically in the consumer goods segment. The company is a key player in both the national and international market, managing a large portfolio of stores and offering a diverse range of products, including groceries, household items and personal care products. With an extensive store network and a high transaction volume, the company serves millions of customers annually across different locations, requiring careful workforce management to maintain operational efficiency.

Given the dynamic nature of consumer demand and the complexity of managing a large workforce, ensuring strategic workforce planning is a key challenge. For this study, the focus was placed on three primary store formats that constitute the core of the company's operations:

- Large hypermarkets, characterized by high transaction volumes and requiring a large, stable workforce;
- Medium-sized supermarkets, which balance product variety and efficiency, necessitating precise workforce planning;
- Neighborhood convenience stores, where agility and workforce flexibility are critical for operations.

Despite structured workforce planning efforts, employee turnover remains a major challenge, affecting workforce stability, service quality and operational costs. Currently, turnover predictions are based on empirical methods, relying on historical workforce variations rather than predictive analytics. This approach lacks analytical depth and limits the company's ability to anticipate workforce gaps proactively. Addressing this gap in turnover forecasting is essential for enhancing workforce planning and optimizing resource allocation, ensuring that stores remain well-staffed to meet business demands.

1.3 Project Objectives and Expected Results

The primary objective of this dissertation is to develop a turnover prediction model using advanced machine learning techniques to support strategic workforce planning. Given the company's reliance on historical workforce variations to estimate turnover, this model seeks to introduce a data-driven approach that incorporates both intrinsic (e.g., employee attributes) and extrinsic (e.g., market conditions) factors.

The model will generate year-by-year turnover predictions over a five-year period, providing granular insights into employee departure probabilities. Unlike the existing empirical approach, which primarily uses Full-Time Equivalent (FTE)-based estimations, this model will allow the company to proactively address workforce gaps, particularly in key operational functions.

From the second year onwards, the model does not aim to predict individual exits but rather forecasts the expected number of voluntary departures aggregated by employee profiles, role cluster, store type and region — supporting group-level workforce planning instead of employee-level forecasting.

The expected outcome of this dissertation is the development of a high-precision turnover prediction model, capable of identifying patterns in employee attrition and providing actionable insights to support HR planning and decision-making. This predictive capability will not only enable HR and management teams to develop targeted retention strategies and adjust recruitment plans, but will also serve as a key input for an already existing workforce optimization model developed by the company or project team. Through this integration, the predicted turnover rates will support more efficient workforce structuring, helping to balance hiring, internal mobility and promotion decisions across different employee segments and geographical locations.

In addition to model development, this dissertation also aims to reflect on how predictive tools and People Analytics can address the challenges of managing voluntary turnover in high-variability environments, such as retail and how such tools can support more effective, data-driven workforce strategies.

To guide the investigation and structure the narrative across the different chapters, the following research questions (RQs) are proposed:

- RQ1: What are the main individual and contextual factors that influence voluntary employee turnover in the retail sector?
- RQ2: How can predictive analytics be used to forecast voluntary turnover and support strategic workforce planning in high-turnover environments?

1.4 Thesis Outline

The present dissertation is organized into six main chapters, each contributing to a structured understanding of the turnover prediction problem in the retail sector and the development of a data-driven solution. Chapter 1 introduces the thesis by presenting the motivation, project background, objectives and expected results, as well as outlining the document structure. Chapter 2 provides a literature review that contextualizes the issue of employee turnover and highlights the relevance of workforce planning and predictive analytics. It also explores traditional and modern approaches to turnover prediction, including machine learning methods, with special emphasis on the XGBoost algorithm. Chapter 3 describes the problem addressed in this project, beginning with an analysis of the current decision-making context, identifying the limitations of the existing approach and providing a detailed overview of the available data and exploratory analysis. Chapter 4 outlines the methodological approach adopted to develop the predictive models. It includes details on data preparation, cleaning, processing and structuring, followed by the deployment of two models: a one-year individual-level model and a five-year aggregate-level model. Chapter 5 presents the results obtained from each model. It evaluates their predictive performance and interprets the most important variables driving turnover, based on their relevance in the XGBoost framework. Chapter 6 delivers a critical analysis of the models' strengths and limitations, followed by a set of short-term and medium/long-term recommendations for future research and development. The dissertation concludes with references and an appendix that includes additional analyses and data used to support the work.

2. Literature Review

This chapter provides a structured literature review to support the development of a predictive model for employee turnover in the retail sector. It begins by clarifying the concept of turnover and its organizational implications, followed by a review of traditional approaches and the key intrinsic and extrinsic factors that influence voluntary turnover. The chapter then explores how workforce planning can benefit from predictive analytics, highlighting both its strategic relevance and practical challenges. Finally, various machine learning techniques are analyzed, with a special focus on decision tree models such as XGBoost, to evaluate their suitability for turnover prediction in high-variability environments. This review lays the theoretical foundation for the methodology adopted in this dissertation.

2.1 Introduction to Employee Turnover

Employee turnover refers to the rate at which employees leave an organization and must be replaced. It is a critical workforce metric that significantly impacts organizational stability, financial performance and operational efficiency (Olubiyi et al., 2019). Turnover can occur for various reasons, ranging from job dissatisfaction and career progression to organizational restructuring and layoffs, making it a key concern in Human Resource Management (HRM) (Al-Suraihi et al., 2021).

There are two primary types of turnover: voluntary and involuntary. Voluntary turnover occurs when employees choose to leave their jobs, often in search of better career opportunities, higher salaries, improved work-life balance or greater job satisfaction. On the other hand, involuntary turnover happens when employees are terminated due to factors such as layoffs, underperformance or structural changes within the company (Al-Suraihi et al., 2021). Additionally, turnover can be classified as functional or dysfunctional. Functional turnover refers to the departure of underperforming employees, which can benefit the organization by improving team efficiency, whereas dysfunctional turnover involves the loss of high-performing employees, leading to potential skill shortages and disruptions in business operations (Wallace & Gaylor, 2012). Another important distinction is between avoidable and unavoidable turnover. Avoidable turnover results from factors that organizations can control, such as poor management, lack of career growth or inadequate compensation, while unavoidable turnover is driven by external factors, such as retirement, personal relocation or health issues (Barrick & Zimmerman, 2005).

Taking turnover into account is essential because of its financial, operational and cultural implications. Financially, replacing an employee is costly, with expenses related to recruitment, onboarding and training. The Work Institute (2024) estimates that replacing a single employee costs approximately 33% of their annual salary, making turnover a substantial financial burden, particularly in industries with high turnover rates such as retail. Operationally, frequent employee departures disrupt workflow and reduce productivity, as new hires require time to adapt and reach peak performance levels (Rahaman & Bari, 2024). In customer-facing industries like retail, turnover negatively impacts service quality and customer satisfaction, as experienced employees are more effective in handling customer interactions and operational demands (Bar-Gil et al., 2024).

Beyond financial and operational concerns, turnover also has psychological and cultural effects within organizations. A high turnover rate can lower employee morale, increase workplace stress and reduce overall job engagement. The departure of experienced employees leads to the loss of institutional knowledge, further weakening long-term strategic goals and business continuity (Gamba et al., 2024). Consequently, organizations are increasingly focusing on proactive strategies to manage turnover effectively.

One of the most effective approaches to mitigating turnover is the use of People Analytics, which enables HR departments to analyze key workforce indicators — such as job satisfaction, career progression and workplace sentiment — to anticipate employee departures and develop targeted retention strategies. Predictive analytics in HR allows organizations to shift from reactive turnover management to proactive workforce planning, ensuring business continuity and optimizing talent retention (Rahaman & Bari, 2024). Companies that strategically incorporate predictive modeling into their workforce planning processes can enhance workforce stability, reduce hiring costs and create a more resilient and engaged workforce (Levenson, 2018).

2.2 Traditional Models for Employee Turnover

The study of employee turnover has gained significant attention over the years, leading to the development of several theoretical models that explain why employees decide to leave their organizations. Among the most influential contributions in this field are the works of Mobley (1977), Lee (1988), Lee & Mitchell (1994), Morrell et al. (2008) and Wöcke & Heymann (2012), each offering different perspectives on turnover decision-making.

One of the earliest and most foundational models is Mobley's framework, which emphasizes the relationship between job dissatisfaction and voluntary turnover. According to this model, employees who experience dissatisfaction begin evaluating alternative job opportunities, comparing offers and assessing the potential costs and benefits of leaving their current role. This perspective suggests that the decision to leave an organization is largely driven by perceived job alternatives and the weighing of risks associated with the transition (G. J. Lee & Rwigema, 2005; Wöcke & Heymann, 2012).

Building upon Mobley's findings, Lee introduced a refined approach that shifted the focus from job satisfaction to job commitment and involvement as the key determinants of voluntary turnover. His research highlights that employees' level of attachment to the organization plays a crucial role in influencing their decision to stay or leave, reinforcing the idea that commitment, rather than satisfaction alone, dictates turnover behavior (Lee, 1988).

A more dynamic and comprehensive approach was later proposed by Lee & Mitchell (1994) with the development of the *Unfolding Model of Turnover*. This framework suggests that turnover decisions are often triggered by external or internal shocks that lead employees to reassess their employment situation. The model outlines different decision-making paths, considering scenarios in which employees undergo a reevaluation process either due to a significant external event, an unexpected career opportunity or a gradual reassessment of their current role. In some cases, employees might be prompted to leave by a sudden career disruption, while in others, turnover results from a long-term decline in job commitment without an external trigger.

Seeking to refine and improve the predictive accuracy of turnover models, Morrell et al. introduced the *Mapping the Decision to Quit* framework, which builds upon the *Unfolding Model* by emphasizing that turnover is rarely an impulsive decision but rather a gradual and structured process. This model argues that employees typically go through multiple stages before making a final decision to leave, reinforcing the idea that turnover should be seen as a series of evolving considerations rather than a single event. Compared to its predecessors, this model is considered to provide a more accurate and structured depiction of turnover behavior, allowing organizations to anticipate employee departures with greater precision (Morrell et al., 2008).

These theoretical models have significantly shaped turnover research by providing insights into the various factors that drive employee departures. While early models primarily linked turnover to dissatisfaction and commitment, later frameworks integrated external influences, career shocks and long-term psychological processes, emphasizing the complex and multi-faceted nature of turnover. Understanding these models remains essential for organizations seeking to develop effective retention strategies and minimize workforce disruptions.

Despite the significant contributions of traditional models in explaining employee turnover, they also exhibit limitations that reduce their predictive power and practical applicability. One of the primary shortcomings is their limited predictive accuracy, as these models often focus on individual psychological factors such as job satisfaction and organizational commitment but fail to incorporate a comprehensive set of variables that can enhance prediction accuracy (Morrell et al., 2008). Research has shown that turnover decisions are influenced by a wide array of factors, many of which are not adequately captured by traditional frameworks. Additionally, traditional models fail to consider external influences, such as macroeconomic conditions, labor market trends and organizational restructuring, which can significantly impact an employee's decision to leave. While the *Unfolding Model* introduced the concept of external "shocks" triggering turnover, earlier models primarily treated turnover as a gradual process based on dissatisfaction, overlooking abrupt decision-making scenarios (Lee & Mitchell, 1994).

Another critical limitation of many traditional turnover models is their tendency to assume linear relationships between variables, whereas real-world turnover decisions often exhibit nonlinear patterns influenced by multiple interacting factors (Lee & Mitchell, 1994). These models tend to simplify turnover decision-making into predefined stages, which may not fully reflect the dynamic and evolving nature of employee exit decisions. Additionally, conventional statistical approaches, such as logistic regression, dominate turnover research but face limitations when applied to real-world workforce data. These methods struggle with imbalanced datasets, where the number of employees who leave is significantly smaller than those who stay, leading to biased predictions and reduced generalizability. Furthermore, traditional approaches often rely on predefined assumptions about variable relationships, which limits their ability to detect hidden correlations and capture the interplay between multiple factors that influence turnover, particularly when these interactions evolve over time (Park et al., 2024).

Modern approaches leveraging machine learning techniques, such as decision trees and ensemble models, have demonstrated superior predictive performance by automatically learning from data patterns rather than relying on fixed assumptions. Unlike traditional methods, machine learning models can process high-dimensional datasets, capturing intricate relationships between employee attributes, workplace conditions and external labor market factors (Park et al., 2024). Consequently, as workforce planning becomes increasingly data-driven, traditional turnover models face growing challenges in delivering actionable and precise insights, reinforcing the need for more sophisticated predictive methodologies.

2.3 Factors affecting Employee Turnover

Employee turnover is influenced by multiple factors, which can be categorized into financial, organizational, demographic and economic aspects. Understanding these determinants is crucial for developing predictive models and implementing retention strategies. While no single factor can fully explain why employees leave, research has identified several key drivers of turnover that interact in complex ways. Below are outlined the most significant factors affecting employee retention:

- Salary and Compensation: Employees who feel underpaid or perceive wage disparities compared to competitors are more likely to leave in search of better financial rewards. Research consistently highlights pay dissatisfaction as a leading cause of voluntary turnover (Solomon et al., 2024).
- Non-Monetary Benefits: While salary alone does not ensure retention, benefits such as flexible work arrangements, bonuses and health benefits contribute to job satisfaction and play a role in employee retention (Orujaliyev, 2024).
- Workload and Burnout: Employees experiencing excessive workloads and high-pressure environments report increased stress, leading to disengagement and higher turnover rates. Industries like retail, where job intensity is high, show a particularly strong correlation between workload and turnover (Coombs, 2024; H. Kim & Stoner, 2008; Koo et al., 2020).
- Career Development Opportunities: Employees are more likely to stay in organizations that provide structured career progression and professional development programs. A lack of career growth is a strong predictor of voluntary turnover (Chow et al., 2007).
- Tenure: Employees with shorter tenures exhibit a higher likelihood of leaving, whereas those with longer tenures tend to stay unless they perceive a lack of career advancement (Ju & Li, 2019).
- Job Satisfaction: Employees dissatisfied with their daily responsibilities, workplace environment or interpersonal relationships are more prone to leaving. Poor leadership and ineffective communication contribute to higher turnover, whereas a positive workplace culture reduces these risks (Judge & Kammeyer-Mueller, 2012; E. G. Kim & Kim, 2021; Tian et al., 2020).
- Leadership and Organizational Culture: Leadership style and workplace culture significantly influence employee retention. Supportive leadership, transparent communication and a culture of recognition contribute to a positive work environment, fostering employee commitment. In contrast, poor leadership, lack of managerial support and a toxic work culture increase turnover risks, as employees may feel undervalued or unmotivated (Bass, 1990; Tian et al., 2020).
- Demographics (Age, Gender and Education Level):
  - Age: Younger employees tend to seek new opportunities, while older employees prefer stability (Ng & Feldman, 2009).
  - Gender: Research suggests male employees are more likely to leave than female employees, who may prioritize job stability (Grissom et al., 2012).

- Education Level: Highly educated employees are more mobile and seek roles aligned with their qualifications, whereas lower-educated employees tend to remain in their positions longer (Benson, 2006; G. Chen et al., 2021).

- Economic Conditions and Labor Market Trends: In periods of economic uncertainty, employees tend to remain in their jobs for security, whereas in times of economic growth, they are more likely to leave for higher salaries and better benefits (Wynen & Op De Beeck, 2014).
- Shifting Societal and Cultural Expectations: Recent trends, particularly post-COVID-19, show employees prioritizing flexibility, meaningful work and work-life balance. Organizations failing to adapt to these expectations experience higher turnover rates (Tessema et al., 2021).

Given the complexity of these factors, organizations are increasingly turning to predictive analytics as a strategic tool to anticipate and mitigate turnover risks. By analyzing workforce data, HR professionals can identify employees who are at risk of leaving and implement targeted retention strategies, such as career development plans, performance recognition programs and flexible working conditions. Predictive models enable companies to transition from reactive to proactive turnover management, optimizing talent retention and workforce stability. As machine learning models continue to evolve, organizations that leverage data-driven decision-making are better positioned to enhance employee engagement, minimize turnover and improve overall business performance (Punnoose & Ajit, 2016; Solomon et al., 2024).

Ultimately, employee turnover is driven by a complex combination of financial, organizational, demographic and economic factors. Organizations that prioritize competitive compensation, manageable workloads, career growth opportunities and a positive workplace culture tend to achieve higher retention rates. At the same time, external influences, such as labor market trends and workforce expectations, shape turnover patterns, requiring companies to adopt more data-driven and adaptive HR strategies. By integrating predictive analytics, businesses can proactively address turnover challenges, ensuring long-term workforce stability and operational efficiency.

2.4 Workforce Planning and the Role of Predictive Analytics in Turnover Management

Workforce planning has become a fundamental aspect of strategic human resource management, allowing organizations to align talent strategies with long-term business objectives. By anticipating future workforce needs, identifying skill gaps and developing strategies to attract, retain and develop employees, workforce planning ensures organizations remain adaptable in evolving market conditions. In industries such as retail and technology, where workforce demands are particularly sensitive to technological advancements and economic fluctuations, effective workforce planning is essential for maintaining organizational agility and competitiveness (Cappelli, 2008; Jaillet et al., 2022).

A key benefit of workforce planning is risk mitigation. Organizations that proactively assess workforce skills, adaptability and growth potential can address skill shortages before they impact operations. By leveraging predictive analytics, HR teams can optimize talent pipelines through targeted hiring, training and internal mobility strategies, reducing the risks associated with last-minute recruitment efforts or skill mismatches (J. H. Marler & Boudreau, 2017). Another critical advantage is workforce agility, as organizations that engage in continuous workforce assessment are better prepared to navigate economic downturns, shifts in consumer behavior and technological disruptions, ensuring business continuity (Yoon, 2023).

2.4.1 The Impact of Worforce Planning on Turnover Management

Workforce planning plays a pivotal role in enhancing employee retention by aligning talent strategies with long-term business objectives and fostering internal career development opportunities. Employees who see clear opportunities for internal mobility and skills development are more likely to remain committed to their organization. When organizations fail to identify and nurture internal talent, they risk losing high-performing employees to competitors, increasing turnover costs. Thus, organizations that integrate workforce planning into their HR strategies not only reduce turnover but also create a more agile and strategically aligned workforce (Minbaeva, 2018).

Beyond career progression, workforce planning increasingly incorporates workload monitoring and stress prevention as part of proactive workforce management. Employees experiencing sustained high workloads, excessive overtime or job-related stress are at increased risk of reduced well-being and lower job satisfaction. People analytics helps HR teams identify workload imbalances, enabling better task distribution and timely interventions to support employee well-being and performance (Polzer, 2022). By taking a proactive and data-driven approach, organizations can enhance employee well-being, foster better job satisfaction and improve workforce stability.

2.4.2 The Role of Predictive Analytics and Machine Learning

Traditional statistical models have long been used for turnover prediction due to their simplicity and interpretability. However, these models often fall short when applied to workforce data, as they rely on assumptions of linearity, require extensive manual feature engineering and struggle with high-dimensional and imbalanced datasets. Machine learning (ML) techniques have emerged as a more powerful alternative, offering adaptive learning capabilities, improved accuracy and the ability to capture complex interactions among employee attributes (J. Marler et al., 2017; Polzer, 2022; Zhao et al., 2019). In response to these limitations, machine learning has emerged as a more adaptive and data-driven alternative, offering superior analytical power to handle large-scale workforce data, recognize hidden patterns and significantly improve turnover predictions. Unlike traditional regression-based approaches that assume a fixed relationship between variables, ML techniques can dynamically adjust to workforce trends, making them particularly effective in predicting turnover and identifying at-risk employees in real-time (Zhao et al., 2019).

Despite their widespread use in HR analytics due to their interpretability and simplicity, conventional methods such as logistic regression and generalized linear models (GLMs) face several challenges when applied to real-world workforce data, particularly in turnover prediction. The main limitations include:

- Linear Assumptions: Many traditional models assume a linear relationship between employee attributes (e.g., salary, job tenure) and turnover probability. Although interaction terms can be manually incorporated into linear models (e.g., multiple regression), their inclusion requires prior specification by the researcher. In contrast, machine learning models can automatically detect and learn complex and nonlinear interactions, often leading to higher predictive accuracy (Zhao et al., 2019).
- Manual Feature Engineering: Traditional approaches require manual selection and transformation of variables, whereas ML models can automatically learn relevant features from data, improving prediction accuracy (Zhao et al., 2019).
- Poor Performance with High-Dimensional and Unstructured Data: Workforce data often include textual, categorical and hierarchical information, which conventional models struggle to process effectively (Polzer, 2022).
- Difficulty Handling Class Imbalance: In turnover datasets, the proportion of employees who leave is often much smaller than those who stay, leading traditional models to produce biased predictions toward the majority class (Zhao et al., 2019).

These challenges underscore the necessity of more advanced techniques capable of capturing complex interactions, uncovering hidden correlations and significantly improving predictive accuracy. Turnover is an inherently complex phenomenon influenced by a dynamic interplay of intrinsic employee factors and external labor market conditions. Traditional statistical methods, which rely on predefined assumptions, struggle to capture these multifaceted relationships, often leading to oversimplified models and suboptimal predictions. Turnover prediction requires accounting for multiple interacting factors such as job satisfaction, career progression and external labor market conditions. Machine learning models are particularly effective in this context because they can analyze structured and unstructured data simultaneously, uncovering non-obvious patterns that traditional statistical models might overlook (Fukui et al., 2023; Zhao et al., 2019).

Machine learning models address the shortcomings of traditional approaches by automating feature learning, processing high-dimensional data and improving prediction accuracy through adaptive learning techniques. Some of the key advantages include:

- Capturing Nonlinear Patterns: Unlike regression-based models, Decision Trees, Random Forests and Gradient Boosting models (such as XGBoost) can identify complex decision pathways in employee behavior, improving turnover prediction (Zhao et al., 2019).
- Enhanced Handling of Class Imbalance: Machine learning algorithms leverage techniques such as cost-sensitive learning, oversampling (SMOTE) and ensemble methods to improve the prediction of minority classes, making them more effective in identifying at-risk employees (Li & Fox, 2023).
- Integration of Multiple Data Sources: ML models can simultaneously analyze structured (e.g., salary, tenure) and unstructured data (e.g., performance reviews, survey responses) to uncover deeper insights into turnover patterns (Fukui et al., 2023).
- Adaptive Learning and Real-Time Prediction: Traditional models provide static insights based on past data, while ML models continuously refine predictions as new data becomes available. This adaptability ensures that retention strategies remain relevant even as workforce trends evolve, offering a significant advantage over conventional approaches (Gomes et al., 2017).

Furthermore, deep learning techniques such as recurrent neural networks (RNNs) are increasingly being explored in workforce analytics to detect long-term trends and sequential patterns in employee tenure and engagement data (de Oliveira et al., 2019).

Beyond enhancing turnover prediction, one of the most transformative aspects of ML-driven workforce planning is its ability to enable personalized retention strategies tailored to individual employees. Unlike traditional HR interventions that apply broad policies to all employees, machine learning allows organizations to identify specific employees at risk and recommend individualized retention strategies based on their profile (Hall, 2021). Additionally, ML enhances career development planning by predicting which employees would benefit from upskilling opportunities or role transitions, thereby fostering talent growth within the organization (Rathore & Rathore, 2024). Another critical advantage is its capacity for real-time intervention, as machine learning models can detect early warning signs of disengagement, providing managers with actionable insights before employees decide to leave. By transforming HR from a reactive crisis management approach into a proactive and data-driven strategic function, machine learning significantly enhances employee satisfaction, talent retention and long-term organizational stability, enabling companies to make well-informed, forward-thinking workforce decisions (Hall, 2021; Rathore & Rathore, 2024).

2.4.3 Challenges in Implementing Workforce Planning Strategies

Despite the advantages, implementing workforce planning strategies presents several challenges. Organizations often struggle with fragmented data sources, making it difficult to generate accurate predictive insights. The effectiveness of workforce planning depends on data quality and integration across HR systems, requiring investment in robust analytical infrastructure (Minbaeva, 2018). Additionally, ethical concerns and algorithmic biases must be addressed when utilizing machine learning for workforce decisions. biases in historical data can reinforce existing inequalities in hiring, promotions and career development, requiring organizations to implement transparent and fair AI models to ensure ethical workforce planning (Polzer, 2022).

Another significant challenge is organizational resistance to data-driven decision-making. Many HR professionals remain reliant on traditional workforce planning methods and may lack the analytical expertise to interpret predictive models effectively. Addressing these barriers requires investment in HR analytics training and fostering a data-driven culture within organizations (Jaillet et al., 2022). Furthermore, integrating workforce planning into broader business strategies demands alignment with leadership priorities and continuous refinement of analytical models to account for evolving market dynamics.

2.4.4 Workforce Planning as a Competitive Advantage

Organizations that successfully integrate predictive analytics into workforce planning gain a strategic advantage by ensuring long-term talent sustainability. Companies that anticipate workforce needs and develop agile HR strategies are better positioned to navigate market disruptions, retain key employees and enhance workforce productivity (Cappelli, 2008; J. H. Marler & Boudreau, 2017).

Looking ahead, the future of workforce planning will increasingly rely on AI-driven insights to proactively address retention challenges, optimize workforce distribution and maintain a resilient and engaged workforce (Solomon et al., 2024; Zhao et al., 2019). Organizations that invest in data-driven workforce planning today will be better equipped to attract, retain and develop top talent in an increasingly competitive job market. The shift towards predictive workforce management not only enhances business continuity but also fosters a culture of innovation and strategic agility, reinforcing long-term organizational success.

2.5 Machine Learning Models for Turnover Prediction

As organizations seek to improve workforce stability, machine learning (ML) has emerged as a powerful tool for predicting employee turnover. Traditional statistical models, while useful, often fail to capture the complex and non-linear relationships that influence employee retention. ML techniques, particularly ensemble learning models, provide superior predictive accuracy and allow HR professionals to develop proactive retention strategies based on data-driven insights (Zhao et al., 2019).

2.5.1 Comparative Analysis of Machine Learning Models in Turnover Prediction

The study by Kovvuri & Dommeti (2022) aimed to evaluate four machine learning models, including XGBoost, in predicting employee turnover using employee data that included demographic characteristics and employment history. The focus was to identify the model most capable of predicting employees likely to leave the organization, assisting HR departments in planning staffing needs effectively. To achieve this, the authors compared the models based on several key metrics: accuracy, precision, recall, F1 score and AUC (Area Under the Curve). These metrics are crucial in turnover prediction, evaluating how well each model identifies employees at risk of leaving (recall) while minimizing false predictions of turnover (precision). The study demonstrated that XGBoost was the best-performing model, achieving 85.3% accuracy, 88% precision, 69% recall, 77% F1 score and 88% AUC. Random Forest Classifier came second with 83.2% accuracy, 82% precision, 69% recall, 75% F1 score and 85% AUC. Naive Bayes and Logistic Regression performed less effectively, with Gaussian Naive Bayes achieving 70.7% accuracy, 61% precision, 53% recall and 72% AUC, while Bernoulli Naive Bayes performed slightly better with 79.4% accuracy, 76% precision, 63% recall and 82% AUC.

In Juvitayapun (2021), the study aimed to improve the prediction of employee turnover by comparing the performance of four machine learning methods — Logistic Regression (LR), Random Forest (RF), Gradient Boosting Tree (GBT) and Extreme Gradient Boosting (XGBoost). The goal was to assess the effectiveness of these models in identifying employees at risk of leaving the organization, with a particular focus on integrating employee event features such as leave records and merit increases compared to market benchmarks. By incorporating these features, the study sought to enhance the predictive power of the models and provide HR professionals with actionable insights for better workforce management. The results showed that XGBoost outperformed all other models, achieving 98.03%

accuracy, a 90.79% F1 score, 97.18% precision and 85.19% recall. GBT also performed well, with an F1 score of 87.42% and recall of 81.48%. These results highlight the superior performance of XGBoost in predicting employee turnover, making it the most suitable model for this task, followed closely by Gradient Boosting Tree.

The study by Punnoose & Ajit (2016) aims to improve the prediction of employee turnover using machine learning techniques, with a focus on XGBoost. The background of the study emphasizes that employee turnover is a significant issue for organizations, impacting productivity and business continuity. Accurate turnover prediction helps companies implement more effective retention strategies. However, the main challenge lies in the quality of the data used, which often contains noise due to insufficient investment in Human Resource Information Systems (HRIS). Despite this, this model demonstrated excellent results, outperforming all other models, including Logistic Regression, Random Forest and Naïve Bayes. The study found that XGBoost achieved an AUC score of 0.86 and showed high accuracy in predicting turnover. Furthermore, it excelled in terms of low memory usage and fast training times, which makes it highly efficient for turnover prediction tasks. This study, based on data from a global retailer, confirms the robustness of XGBoost, solidifying it as a strong choice for turnover forecasting, particularly in noisy data scenarios.

Similar findings were reported by Zhao et al. (2019), who expanded the scope by testing additional models, including Neural Networks (NNs) and Support Vector Machines (SVMs), reinforcing the efficiency of XGBoost in handling larger datasets. In this article, a comprehensive evaluation of supervised machine learning algorithms to predict employee turnover across small, medium and large-scale organizational datasets was conducted. The study examined multiple models, including Decision Trees, Random Forests, Gradient Boosting Trees, Extreme Gradient Boosting (XGBoost), Logistic Regression, Support Vector Machines, Neural Networks, Linear Discriminant Analysis, Naïve Bayes and K-Nearest Neighbors. The findings highlighted the superior predictive performance of XGBoost and Gradient Boosting Trees, particularly for medium and large datasets, due to their ability to handle complex interactions, reduce overfitting through regularization and rank feature importance automatically. Additionally, the study found that XGBoost outperformed traditional Gradient Boosting Machines (GBM) in terms of computational efficiency, as it required shorter training times while maintaining high predictive accuracy. This advantage was attributed to XGBoost's optimized parallelization, memory-efficient tree structure and integrated regularization techniques. The study also emphasized that small datasets often introduce high variance and randomness, making model selection less reliable in those cases. Furthermore, Zhao et al. underscored the importance of appropriate data preprocessing and model interpretability, advocating for feature ranking and classification rule visualization to enhance the practical application of turnover prediction models.

The study by Park et al. (2024) evaluated the performance of multiple machine learning models in predicting employee turnover intention, highlighting the advantages of XGBoost over traditional statistical approaches. Among the tested models, XGBoost demonstrated the highest accuracy (78.5%), outperforming Logistic Regression (78.3%) and K-Nearest Neighbors (KNN) (76.1%). A key strength of XGBoost noted in the study was its ability to rank feature importance effectively, allowing organizations to identify the most influential factors driving turnover. Notably, job security emerged as the most critical predictor of turnover intention, reinforcing the role of workplace stability in employee retention. Furthermore, the study integrated machine learning techniques with traditional econometric analysis, demonstrating how advanced predictive models can enhance workforce planning by providing data-driven insights into employee behavior.

Table 1 presents a comparative summary of the methodologies used in the reviewed studies. It highlights the machine learning models evaluated, the performance metrics considered, the validation strategies applied and the specific contexts in which these models were tested.

Table 1 - Overview of Turnover Prediction Studies: Models, Metrics and Applications

Article	Models Compared	Evalution Metrics	Evaluation Strategy	Application Context
Kovvuri & Dommeti (2022)	Logistic Regression, Naïve Bayes, Random Forest, XGBoost	Accuracy, Precision, Recall, F1 Score, AUC-ROC	75/25 Train-Test Split	Employee Turnover Prediction using Kaggle dataset
Juvitayapun (2021)	Logistic Regression, Random Forest, Gradient Boosting Trees, XGBoost	Accuracy, Precision, Recall, F1 Score, ROC AUC, Prediction Cost	10-fold Cross-Validation, Grid Search for Hyperparameter Tuning	Employee Turnover Prediction in a Packaging Manufacturing Firm (Thailand)
Punnoose & Ajit (2016)	XGBoost, Logistic Regression, Naïve Bayes, Random Forest, SVM, LDA, KNN	AUC-ROC, Runtime, Memory Utilization	10-fold Cross-Validation, 80/20 Train-Test Split	Employee Turnover Prediction in a Global Retailer (US workforce)
Zhao et al. (2019)	Decision Tree, Random Forest, Gradient Boosting Trees, XGBoost, Logistic Regression, SVM, Neural Networks, LDA, Naïve Bayes, KNN	Accuracy, Precision, Recall, F1 Score, AUC-ROC	10-fold Cross-Validation, Grid Search for Hyperparameter Tuning	Employee Turnover Prediction using datasets from a US Bank and IBM Watson Analytics
Park et al. (2024)	Logistic Regression, K-Nearest Neighbors, XGBoost	Accuracy, Precision, Recall, F1 Score	70/30 Train-Test Split, 4-fold Cross-Validation	Prediction of turnover intention among new college graduates in South Korea

As summarized in *Table 1*, all reviewed studies included XGBoost in their comparisons, consistently demonstrating its superior predictive performance over models such as Random Forest, Logistic Regression and Naïve Bayes. While the choice of evaluation metrics varied, accuracy and AUC were the most commonly reported. Additionally, validation strategies differed across studies, with some employing k-fold cross-validation and others relying on simple holdout validation, emphasizing the need for standardized benchmarking approaches in employee turnover prediction research.

Overall, these studies reinforce XGBoost's effectiveness in employee turnover prediction, as it consistently outperforms traditional models across diverse datasets and validation settings.

2.5.2 XGBoost: A Benchmark Model for Turnover Prediction

XGBoost is a machine learning model based on the Gradient Boosting framework, widely recognized for its efficiency and ability to handle large datasets effectively. In the context of turnover prediction, XGBoost is particularly valuable due to its strong performance in handling complex, high-dimensional and noisy data while maintaining both accuracy and computational efficiency (Punnoose & Ajit, 2016).

XGBoost constructs decision trees sequentially, with each new tree learning from and correcting the errors made by the previous trees. This process, known as boosting, enhances predictive accuracy by focusing on the residuals — the differences between the actual and predicted values. Additionally, XGBoost incorporates both L1 and L2 regularization techniques, which are crucial to prevent overfitting, particularly in turnover prediction datasets where many features and potential collinearity could otherwise compromise model generalization (T. Chen & Guestrin, 2016).

A key advantage of XGBoost is its scalability, as it optimizes both time and memory usage, enabling it to efficiently process large volumes of data (T. Chen & Guestrin, 2016). This scalability is especially important in turnover prediction scenarios, where data often spans several years and includes numerous features capturing both individual employee characteristics and organizational factors (Punnoose & Ajit, 2016). In addition, XGBoost handles missing data natively, automatically learning the optimal split direction for records with missing values, ensuring they are effectively incorporated into the training process without requiring prior imputation (T. Chen & Guestrin, 2016). Another important feature of XGBoost is its ability to handle imbalanced data, which is common in turnover prediction, where turnover events typically represent a smaller proportion of the total dataset. The parameter *scale_pos_weight* can be adjusted to give higher importance to the minority class (turnover), ensuring the model remains sensitive to these critical cases (T. Chen & Guestrin, 2016; Punnoose & Ajit, 2016). Furthermore, XGBoost offers extensive flexibility through its hyperparameter tuning options, including *learning rate*, *max_depth*, *subsample* and *n_estimators*. These parameters control the learning process, tree complexity and data sampling strategy, allowing the model to be carefully optimized for maximum predictive accuracy (T. Chen & Guestrin, 2016).

Overall, this chapter has explored the conceptual foundations and recent advances in employee turnover prediction, emphasizing the importance of incorporating both intrinsic and extrinsic factors into workforce planning. It also highlighted the limitations of traditional statistical approaches and the advantages of machine learning techniques — particularly XGBoost — in capturing complex, non-linear patterns in employee data. These insights provide a solid basis for the methodological approach adopted in this dissertation. The next chapter will contextualize these findings by examining the company's current decision-making process and detailing the dataset used to build the proposed predictive model.

2.6 Summary of Key Findings

The literature reviewed in this chapter highlights several critical insights relevant to the development of a turnover prediction model in the retail sector. First, employee turnover is a multifaceted phenomenon influenced by both intrinsic factors — such as compensation, performance, seniority and job satisfaction — and extrinsic factors, including labor market trends and organizational culture. Traditional models of turnover, while valuable for understanding individual decision-making processes, are limited in their predictive accuracy and scalability. In contrast, the growing use of People Analytics and machine learning techniques has demonstrated significant potential for improving turnover prediction accuracy and enabling proactive workforce planning. XGBoost, in particular, stands out due to its ability to handle imbalanced and high-dimensional data, perform regularization and provide interpretable feature importance rankings. Furthermore, predictive analytics not only enhances strategic HR decision-making but also contributes to cost efficiency, retention planning and long-term workforce stability.

Given the volatility of the retail workforce and the operational challenges posed by high turnover rates, the literature also emphasizes the need for robust, data-driven decision-support tools capable of addressing this complexity. These findings underscore the value of integrating advanced analytical models into HR practices and provide the conceptual foundation for the methodology adopted in this dissertation.

3. THE PROBLEM

Following the literature review, which highlighted both the theoretical foundations and modern advances in turnover prediction, this chapter introduces the real-world context in which the proposed solution will be applied. The aim is to bridge the gap between academic models and practical implementation by detailing the problem that motivated the development of the predictive framework.

In particular, this chapter begins by describing the current decision-making process used by the company to estimate voluntary turnover, highlighting its empirical nature and inherent limitations. It then provides an overview of the available data, including its structure, key variables and sources. Finally, an exploratory data analysis is conducted to identify trends, validate data quality and support the design of the predictive model in the next chapter.

3.1 Problem Description

3.1.1 As-Is

In strategic workforce planning, having a reliable approach to anticipating workforce needs is essential for ensuring operational efficiency. While well-structured workforce plans help mitigate staffing risks, natural employee turnover remains an inherent challenge, making it crucial to integrate workforce projections into decision-making processes. By proactively estimating future employee departures, companies can ensure a balanced workforce structure, reducing inefficiencies and optimizing recruitment efforts.

Currently, workforce adjustments are primarily informed by historical workforce trends, using past variations to approximate turnover expectations. One of the key references in this process is the FTE-based method, which monitors turnover patterns over recent years to provide insights into workforce dynamics. While this approach offers a practical benchmark for workforce planning, it does not incorporate individual employee characteristics or external labor market conditions, making it less adaptable to evolving workforce trends.

In addition to historical trends, the company also utilizes the Value-Based Management (VBM) system, which provides detailed workforce projections per role cluster, store and year. These projections help align staffing needs with expected workforce changes, ensuring that adjustments can be made proactively rather than reactively.

As workforce dynamics become increasingly complex, enhancing workforce planning with predictive models presents an opportunity to refine turnover estimations, integrating both intrinsic (e.g., employee attributes) and extrinsic (e.g., market conditions) factors. This approach aims to complement existing workforce planning strategies, allowing for a more adaptive and data-driven decision-making process.

3.1.2 Point for Improvement

Despite these structured processes, there are two key areas where improvements could significantly enhance the company's ability to accurately predict and manage workforce changes, thereby fostering a more strategic and efficient workforce planning approach.

First, the current turnover prediction method does not incorporate employee-related variables, such as job satisfaction, tenure, performance evaluations, contract type, work schedule and external labor market conditions. These intrinsic and extrinsic factors are key drivers of voluntary turnover, influencing employees' decisions to stay or leave. However, they are not captured in the FTE-based estimation model, which solely relies on historical workforce variations. This limitation reduces the company's ability to anticipate workforce gaps with precision, making it difficult to implement proactive retention strategies, optimize staffing levels and mitigate operational disruptions. As a result, the company remains in a reactive stance, where adjustments to workforce planning are made after turnover occurs rather than in anticipation of it.

Second, the company's initial workforce needs are determined by Senior People Managers based primarily on experience and managerial intuition, rather than on structured, data-driven methodologies. While this expertise is valuable, it introduces an element of subjectivity, leading to potential inconsistencies in workforce allocation across different store formats, regions and job roles. This reliance on human judgment may result in overstaffing in some areas, leading to unnecessary labor costs or understaffing in critical roles, impacting operational efficiency and service levels. Furthermore, this manual decision-making process demands a significant time investment from managers, which could otherwise be allocated to more strategic workforce initiatives.

By integrating employee-specific variables into turnover prediction models and transitioning towards a data-driven approach in workforce planning, the company can gain a more accurate, granular understanding of the factors driving employee departures. This shift would enable proactive workforce management, where HR teams can anticipate turnover risks, adjust hiring strategies accordingly and implement targeted retention efforts to mitigate unnecessary attrition. Additionally, leveraging predictive analytics in workforce planning would establish a more consistent and objective methodology, reducing reliance on managerial intuition and ensuring that staffing decisions are aligned with real operational demands.

Implementing these improvements would lead to multiple benefits, including:

- Optimized resource allocation, ensuring that each store format has the necessary workforce capacity.
- Cost reduction, by minimizing unexpected hiring needs and excessive labor costs.
- Greater workforce stability, reducing operational disruptions caused by unexpected employee departures.
- Enhanced strategic decision-making, allowing HR and management teams to focus on long-term workforce planning rather than reactive adjustments.

By refining its turnover prediction approach and adopting a more structured, analytics-based workforce planning model, the company can enhance operational efficiency, reduce workforce management costs and ensure sustainable business growth in an increasingly competitive retail landscape.

3.2 Data Overview and Exploratory Analysis

3.2.1 Data Overview

The dataset used for this analysis comprises a total of 189,160 employee records per year, spanning from 2015 to 2023. It is organized by "employee x year", meaning that for each year, there is a record for every employee, providing a robust and comprehensive longitudinal view of the workforce over time. The dataset consists of various variables that capture different aspects of employee characteristics, turnover and work-related factors. These variables include:

- Employee ID: A unique identifier for each employee.
- Year: The year to which the employee record corresponds.
- Turnover: Indicating whether the employee has left the organization.
- Salary: The fixed salary paid to the employee.
- Store Type: The type of store where the employee works.
- Role Cluster: The job role classification of the employee.
- Academic Qualifications: The educational background of the employee.
- Contract Type: Type of employment contract (e.g., permanent, fixed term).
- Marital Status: Employee's marital status.
- Age: The employee's age.
- Seniority: The length of time the employee has worked with the company.

- Workload Type: The type of workload (e.g., full-time, part-time).

The dataset is generally well-populated, with the highest proportion of missing values observed in the workload variable, which accounts for 10% of the total data. For all other variables, missing data is minimal, with an average of 1.5% missing values across the dataset. This low level of missing data ensures that the analysis can be conducted with a high degree of completeness.

Outliers were identified based on the interquartile range (IQR) method, which allows for detecting extreme values in numerical data. The variable with the highest percentage of outliers is salary, at 11%. This is likely due to a small proportion of employees in higher-level positions who receive significantly higher salaries compared to the majority of employees in lower-level roles. These outliers are considered natural and reflect real differences in pay scales within the company. Overall, 6.4% of the dataset contains outliers across all variables.

3.2.2 Data Analysis

In this section, we present an analysis of key variables and their relationships within the dataset. The goal is to understand the patterns, trends and distributions that will help us build a more effective predictive model for employee turnover. This foundational understanding will guide the modeling approach by highlighting important factors that need to be considered.

The Figure 1 illustrates the annual evolution of the workforce and turnover rate. The main trend observed is the gradual growth of the number of employees over time, except during the *COVID* years (2020 and 2021). During these years, the turnover rate drops significantly, likely due to the economic uncertainty that made employees more inclined to retain their current jobs due to the instability of the job market. Also, the increasing trend in the number of employees reflects the company's expansion, which is crucial for forecasting turnover and staffing needs in the coming years.

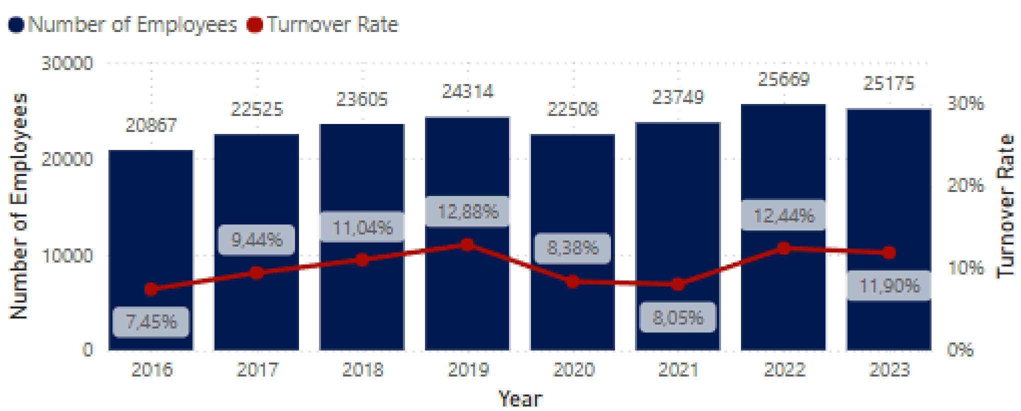

Figure 1 - Annual Workforce Evolution and Turnover Rate

Figure 2 highlights the percentage of employees who left the company compared to those who stayed. It shows that, on average, around 10.3% of employees left, while 89.7% remained. This imbalance is crucial for predictive modeling, as unbalanced data like this (where the number of departures is significantly lower than the number of employees remaining) can lead to biased predictions and reduced model performance, particularly by affecting the model's ability to learn from the minority class (employees who left the company). To improve the accuracy of turnover predictions, it is essential to address this imbalance using techniques such as oversampling, undersampling or adjusting the cost function to give more weight to the minority class.

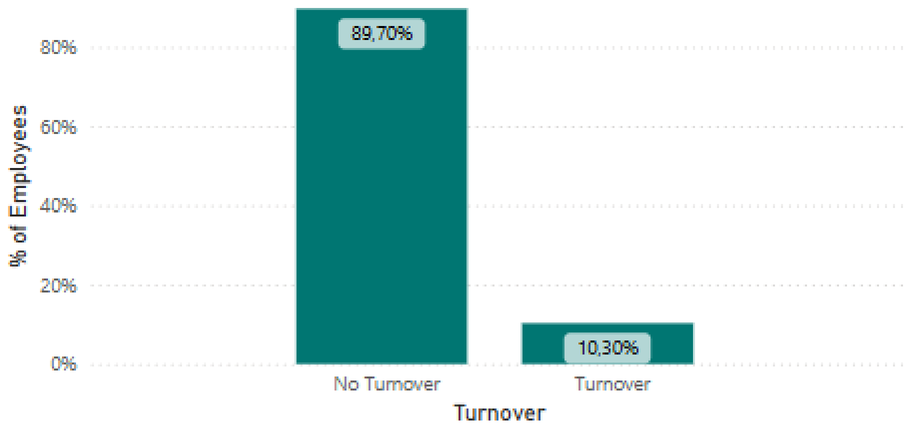

Figure 2 - Percentage of Employees with and without Turnover

Figure 3 shows the workforce distribution and corresponding turnover rates across the ten districts with the highest employee representation, which together account for approximately 87.5% of the total workforce. Lisbon and Porto stand out, representing 28.02% and 21.25% of employees, respectively.

Turnover rates across these districts generally range between 9% and 12%, with a few exceptions. Notably, Porto, despite being the second-largest employing district, records the second-lowest turnover rate (8.35%), ranking 9th among the ten in terms of employee attrition. This contrast may point to greater job stability or stronger employee-company alignment in the region.

The Madeira region presents a particularly low turnover rate of 4.56%, well below all other districts. This difference may be partially explained by its geographic characteristics — being an island — which could reduce external job opportunities and contribute to greater workforce retention.

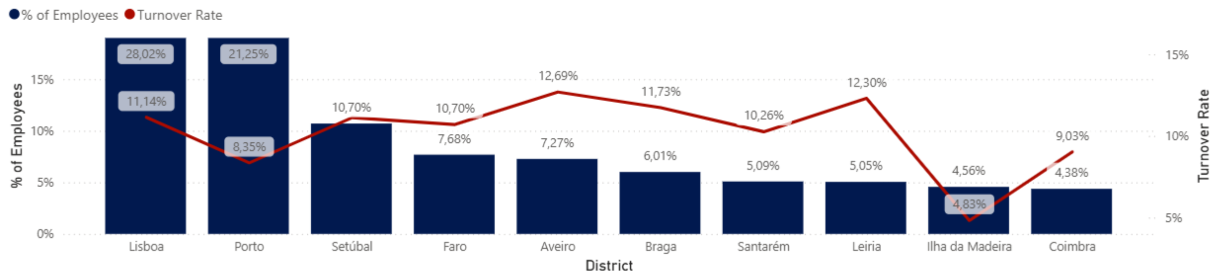

Figure 3 - Workforce Distribution and Turnover Rate by District (Top 10)

Figure 4 presents the distribution of turnover rates across different store types. Convenience stores, typically smaller retail establishments offering a limited selection of products, show the highest turnover rate at 14.99%. Supermarkets, which are larger than convenience stores and offer a broader range of products, exhibit a lower turnover rate of 10.61%. Hypermarkets, the largest store type with extensive product offerings, have the lowest turnover rate at 7.03%.

This data reveals a clear trend: as the size of the store increases, the turnover rate decreases. This suggests that employees in larger store formats, such as hypermarkets, may benefit from more stable working environments, possibly due to greater job security, better career development opportunities or better management practices compared to smaller store formats. This trend highlights the importance of considering store type when analyzing employee turnover and workforce management strategies.

Further insights into how workforce composition varies across store types are explored in Figure 37, which examines the distribution of seniority levels in each store format.

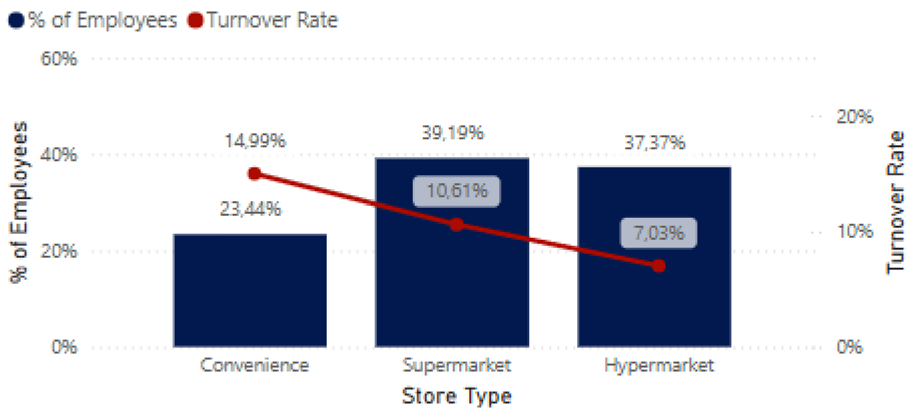

Figure 4 - Workforce Distribution and Turnover Rate By Store Type

Figure 5 illustrates the distribution of employees across different role clusters and their respective turnover rates. While some role clusters exhibit higher turnover rates individually, a pattern emerges: positions with a larger share of employees tend to also have higher turnover rates.

Notably, the roles with the highest number of employees – such as *Cashier Operator* (24.8%) and *Flow Operator* (22.5%) – are predominantly entry-level positions requiring fewer specialized skills. These roles, often associated with operational and repetitive tasks, may offer fewer career advancement opportunities, contributing to higher turnover rates.

In contrast, positions requiring more expertise and responsibility, such as *Main Supervisor* (4.2%) and *Section Coordinator Hmkt* (Hypermarket) (1.8%), tend to have lower turnover rates. These roles involve greater managerial responsibilities and decision-making power, which are often accompanied by greater job stability and career development opportunities, helping to foster long-term retention.

The *Other* category encompasses various roles that individually represent less than 1,5% of the dataset. A more detailed breakdown of these roles is provided in the appendix, offering additional insights into their respective turnover dynamics.

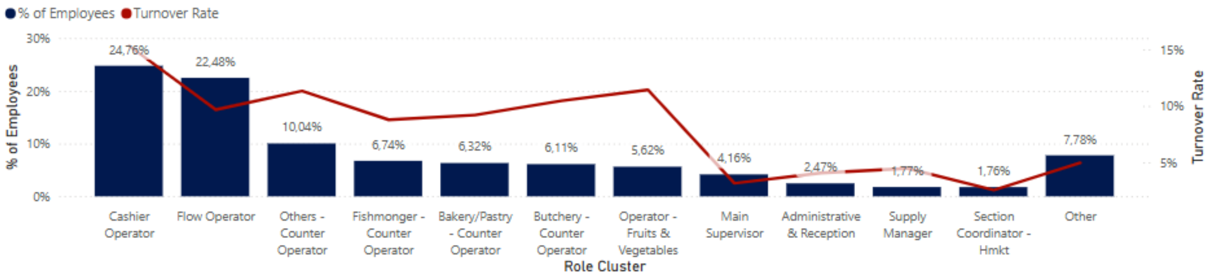

Figure 5 - Workforce Distribution and Turnover Rate by Role Cluster

Turnover rates tend to decline as employees gain seniority within the company, as shown in Figure 6. The most junior employees (0-1 years of seniority) have a turnover rate of 21.39%, which decreases to 20.77% for 1-3 years. This decline continues significantly, reaching 11.10% for 3-6 years, 6.56% for 6-10 years and dropping to only 2.57% for employees with over 10 years of seniority.

This represents an 87.97% relative reduction in turnover from the least experienced employees to the most experienced ones. This suggests that the early stages of employment are the most critical in terms of retention, possibly due to employees assessing job fit, career opportunities or facing adaptation challenges. Additionally, the distribution of employees by seniority shows that 45.33% of the workforce has been with the company for over a decade, reinforcing the notion of workforce stability at higher seniority levels. Conversely, newer employees (0-1 years and 1-3 years) account for over 34% of the workforce, aligning with the observed higher turnover rates in early career stages.

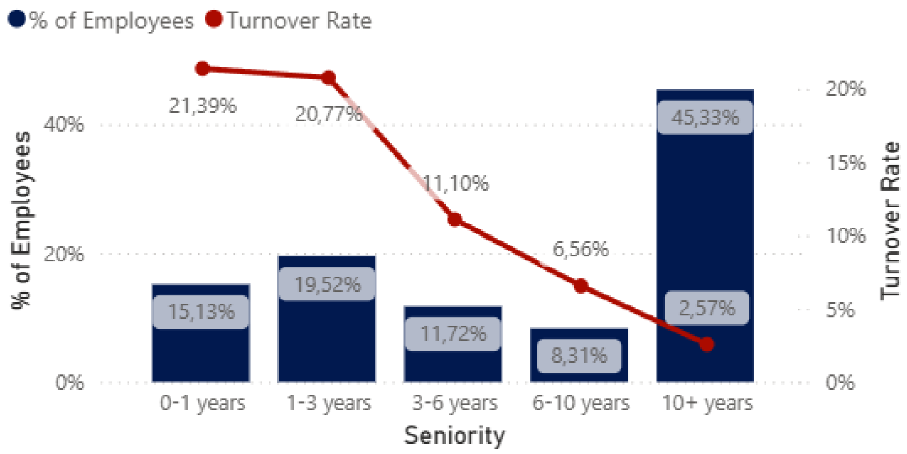

Figure 6 - Workforce Distribution and Turnover Rate by Seniority

Figure 7 illustrates a downward trend in turnover rates as employee age increases. Younger employees exhibit significantly higher turnover, suggesting that early career stages are more volatile, potentially due to career exploration, educational commitments or job transitions. Conversely, older employees tend to remain longer with the company, likely driven by job stability, career progression and fewer external job transitions. Most employees fall within the 35-50 age range, reinforcing the idea that mid-career professionals are more established in their roles, whereas employees above 50 represent the smallest proportion of the workforce. This trend aligns with the previously observed relationship between seniority and turnover, reinforcing the notion that both age and tenure play a critical role in employee retention.

The data also suggests a close relationship between age and seniority, as older employees tend to have higher tenure within the company. This is a logical outcome, given that seniority naturally increases with time and employees who remain longer in the organization tend to progress into more stable career paths. This pattern is further illustrated in Figure 38, which explicitly depicts how seniority levels correlate with employee age distribution, reinforcing the observed trends in Figure 6 and Figure 7.

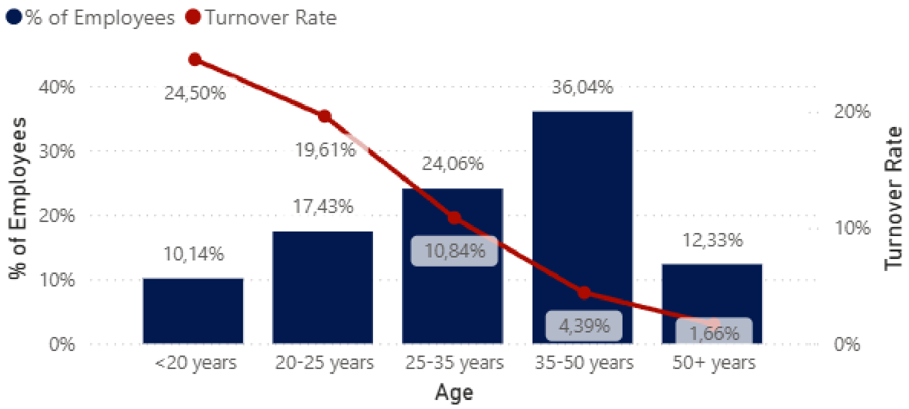

Figure 7 - Workforce Distribution and Turnover Rate by Age

The relationship between contract type and turnover rate reveals a stark contrast between permanent and fixed-term employees. As seen in Figure 8, employees with permanent contracts make up the majority of the workforce (69.44%) and exhibit a relatively low turnover rate of 5.54%. This suggests that job security plays a key role in employee retention, as those with stable contracts are less likely to leave the organization.

Conversely, employees with fixed-term contracts account for 30.56% of the workforce but experience a significantly higher turnover rate of 21.12%. This is likely due to the temporary nature of their employment, fewer career opportunities and the expectation that their contract may not be renewed. The contrast between the two groups highlights the impact of contract stability on workforce retention and suggests that policies aimed at converting high-performing fixed-term employees into permanent staff could help reduce turnover.

This pattern is further examined in Figure 39, which illustrates how contract types are distributed across different age groups.

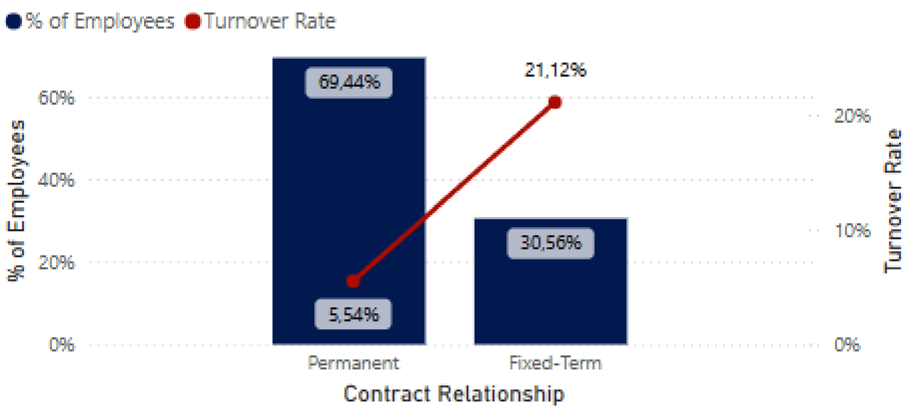

Figure 8 - Workforce Distribution and Turnover Rate by Contract Type

Figure 9 illustrates the relationship between annual salary (expressed as a multiple of the monthly minimum wage) and turnover rates. This standardized approach allows for a clearer comparison of salary levels across different time periods, mitigating the effects of inflation and changes in minimum wage policies. Given that annual salary includes 12 monthly wages plus holiday and Christmas bonuses (equivalent to two additional monthly wages), the lowest possible value observed in this dataset is 14 times the monthly minimum wage, assuming all records are legally compliant and accurately reported.

A strong inverse relationship is observed between salary and turnover. Employees earning 14-15 times the monthly minimum wage exhibit the highest turnover rates (17.84%). Notably, this salary range accounts for more than 40% of the dataset, reinforcing the idea that a significant portion of the workforce is in lower-wage positions, which are more prone to higher turnover. Expanding this view, over 80% of employees earn up to 18 times the monthly minimum wage, meaning that the vast majority of the workforce falls within a relatively low salary bracket, where turnover remains substantially higher compared to higher-paid employees.

As salaries increase, turnover rates decline significantly. Employees earning above 18 times the monthly minimum wage demonstrate much lower turnover rates, suggesting that higher compensation contributes to job stability, increased job satisfaction and long-term retention. This could be due to better career growth prospects, enhanced benefits or the fact that higher-paying positions are generally associated with specialized skills and seniority.

Interestingly, there is a slight increase in turnover within the 20-30 times the minimum wage range. While this remains a minor fluctuation, it may suggest that employees in this salary bracket have stronger external opportunities, potentially moving to better-paying roles in other companies. Another possible explanation is that employees in this range have greater expectations regarding career progression and if these expectations are not met, they may be more likely to seek alternatives. Figure 41 further explores salary distribution across different role clusters, providing additional insight into how wage levels vary depending on job functions within the company.

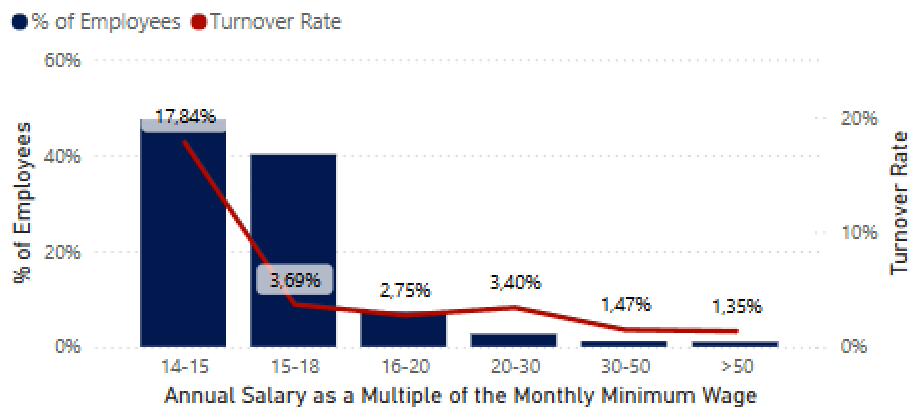

Figure 9 - Workforce Distribution and Turnover Rate by Standardized Wage

Figure 10 presents the distribution of employees based on their academic qualifications and the corresponding turnover rate. The majority of employees have secondary education (51.18%), followed by those with primary education (41.64%), while only 7.19% hold a higher education degree. The turnover rate increases with the level of education, starting at 7.58% for primary education, rising to 11.91% for secondary education and reaching its highest point at 14.56% for higher education. This trend suggests that employees with higher qualifications may have greater external opportunities, leading to higher mobility and a greater likelihood of leaving the company. Conversely, employees with lower educational backgrounds may have fewer job alternatives and, therefore, a stronger tendency to remain in their positions.

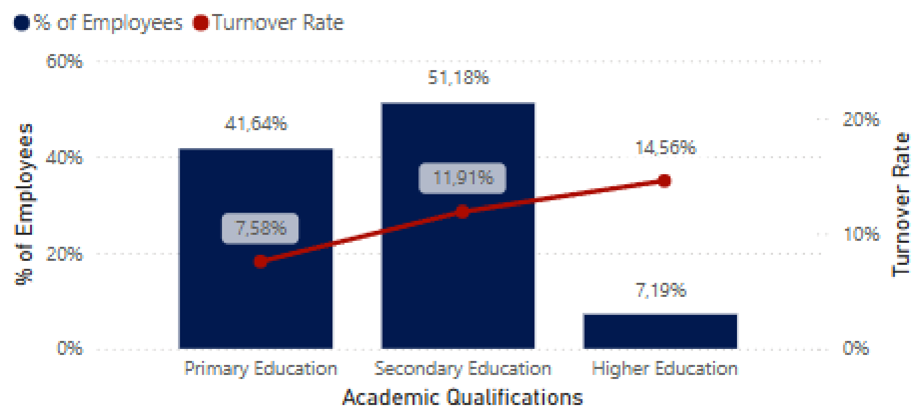

Figure 10 - Workforce Distribution and Turnover Rate by Academic Qualifications

Figure 11 illustrates the relationship between workload type and turnover rate, showing a clear distinction between full-time and part-time employees. The majority of employees, 67.98%, work full-time, whereas 32.02% are part-time. However, the turnover rate differs significantly between these two groups, with full-time employees having a relatively low turnover rate of 5.97%, while part-time employees exhibit a much higher turnover rate of 19.49%.

This disparity suggests that part-time employees are more likely to leave the company compared to their full-time counterparts. Several factors could contribute to this trend, including the typically lower job security, fewer career advancement opportunities and a greater likelihood of part-time workers using these positions as transitional roles before moving on to full-time jobs elsewhere. Additionally, part-time roles often attract students or individuals seeking supplementary income, making them more prone to voluntary turnover as their circumstances change.

On the other hand, full-time employees tend to have greater job stability, possibly benefiting from better compensation, benefits and career progression opportunities, which may contribute to a lower turnover rate. This insight highlights the need for distinct retention strategies for part-time employees, especially if reducing turnover within this group is a priority for the organization.

Figure 40 provides additional insights into the interaction between workload type and contract stability, illustrating how part-time and full-time roles are distributed across different contract types. Given the turnover patterns observed in Figure 8 and Figure 11, understanding this relationship may help clarify whether contract type and workload act as overlapping factors influencing employee turnover.

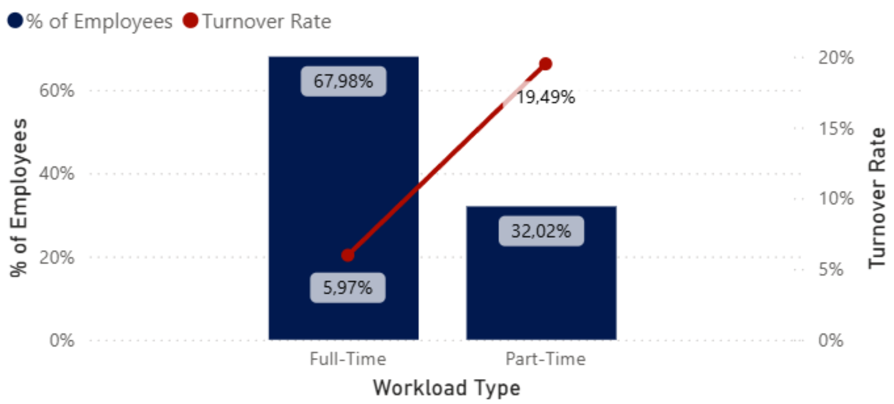

Figure 11 - Workforce Distribution and Turnover Rate by Workload Type

Figure 12 shows the distribution of employees and corresponding turnover rates by performance level (rated on a scale from 1 to 5). The vast majority of employees — over 80% — are evaluated with a performance score of 3, making it the dominant category within the organization. This central tendency may reflect a conservative or standardized evaluation culture.

Turnover rates follow a curved trend, peaking at the mid-performance level (3) with 20.81% and decreasing as performance moves toward either extreme. Employees rated as 4 or 5 exhibit lower turnover (12.59% and 7.63%, respectively), which aligns with expectations, as high-performing individuals are typically more engaged or may benefit from retention efforts. On the other end, although employees rated 1 or 2 represent a minimal fraction of the population, their turnover rates are non-negligible (9.39% and 16.25%, respectively), which could indicate some level of voluntary or involuntary attrition among lower performers.

The fact that most evaluations cluster at level 3, with deviations toward 1 or 5 occurring only in a minority of cases, suggests that performance ratings are only significantly adjusted when an employee is perceived as either clearly underperforming or excelling. As such, scores at the extremes may serve as strong indicators for key workforce decisions, such as promotions or exit prioritization.

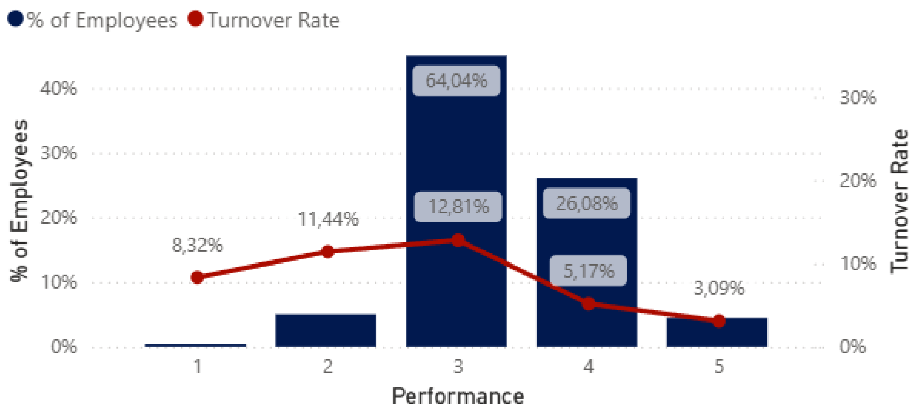

Figure 12 - Workforce Distribution and Turnover Rate by Employee Performance

Figure 13 presents the distribution of employees and corresponding turnover rates by potential level (rated on a scale from 1 to 3). The majority of employees — nearly 80% — are rated with a potential score of 1, followed by 16% at level 2 and just under 5% at level 3. This skewed distribution suggests that higher potential ratings are reserved for a relatively small portion of the workforce, possibly reflecting stricter evaluation criteria or a conservative approach in talent identification.

Turnover rates show a clear decreasing trend as potential increases. Employees with a potential rating of 1 exhibit a turnover rate of 11.54%, which drops to 5.47% for level 2 and 4.12% for level 3. This downward pattern suggests that higher-rated employees are more likely to stay within the organization, potentially due to stronger engagement, growth opportunities or retention initiatives.

The rarity of levels 2 and 3 highlights their possible role in identifying key talent segments for succession planning or career development. As such, the potential variable may be particularly useful in differentiating *personas* within the optimization model, helping to inform strategic decisions such as promotions, role reallocations and targeted retention efforts.

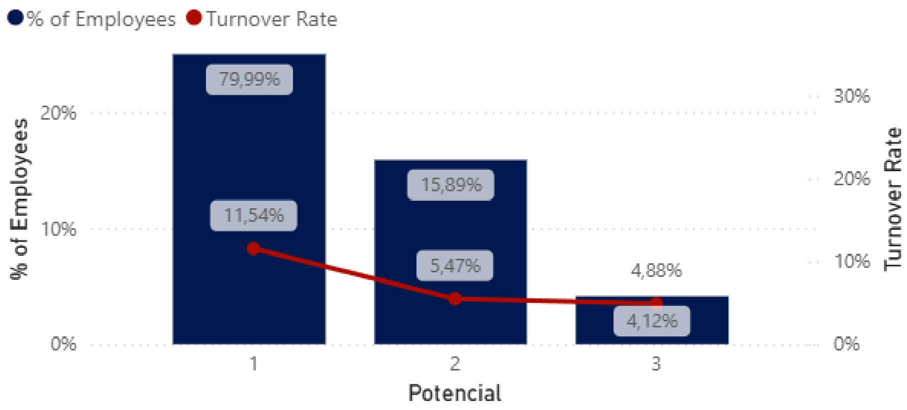

Figure 13 - Workforce Distribution and Turnover Rate by Employee Potential

To further support these findings, additional analyses were conducted on the interactions between tenure, contract type, workload and salary distribution. These complementary insights can be found in Appendix A – Supporting Data and Extended Analysis.

3.3 Final Remarks

This chapter outlines the core problem addressed in this dissertation — the high levels of employee turnover within the company and the limited effectiveness of current heuristic-based approaches in accurately forecasting future workforce needs. The exploratory data analysis revealed a rich and granular dataset, with variables that offer significant potential to enhance turnover prediction, despite certain challenges such as class imbalance and structural variability across stores, roles and regions.

Furthermore, the dual nature of the predictive task — individual-level forecasting for the first year and aggregate-level forecasting for subsequent years — reinforces the need for flexible and robust models tailored to different planning horizons. These findings underscore the limitations of the current forecasting methodology and provide a strong rationale for adopting a data-driven machine learning approach, which will be detailed in the following chapter.

4. Methodology for Turnover Prediction

The objective of this chapter is to describe the methodology used to develop a turnover prediction model that forecasts the voluntary departure of employees over a five-year horizon. Although the predictive model was designed with potential integration into a broader workforce optimization framework, this chapter — and the dissertation as a whole — focuses exclusively on the development, implementation and evaluation of the predictive component. Involuntary exits, such as layoffs or contract expirations, are excluded from the scope, as they are driven by different organizational dynamics.

Two predictive models were developed, each addressing a different time horizon. The first model forecasts turnover for the following year based on the current characteristics of each employee. In contrast, the second model estimates turnover over the subsequent four years (i.e., from the second to the fifth year), providing aggregated forecasts by *persona*, role cluster, store type and region rather than individual-level predictions. Since the characteristics of future employees are not yet known, this second model required a structural adjustment. To manage this uncertainty, a *persona*-based approach was adopted, allowing the model to generalize the expected behavior of future hires based on representative employee profiles.

Personas are defined by a combination of variables: age, tenure, academic qualifications, contract relationship (fixed-term or permanent), workload type (full-time or part-time), performance (scale of 1 to 5) and potential (scale of 1 to 3). Each *persona* is also associated with a specific district x store type x role cluster combination, ensuring that employee profiles are tailored to the organizational and operational context of each business unit. These groupings were defined by the company to reflect internal policies, performance thresholds and development potential.

For example, *Persona 1* might represent employees aged under 25, with secondary education, a fixed-term contract, working part-time and rated with a performance of 4 and a potential of 2. According to company strategy, this *persona* may be considered eligible for vertical mobility (promotion), lateral movement (changing roles) or even flagged for exit if they do not align with long-term organizational needs. The exact value ranges for numerical variables (such as age, performance and potential), as well as the specific categories or groupings for categorical variables (such as contract type or academic qualifications), were defined by the company according to internal criteria and strategic objectives.

Although *personas* are not directly used in the turnover prediction model, they are critical in the subsequent optimization phase. This optimization model uses the *persona* structure to determine which profiles are prioritized for promotion, mobility or exit, aligning workforce transitions with strategic planning objectives.

Additionally, the prediction will be made year by year, considering only the final status of each employee at the end of each year. Internal dynamics that occur within each year, such as temporary employee movements or role adjustments, will not be considered in the predictions. Excluding these internal dynamics allows the model to maintain a simplified structure, focusing on the final status of employees, which has a more direct impact on retention strategies and replacement costs. In this way, the model becomes more efficient and tailored to the company's long-term needs, while also providing more stable and accurate predictions to support strategic decision-making. Moreover, to align the predictive outputs with the requirements of the workforce optimization framework, the individual-level predictions from the one-year model are subsequently aggregated by *persona*, role cluster, store type and region.

Given the large volume of data generated by the client company, a decision tree-based approach was chosen. This type of model is particularly suitable for handling heterogeneous data, including both categorical and numerical variables and for providing transparent, interpretable outputs, which is crucial in human resource management contexts.

Among decision tree-based models, XGBoost was selected due to its efficiency and boosting optimization capabilities. XGBoost is particularly valuable when working with large-scale, high-dimensional and complex data, where it balances predictive power with computational efficiency.

As discussed in the Data Overview and Exploratory Analysis section, the dataset contains nearly 200,000 records, covering several years and including numerous features that capture both employee characteristics and organizational factors. The analysis also identified the presence of some natural outliers, particularly in salary variables, reflecting genuine differences between employee profiles (e.g., managerial vs operational roles). These outliers were retained to preserve the representativeness of the dataset.

Although boosting algorithms, including XGBoost, can be sensitive to outliers in certain contexts, tree-based models are generally less affected by outliers in input features than models relying on distance-based calculations (e.g., linear regression or k-nearest neighbors). This is because decision trees split data based on thresholds, meaning extreme values are treated the same as any other value falling into the same side of the split. This characteristic, combined with XGBoost's regularization mechanisms, helps mitigate the potential influence of outliers on the final model.

4.1 Data Preparation

4.1.1 Data Sources

The data used in this study were extracted from the company's internal Human Resources platforms. These platforms record detailed information about employees, including data on their role, performance, compensation and history. These platforms provide continuously updated data, ensuring an accurate and current picture of the workforce.

Additionally, *personas* were defined by the company as strategic groupings of employees based on role clusters, stores or regions, considering specific criteria such as performance, potential and eligibility for mobility. These *personas* allow the company to better plan workforce transitions and optimize staffing decisions. Though the *personas* do not directly feed into the turnover prediction model, they serve as a basis for future workforce planning and the optimization model that will be applied after the turnover prediction.

External variables, such as unemployment rate, inflation rate, average wage variation and minimum wage variation, were obtained from reliable public sources, such as the National Statistics Institute (INE). These variables were included to account for macroeconomic factors that may influence employee turnover decisions.

4.1.2 Data Integration and Consolidation

The data from various sources were integrated into a master table, where each row represents an employee in a given year, identified by the key "employee x year". This structure allows tracking each employee's history throughout the analysis period, facilitating the application of predictive models.

Table 2 - Dataset Structure for Turnover Prediction

employee code	year	predictor *1*	predictor 2	...	predictor *n-1*	predictor *n*	turnover
100001	2015	pred_1_v1	pred_2_v1	...	pred_n-1_v1	pred_n_v1	0
100001	2016	pred_1_v2	pred_2_v2	...	pred_n-1_v2	pred_n_v2	1
100002	2015	pred_1_v3	pred_2_v3	...	pred_n-1_v3	pred_n_v3	1

- Employee Code: Unique identifier for each employee.
- Year: Year for which the data were recorded.
- Predictors: These columns include various factors that could influence the probability of turnover.

- turnover: The target variable that indicates whether the employee experienced turnover (1) or stayed (0) in the company.

It is important to note that the value in the *turnover* column refers to the employee's turnover status in the year following the one listed in the year column. For example, if *year = 2022* and *turnover = 1*, it means that the employee left the company in 2023. This approach is used because the goal of the model is to forecast turnover in the upcoming years based on historical data up to the present.

This structure facilitates not only the historical tracking of each employee but also the input of various predictors into the model. In a scenario like 2024, the objective is to predict the probability of turnover for the upcoming five years using the information available up to that point.

4.1.3 Variable Structuring

To improve clarity and facilitate understanding, the variables were divided into five main categories. Below are the key variables used in the predictive model, along with their descriptions:

1. Current Position
    - Store Type: Refers to the different types of stores or divisions within the company.
    - Role Cluster: Describes the employee's function within the company.
    - Store Region and Employee Region: The geographical location of the store and the employee's residence.
    - Store District and Employee District: The district where the store and employee are located.
    - Contract Type: Indicates whether the employee has a fixed-term or indefinite contract.
    - Organizational Unit Change: Refers to the number of changes in the organizational unit throughout the year.
    - Leadership Change: The number of leadership changes the employee experienced during the year.
    - Team Exits under same Leader: The number of team members under the same leader who left voluntarily in that year.
    - Workload Type: Whether the employee works full-time or part-time.
2. External Factors
    - Average Wage Variation: The variation in average wages compared to the previous year.
    - Minimum Wage Variation: The variation in the minimum wage compared to the previous year.
    - Inflation Rate: Measures the general price variation in the economy.

- Unemployment Rate: Reflects labor market conditions.

3. Compensation and Performance
    - Theoretical Fixed Compensation and Variable Target Compensation: Annual compensation adjusted to the monthly minimum wage.
    - Target Achievement: The percentage of variable target achievement.
    - Change in Target Achievement: The difference in target achievement compared to the previous year.
    - Performance: The employee's performance, evaluated on a scale of 1 to 5.
    - Potential: The employee's potential, evaluated on a scale of 1 to 3.
    - Performance Change and Potential Change: Performance and potential variations compared to the previous year.
    - Extra Compensation: Additional compensation for extra hours or strenuous work.
    - Time Since Last Promotion: Time since the employee's last promotion.
    - Average Workload: The average weekly workload.
    - Workload Variation: The standard deviation of the employee's weekly workload.
    - Absence Hours and Team Absence Hours: The number of hours of absence for the employee and the team.
    - Training Hours: The number of training hours the employee has received.
4. Employee Characteristics
    - Age: The employee's age.
    - Tenure: The employee's length of service in the company.
    - Academic Qualifications: Refers to the highest level of education attained by the employee.
    - Number of Dependents: The number of dependents the employee has.
    - Marital Status: The employee's marital status.
5. Leadership Performance
    - Self-Assessment by the Leader: The leader's self-evaluation of their performance.
    - Team Assessment of Leadership: The evaluation of leadership made by the team members.
    - Leadership Assessment Difference: The difference between the leader's self-assessment and the team's evaluation.

Some variables underwent specific processes to be accurately represented in the model. These transformations ensure a better reflection of the underlying dynamics. Below are the key variable adjustments:

- Wage Variation: Variables like *average wage variation* and *minimum wage variation* were calculated as the percentage variation between the current year and the previous year using the formula:

$$Wage\ Variation = \frac{Current\ Year\ Value - Previous\ Year\ Value}{Current\ Year\ Value} \quad (4.1)$$

    - Advantage: Calculating wage variations as a percentage between consecutive years allows the model to capture not just absolute changes in wages, but also how significant those changes are relative to previous values. This approach better reflects economic trends, such as inflation or wage stagnation and helps to normalize the data over time. It highlights the impact of wage fluctuations, giving the model a better understanding of how changing economic conditions might influence an employee's decision to stay or leave.
- Compensation Normalization: The variables *theoretical fixed compensation* and *variable target compensation* were divided by the minimum wage for each respective year.
    - Advantage: By dividing *theoretical fixed compensation* and *variable target compensation* by the minimum wage for each year, the model accounts for inflation and purchasing power. This normalization is crucial for comparing compensation across different time periods since a higher nominal salary in a later year might not represent a true increase in value due to inflation. This adjustment ensures that the model is evaluating salaries in real terms, making it more accurate in assessing the potential impact of compensation on employee turnover.
- Target Achievement: This represents the achievement rate of the target variable compensation, calculated as:

$$Target\ Achievement = \frac{amount\ (incentives)}{Variable\ Target\ Compensation} \quad (4.2)$$

    - Advantage: Calculating target achievement as the ratio of incentives to variable target compensation helps quantify how well employees are performing relative to their variable compensation goals. This metric can reveal patterns in employee engagement and motivation. For instance, consistent underperformance might correlate with a higher likelihood of turnover, while exceeding targets may indicate stability.

- Change in Target Achievement: The variation in target achievement between consecutive years, calculated as:

$$Change\ Target\ Achievement = \frac{Target\ achievement\ [n]}{Target\ achievement\ [n-1]} - 1 \tag{4.3}$$

  - Advantage: By calculating the percentage change in target achievement between consecutive years, the model can track whether an employee's performance is improving or declining. Sudden drops in target achievement might signal dissatisfaction or disengagement, which can serve as early indicators of turnover. This dynamic view is more insightful than simply tracking absolute performance.

- Performance and Potential Changes: The variables *performance change* and *potential change* capture the change in performance and potential by subtracting the previous year's score from the current year's score:

$$Performance/Potential\ Changes\ = Current\ Year\ Score - Previous\ Year\ Score \tag{4.4}$$

  - Advantage: Using the difference between the current and previous year's performance and potential scores allows the model to detect shifts in an employee's career trajectory. A consistent decline in performance or potential might indicate burnout or a lack of future opportunities, whereas improvement could signal increased retention likelihood. The model, therefore, gets a clearer sense of an employee's momentum rather than just a static snapshot.

- Additional Compensation: This variable was calculated as the ratio of extra compensation to fixed compensation:

$$Additional\ Compensation = \frac{Total\ Extra\ Compensation}{Fixed\ Compensation} \tag{4.5}$$

  - Advantage: Calculating the ratio of extra compensation to fixed compensation provides insight into how much additional workload or strenuous conditions are being compensated for. If employees are regularly receiving a high amount of extra compensation, it could indicate stressful work conditions. This can be a key factor in predicting turnover, as overwork is often a precursor to voluntary exits.

4.1.4 Variable Classification

The variables used in the model are further classified as either categorical or numerical, as presented in the table below:

Table 3 - Variables Classification for Turnover Prediction Model

Categorical Variables	Numerical Variables
Store Type	Organizational Unit
Role Cluster	Leadership Change
Store Region	Team Exits
Employee Region	Average Wage Variation
Store District	Minimum Wage Variation
Employee District	Inflation Rate
Contract Type	Unemployment Rate
Workload Type	Theoretical Fixed Compensation
Marital Status	Variable Target Compensation
Turnover (target variable)	Target Achievement
	Change in Target Achievement
	Performance
	Potential
	Performance Change
	Potential Change
	Extra Compensation
	Time Since Last Promotion
	Average Workload
	Workload Variation
	Absence Hours
	Team Absence Hours
	Training Hours
	Age
	Tenure
	Number of Dependents

4.2 Data Cleaning and Processing

Given the characteristics of the dataset and the selection of XGBoost as the predictive model, extensive data cleaning was not required.

XGBoost handles missing data natively, as the algorithm automatically determines the optimal split direction (left or right) for records with missing values during the tree construction process. This mechanism allows missing data to be directly incorporated into the model without requiring explicit imputation.

Regarding outliers, the exploratory analysis identified some natural outliers, particularly in salary, reflecting genuine differences between employee profiles across distinct roles and hierarchical levels. These outliers were retained to ensure the dataset preserved the natural variability of the workforce, as they reflect real organizational patterns rather than data errors.

Tree-based models, including XGBoost, are generally less sensitive to outliers in input features because they split data at specific thresholds rather than being influenced by the absolute magnitude of values. This structural characteristic reduces the risk of extreme values disproportionately affecting model performance. Nevertheless, care was taken to assess that these outliers were legitimate observations and not data errors, thus ensuring their retention did not distort the overall dataset.

Additionally, as analyzed in the Data Overview chapter, the dataset contained a low proportion of missing values and a manageable number of natural outliers, particularly in features such as salary. Considering these factors and the capacity of XGBoost to handle both missing values and outliers effectively through its tree-splitting logic, it was decided to proceed with model training using the data as-is, without the application of additional cleaning procedures.

4.3 Predictive Models Design and Evaluation Methodology

4.3.1 One-year Model

The first model developed focuses on predicting employee turnover for the following year, using a one-year horizon. The selected approach is based on the XGBoost algorithm, implemented via the H2O library, due to its robustness and efficiency in handling large volumes of heterogeneous data. XGBoost was chosen specifically for its ability to handle missing values and outliers, as well as its capacity to avoid overfitting through boosting techniques.

Data Splitting

To ensure proper model validation, the data was split into two stages. In the initial stage, the data was divided into training and testing sets. Data from 2023 was reserved for testing, while all previous years — excluding 2020 and 2021 — were used for training. The decision to use 2023 as the test year, rather than performing a random percentage split, was motivated by the model's objective: to predict turnover for the following year based on historical data. By using the most recent year for testing, the model better simulates real-world forecasting scenarios.

Following this, in the calibration stage, 80% of the training data was used to train the model, while the remaining 20% was reserved as a calibration frame. This calibration set was used to fine-tune the predicted probabilities using Platt Scaling — a technique that adjusts the model's raw outputs into well-calibrated probabilities. A more detailed explanation of Platt Scaling and its role in the model's hyperparameters will be provided later in the methodology.

In addition, k-fold cross-validation was applied within the training set during the model construction phase to ensure robust internal validation. This technique helped assess the model's generalization performance before evaluating it on the 2023 data. Further details about the cross-validation setup are presented in the following subsection.

While rolling and growing window techniques are common in time series forecasting or models that explicitly account for temporal trends to evaluate performance across different temporal slices, they were not adopted in this project for a specific reason. The turnover prediction model developed in this study does not incorporate any explicit temporal features — such as the year of observation — as input variables. Instead, it relies exclusively on the intrinsic and extrinsic characteristics of each employee-year record.

As a result, the model is not designed to capture time-evolving trends, but rather to uncover consistent patterns linking employee attributes to turnover probability. Applying a windowed data splitting strategy would not offer additional insights into the model's generalization performance. On the contrary, it would reduce the amount of training data in each iteration, potentially limiting the model's ability to learn from the full range of workforce heterogeneity.

Therefore, a holdout strategy using the most recent year (2023) as the test set was adopted to reflect a realistic, forward-looking prediction scenario — fully aligned with the model's practical application in forecasting future turnover based on available historical data.

Model Construction

The one-year turnover model was built using H2OXGBoostEstimator, fine-tuned to optimize the prediction of voluntary employee exits. XGBoost is a decision-tree-based algorithm that uses boosting to progressively improve accuracy by correcting errors from previous iterations. Given that the target variable is binary (turnover vs. no turnover), the model utilizes a binomial distribution, which is well-suited for classification problems involving two possible outcomes. This choice of distribution aligns with the model's objective to effectively classify employees into two categories based on their likelihood to turnover.

Each hyperparameter was chosen to balance complexity and predictive power, with the following configurations:

- Cross-Validation (nfolds): 5-fold cross-validation was applied to the training data. This process divides the data into 5 parts, trains the model on 4 parts and validates on the remaining part, helping the model generalize better and avoid overfitting.
- Number of Trees (ntrees): This controls how many trees are built. The default of 50 trees was chosen, as recommended by H2O. This number provides a solid balance between complexity and accuracy without overfitting.
- Maximum Depth (max_depth): Limits tree depth. 6 was selected to prevent the model from becoming too specific while capturing enough data complexity. This depth avoids overfitting and works well across datasets.
- Learning Rate (learn_rate): This regulates the contribution of each tree. It was set to 0.2, allowing for gradual improvements. Given the heterogeneity of the dataset, this lower rate reduces the risk of overfitting.
- Subsampling (subsample): Defines the proportion of data used per tree. Set to 1, this ensured all data was used for each iteration. This approach maximizes the dataset's effectiveness, as the data was relatively balanced.
- Scale Pos Weight: This was adjusted to handle class imbalance between turnover and non-turnover cases. The imbalance factor was calculated based on the ratio of non-turnover to turnover instances across the training and calibration datasets. This ensured that turnover cases were appropriately weighted, improving the model's ability to predict turnover accurately.

$$Imbalance\ factor = \frac{\#Train\ dataset_{stay} + \#Calibration\ dataset_{stay}}{\#Train\ dataset_{turnover} + \#Calibration\ dataset_{turnover}} \tag{4.6}$$

- Calibration: The model was calibrated using Platt Scaling on 20% of the training data. Platt Scaling is a method that adjusts the model's predicted probabilities by fitting a logistic regression model to the output of the XGBoost classifier. This allows the raw outputs (logits) to be transformed into well-calibrated probabilities, making the model's predictions more reliable, especially in imbalanced datasets like this one. This fine-tuning ensured that the predicted probabilities were more accurate and aligned with real-world outcomes.

Each of these parameter choices is justified by aligning with the standard recommendations from H2O's XGBoost documentation. These defaults are widely accepted in practice and provide a solid foundation for constructing robust models.

4.3.2 Five-year Model

Five-year dataset structure and variables

The five-year turnover model is designed to forecast employee turnover over an extended period using *personas* to represent groups of employees with shared characteristics. These *personas* allow the model to generalize behavior without needing specific employee data for future years.

From the second year onward, changes such as employees being terminated, new hires, voluntary exits and role changes become inevitable. This makes it impractical to predict turnover at the individual employee level, since we don't know anymore the human resource composition by 100%. Instead, *personas* are used to account for average behavior across groups.

Given that, the structure to train and test the model is the following:

Table 4 - Dataset Structure for Five-year Model

year	*persona*	role cluster	district	store type	predictors	turnover
2016	P1	Cashier Operator	Braga	Medium	...	0.2
2016	P2	Cashier Operator	Braga	Medium	...	0
2016	P2	Flow Operator	Braga	Large	...	0.9

In the five-year model, *personas* are defined using key variables that capture essential employee characteristics relevant to turnover and mobility. The *persona*-defining variables — Contract Type, Workload Type, Academic Qualifications, Performance, Potential, Age and Tenure — were selected because they encapsulate core employee attributes necessary for identifying distinct employee groups. Each Role Cluster consists of multiple *personas* and each Store Type encompasses several role clusters. This structure allows for targeted predictions across various regions, aligning with the objective of forecasting aggregated turnover by role cluster and region, rather than individual turnover.

Dividing predictions by regions is essential, as the optimization model focuses on providing actionable insights regarding workforce needs at the regional level. By forecasting turnover rates for specific role clusters within regions, the model aligns with the strategic goal of managing workforce turnover in a way that supports efficient long-term planning. For example, the model can predict that a specific role cluster, such as the Fresh Counter in the Porto region, may need 20 hires in 2025 and 30 in 2026. However, it does not address store-level or individual movements, like transferring a specific employee from one store to another. This high-level approach ensures that the model serves regional planning needs without delving into individual-level actions, which would be impractical for long-term projections.

To create the five-year model, historical averages were applied to *persona*-defining variables. Specifically, numerical variables such as Performance, Potential, Age and Tenure were aggregated into averages, capturing the core trends within each *persona* over time. For categorical variables — Contract Type, Workload Type and Academic Qualifications — combinations were created across each role cluster, store type and district. This led to a comprehensive table structure with various combinations, ensuring that each category's distinct characteristics are maintained.

The turnover variable is now continuous to achieve more accurate values. By using an average turnover rate for each employee matching a given *persona*, the model can better represent turnover within groups. This approach is more precise than rounding turnover to binary values of 0 or 1, as it captures the nuanced likelihood of turnover across *personas*. It aligns with the model's goal of forecasting aggregated trends rather than individual outcomes, providing a clearer picture for long-term planning.

In this model, external factors and remuneration adjustments play a critical role in supporting long-term scenario planning. By treating variables like inflation, unemployment and wage changes as user-defined inputs, the model offers flexibility to explore different potential outcomes. Predicting these factors over multiple years would be overly complex and prone to error. Instead, allowing the client to modify these variables directly helps avoid inaccuracies while offering strategic insights.

By also incorporating remuneration as a modifiable predictor, the model helps simulate how various salary adjustments might impact employee retention. This enables the client to evaluate different cost-saving strategies and understand how long-term planning decisions might affect turnover.

For example, if the client anticipates a high inflation rate, they can assess whether adjusting wages for specific role clusters or regions could mitigate increased turnover risks. This structure ensures that the model remains adaptable and directly aligned with the client's strategic objectives over a five-year horizon.

So, as a resume, the variables that are taken into consideration in the five-year model are:

Table 5 - Variables Classification for Turnover Five-year Prediction Model

Categorical Variables	Numerical Variables
Store Type	Average Wage
Role Cluster	Minimum Wage
Store District	Inflation Rate
Academic Qualifications	Unemployment Rate
Contract Type	Theoretical Fixed Compensation
Workload Type	Performance
	Potential
	Age
	Tenure
	Turnover (target variable)

The decision to use absolute values for minimum and medium wages, instead of variations, is due to the practicality for long-term scenario planning. With absolute values, the client can directly input projected wage levels without needing to calculate annual changes, simplifying the evaluation of various salary strategies on turnover over the five-year period. This approach provides an intuitive and efficient way to simulate and assess the impacts of different remuneration strategies, supporting more strategic decision-making.

The five-year model aims to predict employee turnover over a longer timeframe using *personas* rather than individual employee data. Like the one-year model, it utilizes the XGBoost algorithm due to its effectiveness with heterogeneous data, handling of missing values and prevention of overfitting. This model forecasts turnover as a continuous variable (average turnover rate) across grouped *personas*, leveraging XGBoost's ability to handle both categorical and continuous target variables, making it well-suited for the nuanced needs of this extended prediction model.

Data Splitting

For the five-year model, the data splitting approach is adjusted to handle the aggregated nature of the model, where predictions are made at the *persona* level rather than the individual employee level. Since it's impractical to predict exact turnover at the employee level over five years, we use aggregated *persona* data. The data from 2023 is reserved for testing, allowing the model to be validated using the most recent data and simulating a realistic five-year forecast.

Unlike the one-year model, a calibration set is unnecessary, as we're dealing with a continuous target variable (average turnover rate) instead of a categorical one. Nonetheless, it's crucial to ensure that the training data reflects the period being forecasted, capturing general turnover trends over time accurately.

Model Construction

The five-year turnover model builds on the same XGBoost framework as the one-year model, retaining many of the core hyperparameters due to their effectiveness in handling complex, heterogeneous datasets. However, given the distinct requirements of forecasting over a longer timeframe and predicting a continuous variable, certain adjustments were made to enhance the model's performance:

The five-year turnover model builds on the same XGBoost framework as the one-year model, retaining many of the core hyperparameters due to their effectiveness in handling complex, heterogeneous datasets. However, given the distinct requirements of forecasting over a longer timeframe and predicting a continuous variable, certain adjustments were made to enhance the model's performance:

- Number of Trees (ntrees): Increased from 50 to 100. With the five-year model forecasting over a longer period, having more trees allows for greater complexity and enables the model to capture intricate patterns within the aggregated data. This depth is particularly useful when trying to model nuanced, long-term trends across diverse *personas*, where more decision paths contribute to higher accuracy in continuous predictions.
- Maximum Depth (max_depth): Raised from 6 to 10. By increasing the tree depth, the model can explore deeper interactions between variables, which is essential for a model operating on *persona*-level aggregations rather than individual data. This allows the model to capture detailed relationships within the data that might influence turnover over a five-year horizon, such as how different *persona* traits interact with external factors like inflation or wage changes.
- Learning Rate (learn_rate): Adjusted from 0.2 to 0.3. With the longer forecast horizon and continuous target, a slightly higher learning rate enables faster adjustments in each boosting round, helping the model converge more quickly on meaningful patterns in the data. This adjustment strikes a balance between allowing the model to learn effectively from the data while avoiding overfitting, which is crucial given the model's long-term perspective.

- Distribution (distribution): Set to gaussian, reflecting the continuous nature of the target variable. Unlike the one-year model, which forecasts a binary outcome (turnover or no turnover), the five-year model predicts an average turnover rate as a continuous variable. The gaussian distribution is optimal for this type of output as it models continuous values and helps the model generate smoother, more realistic predictions over time.
- Calibration: Not used in this model. Calibration is generally applied to refine probability estimates in categorical models, particularly in imbalanced datasets. Since the five-year model's target is a continuous average turnover rate, calibration would not provide additional benefit. The continuous nature of the predictions aligns naturally with the model's objective to estimate average turnover, making calibration unnecessary.

By tailoring these hyperparameters, the five-year model remains robust enough to handle large, aggregated datasets while offering the precision required for long-term turnover forecasts. The other hyperparameters remain aligned with those of the one-year model, as they already support the model's ability to handle large volumes of diverse data effectively.

4.3.3 Evaluation

Both models were evaluated using Mean Absolute Error (MAE), Brier Score and detailed subgroup analyses to compare predicted and actual turnover. Additionally, the one-year model was also assessed using AUC-ROC, given its binary classification nature.

AUC-ROC

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate the model's ability to distinguish between turnover and non-turnover cases. The Receiver Operating Characteristic (ROC) curve plots the true positive rate (the proportion of correctly predicted turnover cases) against the false positive rate (the proportion of non-turnover cases incorrectly classified as turnover) across different probability thresholds. This curve illustrates the trade-off between sensitivity and specificity at various thresholds.

The AUC-ROC provides a single value summarizing the model's overall performance in distinguishing between classes. An AUC-ROC value closer to 1 indicates that the model has a high capacity to correctly differentiate between employees who leave and those who stay, while a value closer to 0.5 indicates performance no better than random guessing.

Mean Absolute Error (MAE)

The MAE was used to measure the average absolute errors between the predicted turnover and the actual values. This metric provides insight into how close the model's predictions are to the real turnover rates. The lower the MAE, the better the model's performance, as it means that the predicted values are closer to the actual outcomes.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (4.7)$$

The MAE was used to measure the average absolute errors between the predicted turnover probabilities and the actual turnover values. Although MAE is typically applied to continuous variables, it is also suitable in this context to evaluate both models — including the one-year prediction model, where the target variable is binary (0 or 1). This is because, even though the actual values are binary, the model outputs probabilities, representing the likelihood of each employee leaving. If the model was perfect, the predicted probabilities would always be either 0 or 1, matching the actual outcomes. By using MAE, we can assess how close the predicted probabilities are to the correct binary values, providing valuable insight into the model's overall performance.

Brier Score

The Brier Score was used to evaluate the accuracy of the probabilistic predictions produced by the models. It measures the mean squared difference between the predicted probabilities and the actual outcomes. As such, it is particularly suitable for assessing the calibration of classification models where the output is a probability rather than a discrete class.

$$Brier\ Score = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (4.8)$$

In the one-year model, where the target variable is binary (0 or 1), the Brier Score captures how close the predicted probabilities are to the actual binary outcomes. A lower Brier Score indicates better calibration and higher confidence in the predicted probabilities. While AUC-ROC focuses on the model's ranking ability, the Brier Score offers complementary information by assessing whether the predicted probability values are well-calibrated — for example, whether a 70% predicted probability of turnover corresponds to approximately 70% of actual turnover in similar cases.

In the five-year model, where turnover is expressed as a continuous rate at the group level, the Brier Score also remains applicable. It continues to measure the squared error between predicted and observed rates, functioning as a calibration metric even in the context of aggregated predictions.

Subgroup analysis

In addition to the overall metrics, the model was analyzed in detail by focusing on specific categorical variables. This analysis compared the predicted turnover and the actual turnover within specific subgroups to ensure that the model captures all nuances across different groups of employees.

For instance, while the total predicted turnover might align well with the actual total turnover, significant differences could occur within specific subgroups such as different role clusters. The goal is not only to predict the overall turnover but also to assess the accuracy within each category, ensuring that the model performs well across diverse employee functions.

This detailed analysis by subgroups (such as role clusters, *personas* and districts) is also essential to feed the workforce optimization model, which relies on accurate turnover predictions across these different segments to support strategic workforce planning decisions.

4.4 Final Remarks

This chapter presented the methodology developed to predict voluntary employee turnover across two time horizons: a short-term individual-level model (one-year) and a long-term aggregate-level model (five-year). The first model leverages detailed employee-level data to estimate the probability of turnover in the following year, offering granular, operational insights. The second model, based on aggregated *personas* across store types, role clusters and districts, supports strategic planning by forecasting expected turnover volumes over a longer horizon.

Both models were implemented using the XGBoost algorithm, selected for its ability to handle heterogeneous and imbalanced data while maintaining scalability and interpretability — essential features in a complex, data-rich environment such as retail. The methodology incorporated class balancing strategies, careful variable design and robust evaluation metrics — including MAE, Brier Score and AUC-ROC — complemented by subgroup analyses to ensure performance across diverse workforce segments.

Together, these methodological choices aim to support informed, data-driven decision-making at both tactical and strategic levels, aligning predictive power with real-world workforce planning needs. The next chapter builds upon this foundation by presenting the predictive performance of each model and identifying the most influential turnover drivers, providing actionable insights to improve employee retention and organizational stability.

5. RESULTS ANALYSIS AND DISCUSSION

This chapter presents the results of the predictive models developed to estimate voluntary turnover over one- and five-year horizons. It begins by evaluating the performance of the one-year model, which predicts individual-level turnover based on detailed employee features. Subsequently, the chapter discusses the results of the five-year model, which forecasts aggregate turnover at a *persona*-level across different organizational segments. For both models, the analysis considers predictive accuracy, error distribution across subgroups and the importance of key features driving turnover. These results are then interpreted in light of the project's objectives, providing insights into the model's strategic applicability in workforce planning.

5.1 One-year Model

5.1.1 Assessment of Predictive Performance

The one-year turnover prediction model was evaluated using four key metrics: AUC-ROC, Mean Absolute Error (MAE), BIAS and Brier Score. Each of these metrics captures different aspects of model performance, from its ability to rank employees by turnover risk to how well the predicted probabilities align with actual turnover rates. Table 6 summarizes the results.

Table 6 - Overall Performance Metrics for the One-Year Turnover Prediction Model

AUC-ROC	BIAS	MAE	Brier Score
93 %	1.17 pp	10.36 pp	0.046

An AUC-ROC of 93% suggests that the model effectively ranks employees based on their likelihood of leaving. This means that individuals with higher predicted turnover probabilities are indeed more likely to leave. However, a high AUC does not guarantee well-calibrated probabilities, meaning that while the model can order employees correctly, the actual probability values may still require refinement.

The MAE of 10.36 percentage points (pp) indicates that, on average, predicted turnover probabilities deviate from actual outcomes by about 10 pp. Considering that the overall turnover rate is 11.9%, this error represents a substantial proportion of the expected value and is therefore considered high in this context. While some degree of error is to be expected in probabilistic models, such a deviation limits the model’s usefulness for accurate planning at the individual level and calls for careful aggregation and validation at the group level to ensure reliable workforce planning outputs.

To complement this, the Brier Score for the model is 0.046, indicating that the predicted probabilities are reasonably well-calibrated. This supports the model's suitability for aggregated forecasting, even if individual-level predictions carry higher error margins.

With this in mind, the overall turnover prediction for the test set can be visualized in Figure 14, which compares the model's forecast with the actual turnover rate for 2023.

The model predicted a turnover rate of 13.0%, compared to the actual turnover of 11.9%, resulting in a small overestimation of 1.1 pp (Figure 14). This suggests that the model successfully captures the overall trend in workforce attrition, making it a valuable tool for workforce planning. However, while the prediction is accurate at a global level, it is essential to further investigate how well the model performs across different employee segments.

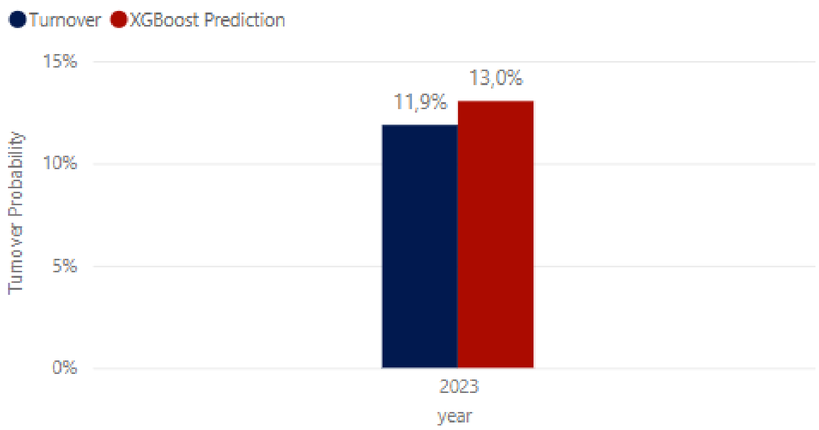

Figure 14 - Actual Turnover vs Turnover Prediction (One-year model)

While the global turnover prediction provides an overall assessment of the model's accuracy, it is essential to analyze its performance at the segment level. Since workforce planning decisions will be based on aggregated turnover estimates per district, store type, role cluster and *persona*, evaluating the model's effectiveness within these key groupings is crucial.

To assess this, the model's errors were analyzed at two levels of aggregation:

- District x Store Type x Role Cluster: A core level used in workforce planning.
- District x Store Type x Role Cluster x *Persona*: The most granular level, which directly feeds into the workforce optimization model.

At the district x store type x role cluster level, the model shows an improvement across all metrics compared to the individual-level evaluation (**Error! Not a valid bookmark self-reference.**). The MAE is reduced by more than half, the Brier Score decreases tenfold and the BIAS increases moderately from 1.17 pp to 1.93 pp. This indicates that aggregating predictions at this level helps smooth out individual variability, resulting in more stable and better-calibrated forecasts. These findings support the use of role cluster as a suitable level of granularity for strategic workforce planning.

At the district x store type x role cluster × *persona* level, however, the benefits of aggregation are less clear. While the MAE decreases to 9.17 pp, the BIAS increases to 5.31 pp, reflecting a stronger directional error. The Brier Score improves to 0.021, indicating better calibration than at the individual level, but remains substantially higher than the 0.004 observed at the role cluster level. This suggests that the model delivers less consistent and reliable predictions when applied at the *persona* level.

Table 7 – Turnover Prediction Model Performance at Aggregated Segment Levels (One-year model)

Level	BIAS	MAE	Brier Score
District x Store Type x Role Cluster	1.93 pp	4.42 pp	0.004
District x Store Type x Role Cluster x *Persona*	5.31 pp	9.17 pp	0.021

Given that the optimization model operates on *persona*-level forecasts, it becomes important to assess whether the observed discrepancies stem from model limitations or from specific characteristics associated with each employee segment. To support this analysis, the following section evaluates model performance across a set of relevant segments, including key *persona*-defining variables — such as academic qualifications, contract relationship, workload, performance, potential, tenure and age — as well as organizational segments like store type and role cluster. This analysis aims to identify potential sources of error and assess whether the current segmentation strategy sufficiently captures employee behavior.

Figure 15 compares the predicted turnover rates generated by the model with the actual turnover observed in 2023 across the top 10 districts, which together represent approximately 87.5% of the workforce.

In these core districts, prediction errors range from 1 to 3 percentage points — well below the lowest MAE recorded across previously evaluated segmentations (4.42 pp at the role cluster level). This indicates that, even though district-level predictions were not the direct target of model calibration, the results are highly accurate and stable across key regions.

The error is consistently positive, with the model overestimating turnover across all districts. For instance, the deviations reach 3 pp in Porto and Aveiro, while in districts like Lisboa, Faro, Braga and Leiria the gap is 1 to 2 pp. In Ilha da Madeira, the model shows an especially small deviation (1 pp), aligning closely with the observed turnover in that region.

This systematic overestimation mirrors the directional bias previously identified (BIAS = 1.93 pp), but the low magnitude of errors across districts reinforces the model's robustness for geographic segmentation. These results support the reliability of district-level predictions as an input for regionally targeted workforce planning and turnover mitigation strategies.

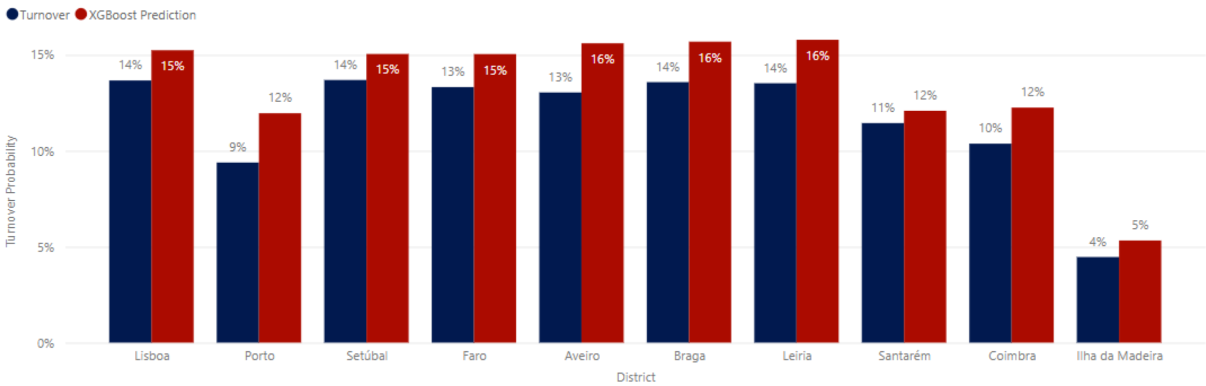

Figure 15 - Actual Turnover vs Turnover Prediction by Top 10 District (One-year model)

As seen in the Figure 16, the model once again performs better for the segments with higher employee representation. Predictions are closest for hypermarkets and supermarkets, where most employees are concentrated (approximately 80%). The difference between predicted and actual turnover is only 1 pp for hypermarkets and 2 pp for supermarkets, while for convenience stores, the gap increases to 3 pp.

As observed in other analyses, the model tends to systematically overestimate turnover, which could indicate a slight calibration bias but also offers a consistent pattern that can be addressed through postprocessing if necessary. Overall, the results remain reasonably accurate across store formats, particularly for those with the largest workforce impact.

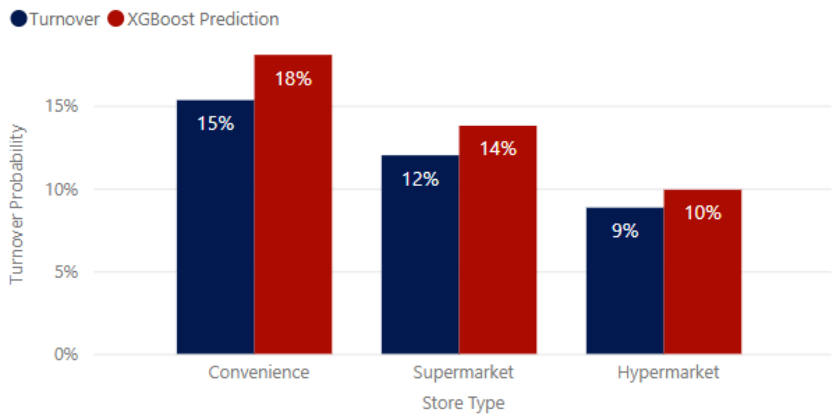

Figure 16 – Actual Turnover vs Turnover Prediction by Store Type (One-year model)

Figure 17 presents a comparison between the model's predicted turnover rates and the actual turnover observed in 2023, across the top 10 role clusters, which together represent approximately 90% of employees. For these core role clusters, the prediction error ranges from 1 to 3 percentage points, a result that is even more favorable than the overall MAE of 4.42 pp calculated at the aggregated role cluster level.

This demonstrates that the model performs particularly well for the majority of the workforce, capturing turnover behavior with a high degree of accuracy in the most representative segments. The fact that the errors remain small and consistent across these clusters strengthens the case for using this level of granularity in workforce planning.

The errors observed are also consistently positive, indicating a tendency to slightly overestimate turnover across all major role clusters. This pattern is coherent with the overall BIAS of 1.93 pp found at the role cluster level. Despite this directional deviation, the magnitude of the errors remains low and stable, reinforcing the model's robustness and its suitability for workforce planning at this level of segmentation.

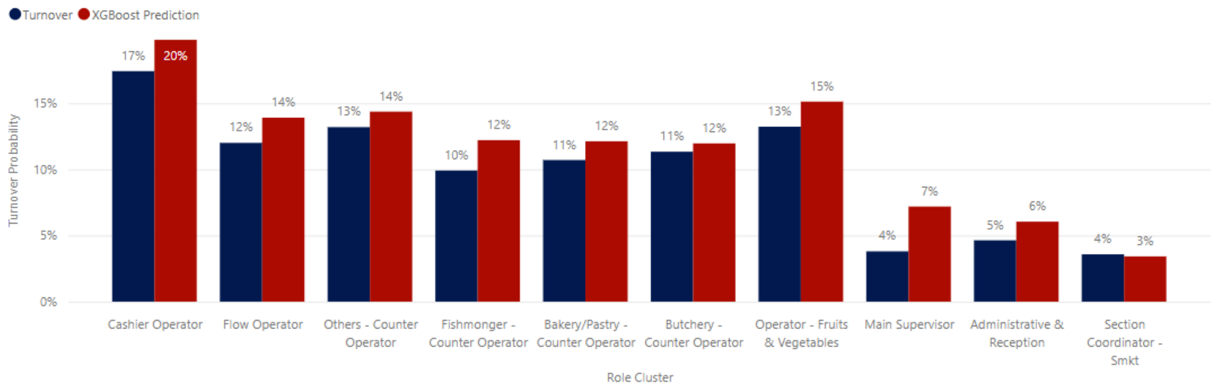

Figure 17 - Actual Turnover vs Turnover Prediction by Top 10 Role Cluster (One-year model)

As shown in Figure 18, the model successfully captures the turnover gradient across education levels, predicting higher turnover for employees with more advanced qualifications. The predicted values closely align with actual figures, differing by 1 pp for those with higher education and 2 pp for the other groups. This suggests the model is able to internalize meaningful behavioral patterns associated with academic background.

It is also worth noting that approximately 90% of employees have primary or secondary education, making the model's performance in these segments particularly relevant for workforce planning.

Once again, a slight systematic overestimation is observed across all education levels. However, the magnitude of error remains low and consistent, reinforcing the model's robustness when segmented by academic qualifications.

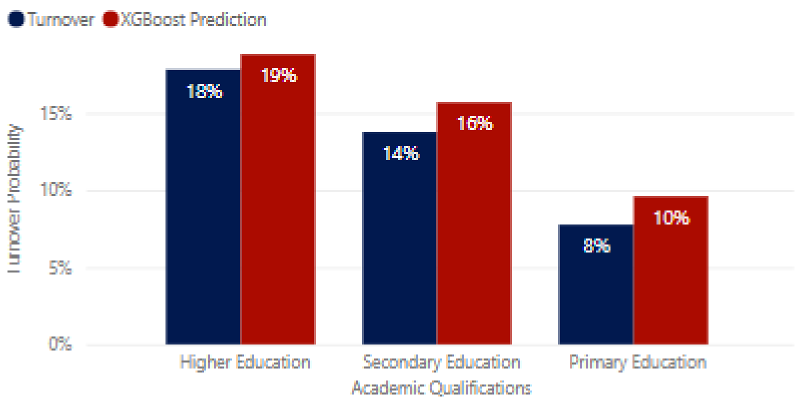

Figure 18 – Actual Turnover vs Turnover Prediction by Academic Qualifications (One-year model)

The model clearly distinguishes between the two contract types, capturing the significantly higher turnover rate among fixed-term employees (Figure 19). While it slightly overestimates turnover in both groups, the deviation is larger for the less common contract type: a difference of 4 pp for fixed-term contracts (28% predicted vs. 24% actual), compared to just 1 pp for permanent contracts (9% vs. 8%). Notably, approximately 80% of employees hold permanent contracts, while only 20% are on fixed-term contracts.

These results suggest that the model is able to capture the structural turnover differences associated with contract type, although the prediction error is more pronounced in the smaller group.

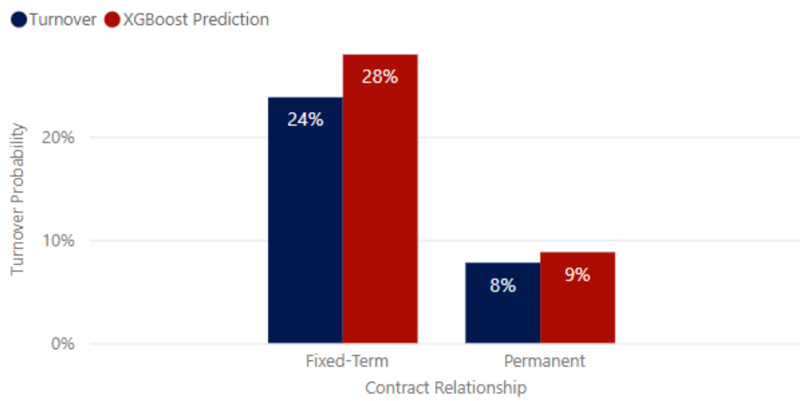

Figure 19 - Actual Turnover vs Turnover Prediction by Contract Relationship (One-year model)

Figure 20 compares the predicted and actual turnover probabilities for employees with part-time and full-time contracts. The model successfully differentiates between the two workload types, capturing the fact that part-time workers are more likely to leave the organization. However, the prediction error is higher for the part-time group, with a difference of 2 percentage points compared to the actual value. For full-time employees, the predicted turnover rate is overestimated by 2 pp as well. It is worth noting that full-time contracts represent approximately 70% of the workforce, meaning that the model's performance on this group is especially relevant for planning purposes.

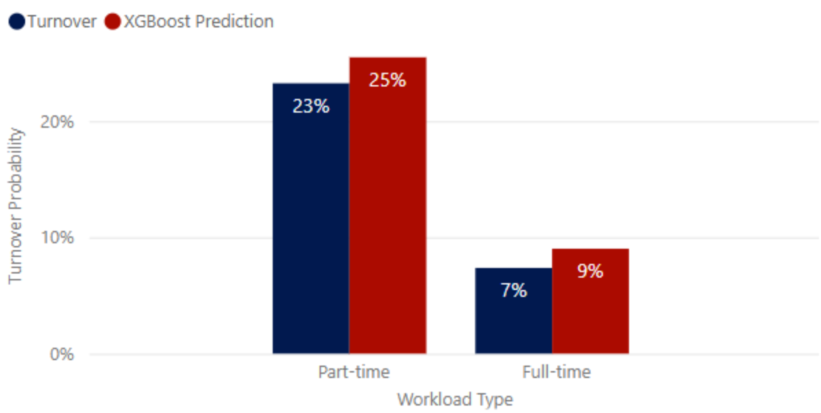

Figure 20 - Actual Turnover vs Turnover Prediction by Workload Type (One-year model)

Figure 21 illustrates the predicted and actual turnover probabilities across different age groups. The model accurately captures the decreasing trend in turnover as age increases, reflecting behavioral patterns typically observed in the retail sector. Younger employees, particularly those under 25, exhibit the highest turnover rates and the model reflects this by assigning them the highest predicted probabilities. Across all age groups, the model slightly overestimates turnover, with deviations of 2 and 4 percentage points in the younger segments, while maintaining near-perfect alignment for employees over 50. This reinforces the model's ability to detect relevant behavioral trends, even if some calibration refinements could improve its precision for younger profiles.

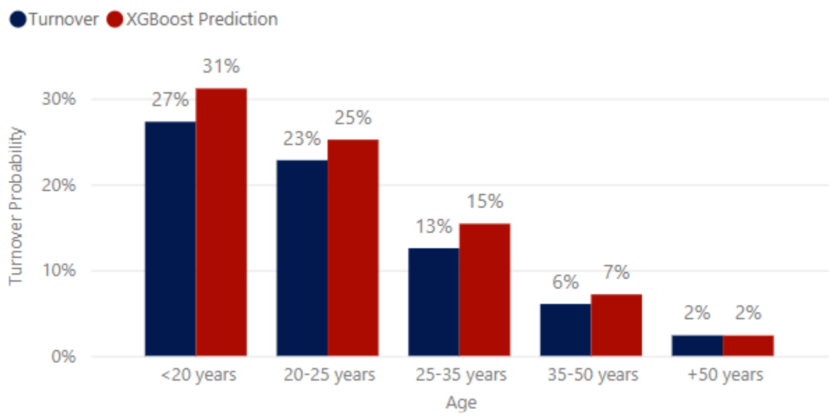

Figure 21 - Actual Turnover vs Turnover Prediction by Age (One-year model)

Figure 22 compares the model's predicted turnover probabilities with actual turnover rates across different seniority groups. The model successfully captures the overall declining trend in turnover as tenure increases — a well-documented behavior in the retail sector.

For employees with 0–1 years of tenure, the prediction is remarkably accurate, perfectly matching the actual turnover rate of 33%. This suggests that the model is highly effective in identifying early-stage turnover risk, which is particularly valuable for designing onboarding and early retention strategies.

However, the model shows a considerable overestimation for the 1–3 year group, predicting a turnover rate of 22% compared to an actual 14% — a difference of 8 pp. This deviation indicates that the model may struggle to capture the behavioral dynamics of mid-tenure employees, possibly due to greater variability or unobserved factors influencing turnover decisions at this stage.

For employees with more than 3 years of tenure, the prediction errors remain small, generally within 1 to 2 percentage points, further reinforcing the model's ability to reflect long-term stability patterns.

Overall, while the model accurately reproduces broad seniority-related turnover trends, the substantial overestimation in the 1–3 year range highlights a specific segment that may benefit from model recalibration or further variable exploration.

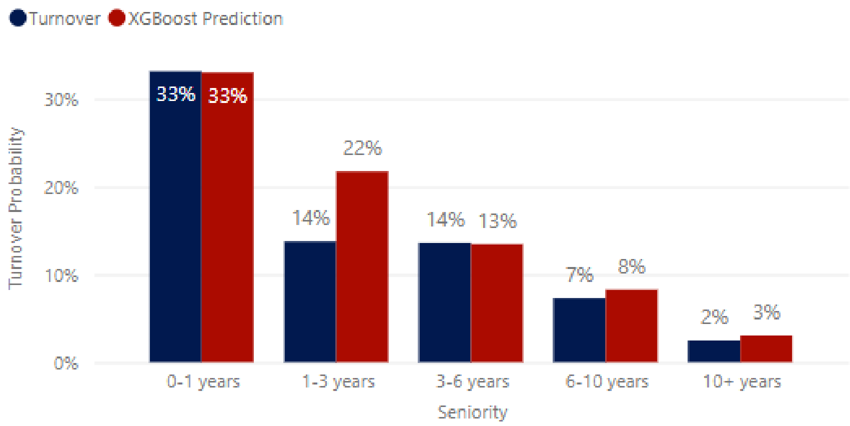

Figure 22 - Actual Turnover vs Turnover Prediction by Seniority (One-year model)

Figure 23 shows the comparison between the predicted and actual turnover rates by performance level (rated from 1 to 5) for the year 2023. The model successfully captures the expected downward trend in turnover as performance increases. Employees with lower performance scores tend to exhibit higher turnover, while high performers are more likely to remain with the organization.

Across all categories, the prediction errors are relatively small, ranging between 1 and 3 percentage points. The model performs particularly well for performance levels 2 and 5, with almost perfect alignment. For employees rated 3 and 4, the model slightly overestimates turnover, while for level 1, it underestimates it (8% predicted vs. 11% actual).

Overall, the model demonstrates strong accuracy across the performance scale, reinforcing its robustness in capturing turnover behavior along this critical *persona*-defining dimension.

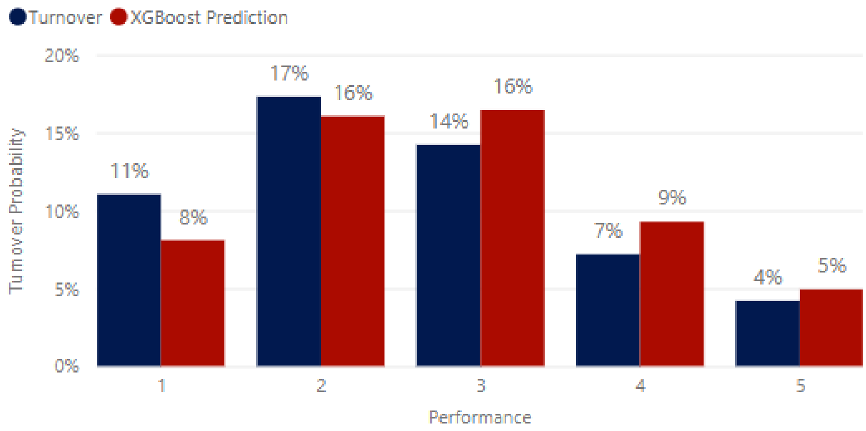

Figure 23 - Actual Turnover vs Turnover Prediction by Performance (One-year model)

Figure 24 presents the comparison between predicted and actual turnover rates by potential level (rated on a scale from 1 to 3) for the year 2023. The model captures the overall structure of turnover across potential levels, with higher turnover observed among employees rated with lower potential.

Employees rated with potential level 1 — the most common group — exhibit the highest actual turnover (13%), which is slightly overestimated by the model (15%). For levels 2 and 3, the prediction errors are also within 2 pp, with the model showing a mild overestimation in both cases.

Although the differences between predicted and actual values are modest, the model successfully reflects the inverse relationship between potential and turnover, reinforcing its ability to internalize meaningful behavioral patterns associated with long-term employee value.

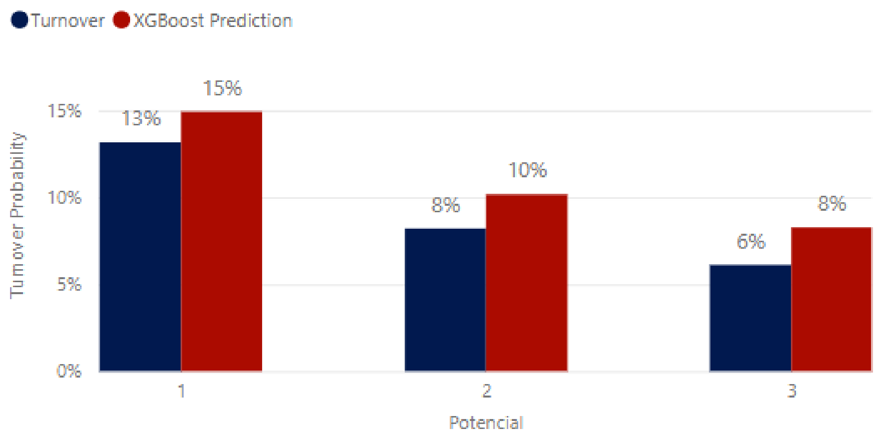

Figure 24 - Actual Turnover vs Turnover Prediction by Potential (One-year model)

Beyond group-level comparisons, it is also essential to understand which individual and contextual features most influence the model's predictions, as explored in the next section.

5.1.2 Drivers of Turnover Prediction

Understanding which features contribute most to the model's predictions is essential not only for interpreting results, but also for guiding strategic HR decision-making. Table 8 presents the ten most important variables in the one-year turnover prediction model, based on their scaled importance (relative to the most impactful variable) and percentage contribution to the model's overall predictive power.

Table 8 - Top 10 Most Important Variables (One-year Model)

Variable	Scaled Importance (%)	Percentage (%)
Theoretical Fixed Compensation	100.00	58.33
Seniority	25.46	14.85
Age	9.52	5.55
Unemployment Rate	3.71	2.16
Training Hours	3.15	1.84
Average Wage Variation	3.11	1.81
Workload Type	2.65	1.54
District	2.50	1.46
Team Absence Hours	2.08	1.22
Absence Hours	1.86	1.09

The ten most important variables, shown in Table 8, are ranked according to their scaled and percentage importance within the XGBoost framework. Collectively, these ten variables account for 89.9% of the model's total predictive weight, indicating that the model's decisions are highly concentrated in a small group of features.

Among these, the top three variables — Theoretical Fixed Compensation, Seniority and Age — are particularly dominant, together representing 78.73% of the total importance. This highlights the model's strong reliance on these factors when estimating turnover probability. The Theoretical Fixed Compensation alone contributes 58.33% of the total weight, clearly standing out as the most influential feature. For context, in relative terms (scaled importance), Seniority holds 25.46% of the importance of Compensation, while Age represents 9.52%. These differences illustrate how central compensation is to the model's predictions, while seniority and age, though still relevant, play more complementary roles.

In contrast, the remaining seven variables each contribute less than 3% to the model's decisions, reinforcing the idea that the model focused most of its predictive power on a narrow set of features —

particularly on remuneration. This level of concentration is valuable for interpretation, as it provides clear insight into which dimensions are most strongly associated with turnover risk in the retail workforce context.

Beyond identifying the individual contribution of each feature, it is also possible to draw insights by grouping variables into thematic clusters. These clusters were previously defined based on the nature of the variables: Compensation and Performance, Employee Characteristics, Current Position, External Factors and Leadership Performance. From this perspective, Compensation and Performance emerges as the most dominant group, contributing with over 60% of the total model importance. This is largely due to the weight of the Theoretical Fixed Compensation, but also supported by variables such as Training Hours and Average Wage Variation. Employee Characteristics, represented mainly by Seniority and Age, jointly account for over 20% of the total importance, highlighting the role of employee tenure and lifecycle in turnover risk. Variables related to the Current Position, such as Workload Type, Team Absence Hours, Absence Hours and District, represent a smaller portion of the importance — collectively under 6%. Similarly, External Factors such as Unemployment Rate and Average Wage Variation have moderate relevance (around 4%), suggesting that broader labor market conditions, while not negligible, are secondary to internal characteristics. This breakdown reinforces the idea that compensation and individual employee characteristics are the primary drivers in predicting turnover in the short term, whereas contextual or organizational variables play a supporting role.

5.2 Five-year model

5.2.1 Assessment of Predictive Performance

Since the five-year model predicts turnover as a continuous variable — based on average turnover rates within each district × store type × role cluster × persona — traditional classification metrics such as AUC-ROC are not applicable, as there is no binary ground truth to rank or classify. Instead, model performance is evaluated using regression-oriented metrics: Mean Absolute Error (MAE), BIAS, and Brier Score. These metrics provide complementary insights into the model's behavior, capturing aspects such as average deviation, directional bias, and calibration. The results are presented in Table 9.

Table 9 - Overall Performance Metrics for the Five-Year Turnover Prediction Model

BIAS	MAE	Brier Score
- 1.2 pp	15.97 pp	0.062

The MAE of 15.97 pp indicates that, on average, the predicted turnover rates deviate from the actual values by approximately 16 percentage points. While this error is higher than the 10.36 pp observed in the one-year model, it is expected given the increased uncertainty associated with long-term forecasts and the lack of individual-level data for future employees.

The Brier Score of 0.062 reflects the average squared difference between the predicted probabilities and actual outcomes. Despite the model operating at an aggregated level and over a longer timeframe, this relatively low score suggests that the model maintains a reasonable degree of calibration — meaning that, in most segments, the predicted turnover rates closely reflect actual group behavior.

The BIAS of -1.2 pp reveals a slight underestimation trend, with the model predicting, on average, lower turnover than what was observed. This contrasts with the one-year model, which showed a positive bias (overestimation). This change in direction may be explained by structural differences between the two models, particularly the use of aggregated group-level inputs and the increased complexity of long-term prediction. Additionally, such directional errors could stem from uncertainty propagation, unobserved cohort effects or variation in how *personas* evolve over time.

Together, these results confirm that the five-year model is capable of producing meaningful long-term forecasts, despite working under more constrained conditions. However, due to its higher error and underestimation tendency, it is important that its outputs are interpreted with caution and supplemented with scenario testing — particularly when being used to inform strategic decisions over multi-year horizons.

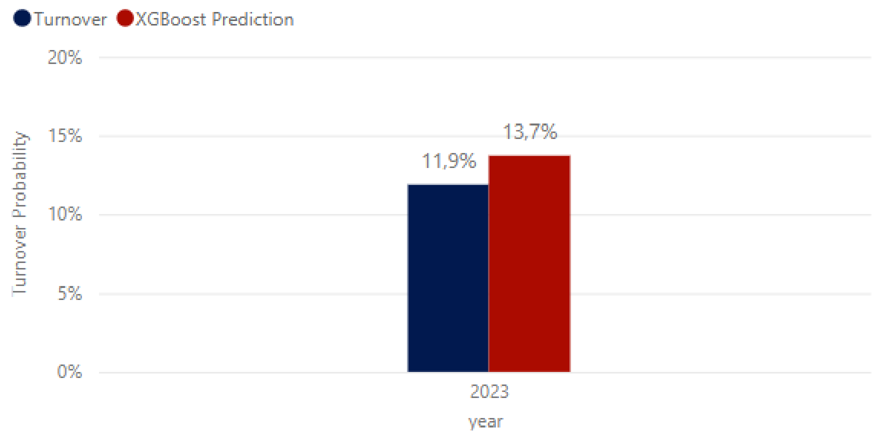

Figure 25 - Actual Turnover vs Turnover Prediction (Five-year model)

Figure 25 displays the comparison between the model's predicted turnover and the actual turnover observed in 2023, based on aggregated values across all segments in the test dataset. The five-year model predicted an average turnover rate of 13.7%, while the actual turnover was 11.9%, resulting in a 1.8 pp overestimation at the global level.

Despite the higher predicted turnover rate (13.7%) compared to the actual turnover observed in 2023 (11.9%), the model presents a negative BIAS of –1.2 pp. While this may initially appear contradictory, it is important to note that BIAS reflects the average direction of the prediction errors across all segments (i.e., district × store type × role cluster × *persona*), rather than the overall difference between aggregated values. In this case, the negative BIAS indicates that the model tends to underestimate turnover in most segments, even though a smaller number of segments with larger overestimations contribute to an overall predicted turnover that exceeds the actual figure. This nuance highlights the need to interpret BIAS and aggregate results together, as they capture different aspects of model performance.

While the global metrics presented earlier offer a first assessment of the model's performance, they are calculated over a fine-grained segmentation that includes all combinations of the categorical variables — not the *personas* defined by the company. In fact, these combinations include all possible groupings across district, store type, role cluster and all *persona*-defining variables individually, rather than the specific *persona* groupings used in workforce planning. For this reason, it is essential to explicitly assess model performance at two additional levels: first, at the district × store type × role cluster level, to analyze a less granular but operationally relevant segment; and second, at the *persona* level, which is the exact aggregation used to feed the optimization model. This distinction is critical, as the metrics at the *persona* level — where attributes like age or performance are grouped into predefined bands — can differ substantially from those calculated on fully disaggregated data. The following results focus on evaluating the model's behavior within these two key levels of aggregation.

The performance results in Table 10 reveal that model accuracy improves significantly when predictions are aggregated at the role cluster level, with Mean Absolute Error (MAE) dropping from 15.97 pp at the global level to just 6.63 pp. This substantial reduction highlights the value of aggregation in smoothing individual-level variation and producing more stable estimates. The Brier Score also decreases sharply to 0.008, reflecting improved calibration of the predicted probabilities at this less granular level.

Moreover, the BIAS shifts from –1.2 pp in the overall model to +0.94 pp when predictions are aggregated at the district × store type × role cluster level, suggesting that while the model tends to underestimate turnover globally, it slightly overestimates turnover at this operationally relevant aggregation. This change in direction reflects how aggregation can balance out localized underestimations and provide more neutral predictions, especially at levels used for high-level workforce planning.

At the *persona* level — the actual granularity used to feed the optimization model — the metrics are less favorable. The MAE increases to 14.09 pp and the Brier Score rises to 0.042, indicating that the model struggles more to maintain accuracy and calibration when targeting specific *persona* groups.

Additionally, the BIAS grows to 4.18 pp, pointing to a more pronounced and consistent overestimation across *personas*.

This degradation in performance is expected, not because the model combines multiple variables — as seen in the one-year model, which handles even more detailed data — but rather due to the way *personas* are constructed. These *personas* are predefined by the client using specific thresholds for numerical variables (e.g., grouping age above or below 35), selected categorical combinations and a fixed set of attributes. While this approach is operationally meaningful and essential for workforce planning, it introduces artificial segmentation that may not align with the natural patterns present in the data. As a result, some predictive accuracy is lost at this level.

Table 10 - Turnover Prediction Model Performance at Aggregated Segment Levels (Five-year model)

Level	BIAS	MAE	Brier Score
District x Store Type x Role Cluster	0.94 pp	6.63 pp	0.008
District x Store Type x Role Cluster x *Persona*	4.18 pp	14.09 pp	0.042

To gain a more comprehensive understanding of the model's behavior, it is therefore important to analyze prediction accuracy across all variables that define the forecast granularity. This includes not only the operational segmentation variables — district, store type and role cluster — but also the characteristics used to construct the *personas*. These *persona*-defining variables (e.g., age, tenure, contract type, workload, academic qualifications, performance and potential) were selected by the client and grouped into predefined categories. As such, analyzing model performance along each of these dimensions provides valuable insights into which specific segments are contributing to error and whether certain combinations require further refinement or targeted scenario testing.

Figure 26 compares the predicted and actual turnover rates for the top 10 districts in 2023, which collectively represent over 87.5% of the workforce. In these districts, the prediction errors range from 1 to 5 pp. Despite being calculated at the district level — a higher level of aggregation than the model's operational unit — the deviations remain relatively modest. The average absolute error across these 10 districts is approximately 2.4 pp, which is substantially lower than the global MAE of 15.97 pp (Table 9) and even below the 6.63 pp MAE observed at the district × store type × role cluster level (Table 10).

While districts are not the direct prediction unit of the model, the results shown here reflect an aggregation of predictions made at the more granular level (district × store type × role cluster × *persona*). Therefore, comparing this average error to the MAE of the district × store type × role cluster level is appropriate and reinforces the model's strong accuracy when results are aggregated geographically — particularly in areas with higher employee concentration.

Despite a consistent tendency to overestimate turnover — echoing the positive gap observed at the global level — only two districts, Porto and Ilha da Madeira, exhibit deviations of 5 pp. These outliers may indicate regional specificities not fully captured by the model and highlight the value of complementing predictive outputs with contextual insights. Still, the overall low error across districts supports the model's robustness for geographically targeted forecasting and strategic workforce planning.

Interestingly, while the five-year model shows moderate deviations in key districts like Porto (5 pp) and Lisboa (2 pp), the one-year model achieves higher accuracy in these same regions — with errors of only 1 pp in Lisboa and 3 pp in Porto (Figure 15). Together, these two districts account for approximately 43.5% of the entire dataset, making their accurate prediction particularly relevant. This contrast highlights how shorter-term models may be more precise for high-density areas in the near future, whereas the five-year model, despite its broader uncertainty, provides valuable strategic insight for long-term planning.

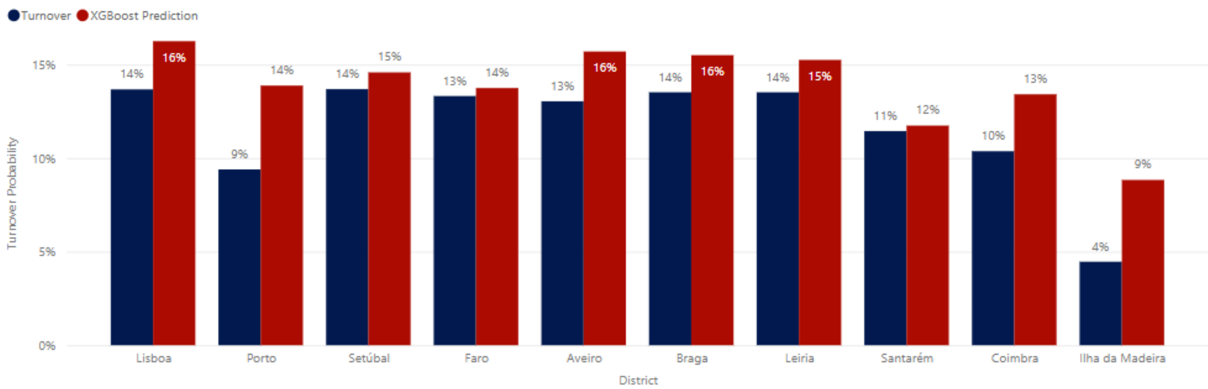

Figure 26 - Actual Turnover vs Turnover Prediction by Top 10 District (Five-year model)

The model's performance by store type — convenience, supermarket and hypermarket — reveals a high level of accuracy, with prediction errors ranging between 1 and 2 percentage points and an average absolute deviation of 1.7 pp. Although store type represents a higher level of aggregation than the model's operational unit, this result is still informative, as turnover planning often considers differences across retail formats.

Notably, the average error across store types is substantially lower than the global MAE of 15.97 pp (Table 9) and also lower than the MAE of 6.63 pp observed at the district × store type × role cluster level (Table 10). However, this difference should be interpreted with caution, as higher levels of aggregation naturally reduce variability and, consequently, error metrics. Still, the ability of the model to approximate actual turnover even at this aggregated level reinforces its consistency and adaptability across organizational layers.

As illustrated in Figure 27, the model slightly overestimates turnover in all three store formats, with deviations of 1 pp in supermarkets and convenience stores and 2 pp in hypermarkets. The small magnitude of these differences supports the model's robustness for store-level forecasting, reinforcing its value for strategic workforce planning across different retail formats. Notably, the five-year model (Figure 16) achieves very similar accuracy levels to the one-year model for supermarkets and hypermarkets — the two dominant store formats — despite the added complexity of long-term forecasting. This reinforces confidence in its applicability across key operational segments.

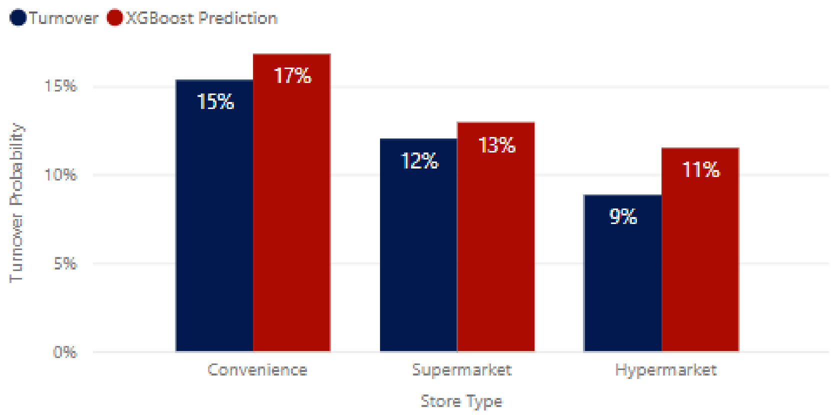

Figure 27 - Actual Turnover vs Turnover Prediction by Store Type (Five-year model)

Figure 28 compares the predicted and actual turnover rates for the top 10 role clusters in 2023, which together account for the majority of the workforce. Across these clusters, prediction errors range from 1 to 6 percentage points, with an average absolute error of approximately 2.4 pp. This is substantially lower than the 6.63 pp MAE observed at the same level of aggregation — district × store type × role cluster — as reported in Table 10. This suggests that, when analyzed by role cluster, the model demonstrates strong predictive accuracy, particularly in the most representative functions within the organization.

The model consistently overestimates turnover across all clusters, in line with the global prediction pattern. However, one function — Bakery/Pastry Counter Operator — stands out with a deviation of 6 pp, indicating a potential area for refinement. Still, the overall low magnitude of errors reinforces the model's suitability for workforce planning when segmenting by role cluster.

These findings are consistent with the one-year model (Figure 17), which also demonstrated strong alignment across core role clusters, with similarly small and stable deviations. This temporal consistency further supports the model's robustness when segmenting turnover by function.

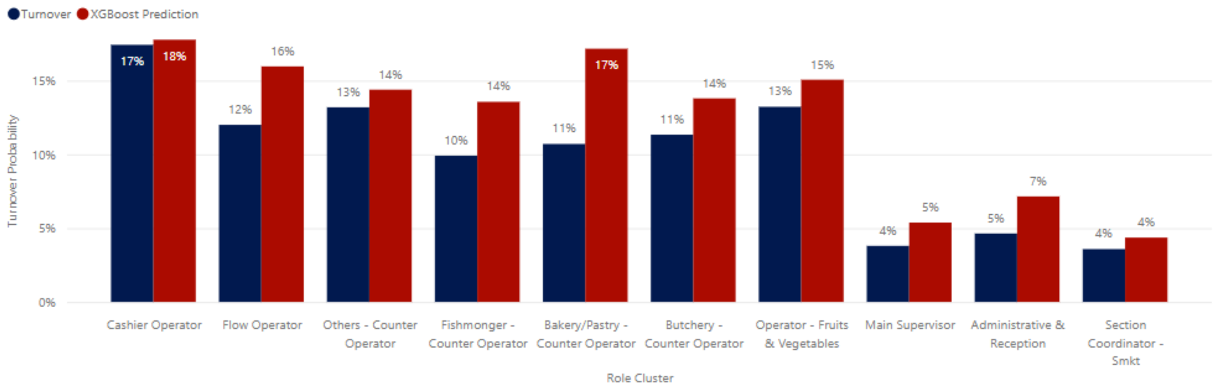

Figure 28 - Actual Turnover vs Turnover Prediction by Role Cluster (Five-year model)

When analyzing turnover by academic qualifications, the model successfully captures the general trend across educational levels. As shown in Figure 29, it slightly underestimates turnover for employees with higher education (by 2 pp), perfectly aligns predictions for secondary education and slightly overestimates for those with primary education (by 3 pp). Despite these modest deviations, the model reflects the expected behavioral pattern — employees with higher qualifications tend to have greater turnover. The relatively small errors reinforce the model's ability to internalize meaningful relationships between educational background and turnover, supporting its applicability across diverse employee profiles.

These results are consistent with those of the one-year model (Figure 18), which also accurately captured the turnover gradient across education levels with similarly low error margins. This reinforces the model's stability in reflecting education-related turnover trends over different forecasting horizons.

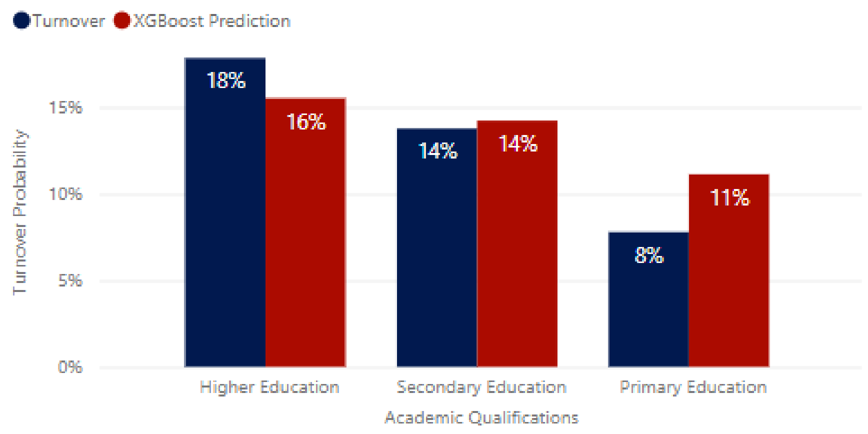

Figure 29 - Actual Turnover vs Turnover Prediction by Academic Qualifications (Five-year model)

The model clearly distinguishes between the two contract types, capturing the structural difference in turnover behavior between fixed-term and permanent employees. As shown in Figure 30, fixed-term contracts exhibit significantly higher actual turnover (24%) compared to permanent contracts (8%) — a trend that the model successfully replicates. While the model slightly underestimates turnover for fixed-term employees (22% predicted), it overestimates it for permanent employees (10% predicted), resulting in a 2 pp directional gap in each group. These deviations are modest and reinforce the model's ability to internalize broad behavioral distinctions across contract types. Given the strong correlation between contract relationship and employee turnover, these results validate the inclusion of this variable as a key driver in *persona* segmentation and long-term forecasting. Notably, the deviations observed are more balanced than those of the one-year model (Figure 19), which exhibited a larger overestimation for fixed-term contracts. This reinforces the five-year model's ability to generalize across segments with varying representation, despite working with more aggregated inputs.

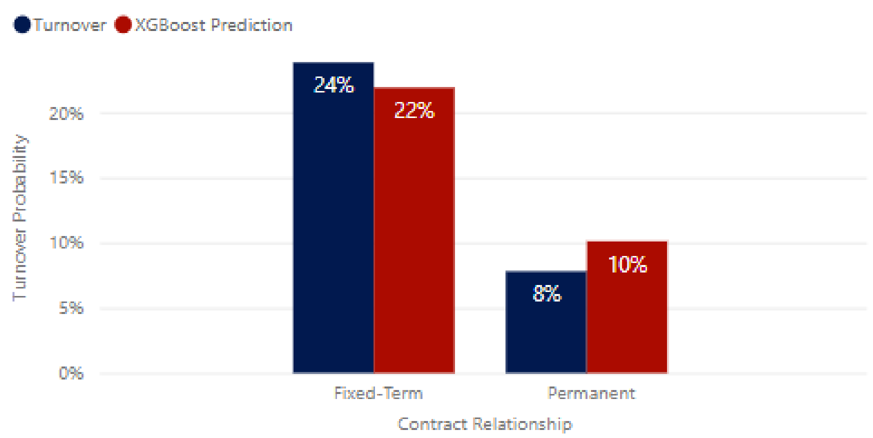

Figure 30 - Actual Turnover vs Turnover Prediction by Contract Relationship (Five-year model)

Figure 31 illustrates the model's performance across different workload types — part-time and full-time contracts. The model effectively captures the substantial gap in turnover behavior between the two groups, correctly predicting a significantly higher turnover rate among part-time employees. For this group, it slightly underestimates turnover by 1 pp (22% predicted vs. 23% actual), while for full-time employees, it overestimates the rate by 3 pp (10% vs. 7%). Overall, the model aligns closely with the one-year model in this segmentation (Figure 20), confirming its ability to differentiate turnover levels by workload type and provide reliable estimates for planning purposes.

These results are aligned with those of the one-year model (Figure 20), which also accurately captured the behavioral gap between part-time and full-time employees, albeit with slightly higher deviations for the part-time group. This consistency across time horizons strengthens the confidence in the model's ability to represent contract-related turnover dynamics.

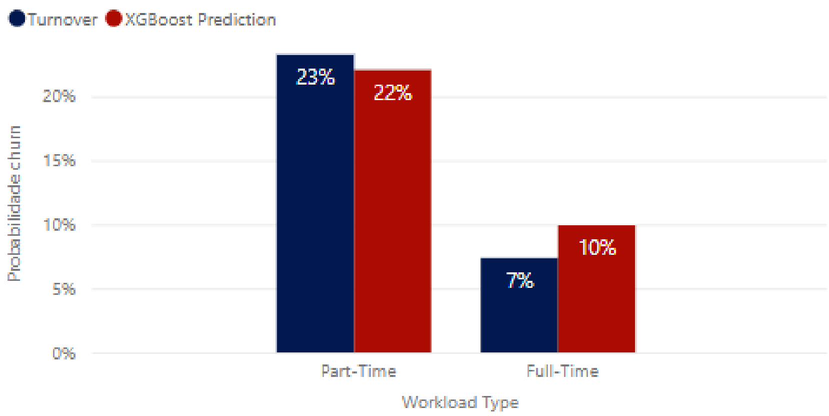

Figure 31 - Actual Turnover vs Turnover Prediction by Workload Type (Five-year model)

Figure 32 compares the predicted and actual turnover rates across age groups. The model accurately captures the general downward trend in turnover as age increases, with prediction errors ranging between 1 and 2 percentage points across all segments. This consistent alignment across age brackets suggests that the model effectively differentiates employee behavior by age, reinforcing its reliability for workforce planning across generational profiles. Notably, these results closely mirror those obtained with the one-year model (Figure 21), confirming the model's ability to replicate known behavioral patterns even in a long-term forecasting context.

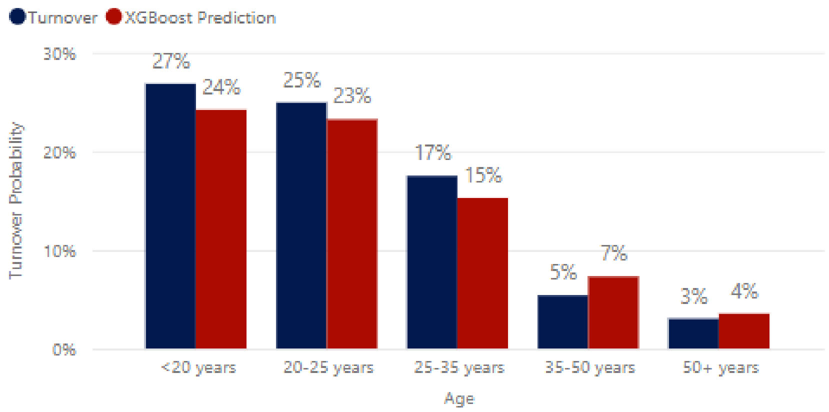

Figure 32 - Actual Turnover vs Turnover Prediction by Age (Five-year model)

Figure 33 illustrates the comparison between predicted and actual turnover rates across different seniority brackets. The model accurately captures the general declining trend in turnover as seniority increases, aligning well with patterns typically observed in the retail sector. However, it exhibits notable deviations in the 1–3 year and 3–6 year brackets. In these mid-tenure groups, the model underestimates turnover by 14 pp and 13 pp respectively — the largest gaps across all categories. For the remaining groups, particularly those with more than 6 years of tenure, prediction errors are much smaller, ranging between 1 and 2 pp. These findings suggest that while the model performs reliably for long-tenured employees, additional refinement may be required to better capture the turnover behavior of employees in their early-to-mid career stages.

Compared to the one-year model (Figure 22), which also exhibited a notable prediction error of 8 pp in the 1–3 year bracket, the five-year model further amplifies this deviation, while additionally underestimating turnover in the 3–6 year group by 13 pp. This suggests that turnover behavior among mid-tenure employees is particularly challenging to capture, with forecasting accuracy declining further over longer time horizons. These findings reinforce the need for targeted refinements when modeling employee behavior in early-to-mid career stages.

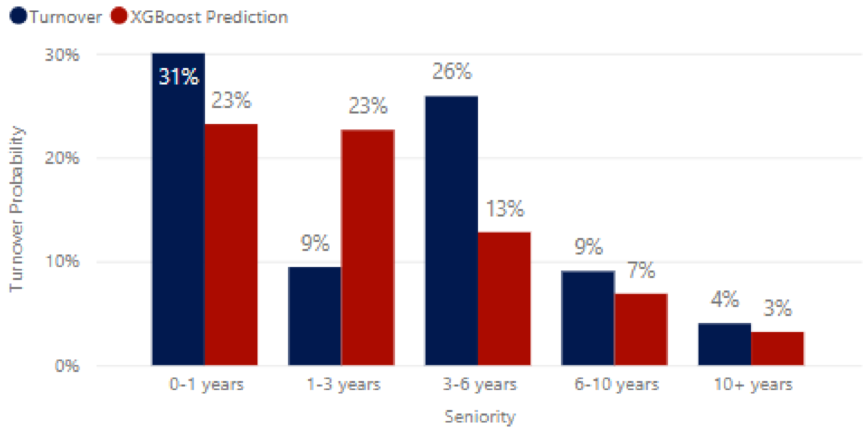

Figure 33 - Actual Turnover vs Turnover Prediction by Seniority (Five-year model)

As shown in Figure 34, the model demonstrates mixed performance when predicting turnover across different performance levels. For ratings 1 and 2, it slightly underestimates the actual turnover, with deviations between 2 and 4 percentage points. In contrast, for higher performance scores (3 and 4), the model tends to overestimate turnover — particularly in score 4, where the predicted rate doubles the observed value (10% vs. 5%). These discrepancies suggest that, although the model performs reasonably well for lower performance categories, it struggles to accurately capture turnover behavior for higher-performing employees.

It is also important to explain the absence of level 5 in this analysis. Since the five-year model is based on aggregated data at the district × store type × role cluster × *persona* level, the performance variable — like other numerical predictors — is averaged within each segment. Given that employees with a performance score of 5 represented only about 3% of the dataset (as observed in Figure 12), these rare values were diluted when combined with more frequent lower scores during the aggregation process. As a result, no *persona*-level segment retained an average performance score of 5, explaining its absence from the graph.

Compared to the one-year model (Figure 23), which managed to predict turnover more consistently across all five performance levels with smaller deviations, the five-year model shows greater difficulty in accurately capturing the behavior of higher-performing employees — particularly those rated 4, for whom the predicted turnover rate doubles the observed value. This contrast underscores how the aggregation required for the five-year horizon, while enabling long-term forecasting, may reduce sensitivity to less frequent profiles and highlight segments that could benefit from further refinement.

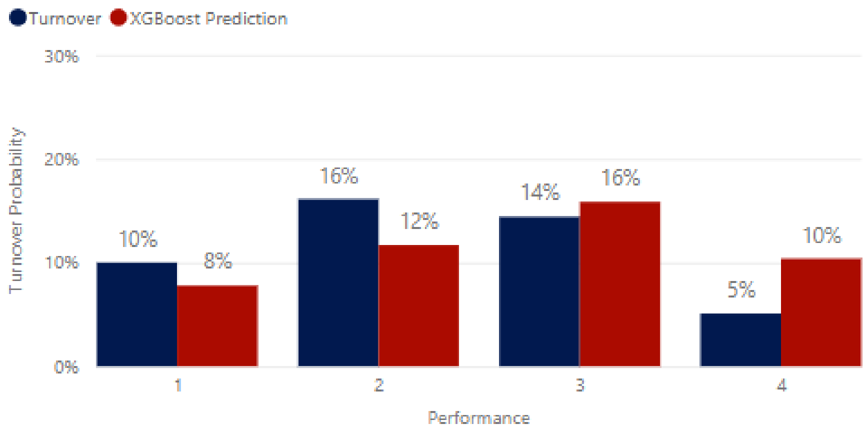

Figure 34 - Actual Turnover vs Turnover Prediction by Performance (Five-year model)

The model's predictions across different potential levels (Figure 35) show a consistent overestimation of turnover, although the magnitude of the errors remains within an acceptable range. For employees rated with the lowest potential (score 1), the model overpredicts turnover by 2 pp (15% vs. 13%). For potential levels 2 and 3, the differences are slightly larger — 4 pp and 2 pp, respectively — with the predicted values exceeding the actual ones in all cases. While the model appears to capture the general ordering of turnover rates by potential, the consistent upward bias suggests that this variable may be more challenging to model accurately under the aggregated *persona* structure. Still, the relatively small deviations highlight a reasonable capacity to reflect turnover tendencies across potential levels.

These results are consistent with the one-year model (Figure 24), which also captured the inverse relationship between potential and turnover, while maintaining similarly modest error margins. This consistency across time horizons reinforces the model's reliability in segmenting turnover risk based on employee potential.

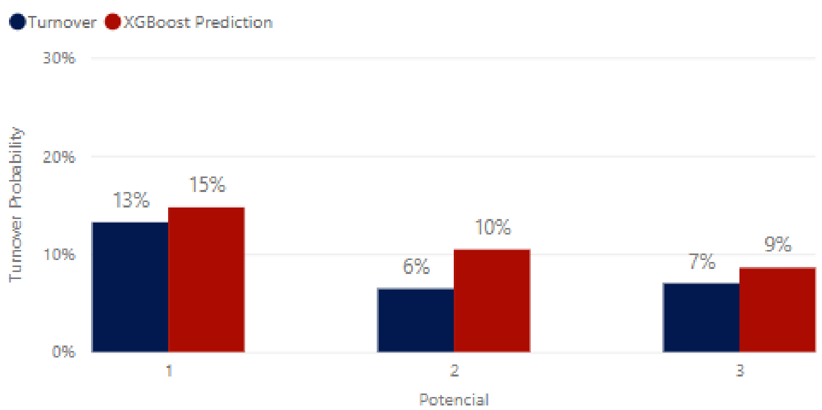

Figure 35 - Actual Turnover vs Turnover Prediction by Potential (Five-year model)

Having evaluated how the model performs across employee potential levels, the next section shifts focus to understanding which features most influence these predictions.

5.2.2 Drivers of Turnover Prediction

In the same way that feature importance was analyzed for the one-year model, it is equally essential to conduct this analysis for the five-year model. This helps unveil the internal mechanisms driving turnover predictions, enhancing transparency and mitigating the risks of relying on a black-box approach.

Table 11 - Top 10 Most Important Variables (Five-year Model)

Variable	Scaled Importance (%)	Percentage (%)
Theoretical Fixed Compensation	100.00	32.38
Seniority	61.84	20.02
Minimum Wage	28.21	9.13
Unemployment Rate	22.91	7.42
Age	22.59	7.31
District	22.39	7.25
Role Cluster	21.77	7.05
Academic Qualifications	2.98	2.37
Store Type	6.20	2.01
Performance	6.11	1.98

The ten most important variables presented in Table 11 are ranked according to their scaled and percentage importance within the five-year XGBoost model. Together, they account for 96.92% of the model's total predictive weight — a strong indication that the model's decisions are concentrated in a defined group of features.

In the one-year model, just three variables were responsible for nearly 80% of the total importance. In the five-year model, a similar concentration is reached — though more gradually — with six variables already accounting for 83% of the model's predictive power. This indicates that, while the long-term model still identifies a core set of influential features, it spreads predictive relevance across a slightly broader base of inputs, reflecting the greater uncertainty and generalization needs in long-term forecasting.

To deepen the interpretation of the model’s internal logic, the variables can be grouped into thematic categories — the same used in the one-year model analysis: Compensation and Performance, Employee Characteristics, Organizational Segmentation and External Factors.

Within this framework, Compensation and Performance remains the most influential group, primarily due to the prominence of Theoretical Fixed Compensation, which alone contributes 32.38% of the model's total importance. Variables such as Performance also reinforce the relevance of this category, confirming the strong link between remuneration and turnover likelihood — even in long-term projections.

Employee Characteristics — particularly Seniority and Age — also represent a significant share of the total importance. Seniority alone accounts for 20.02%, while Age contributes 7.31%, confirming that structural attributes of the employee profile continue to play a critical role in predicting turnover across extended horizons. This mirrors the findings from the one-year model and reinforces the idea that employee lifecycle factors remain key drivers of exit probability over time.

Interestingly, External Factors — such as Minimum Wage and Unemployment Rate — gain more prominence in the five-year model compared to the one-year counterpart. Together, they contribute 16.55% of the total importance, whereas in the one-year model, external variables represented only 3.97%. This shift highlights the model’s increasing reliance on macroeconomic indicators to explain long-term turnover trends and supports its application in scenario planning, where user-defined projections (e.g., economic policies or labor market dynamics) directly influence forecast outcomes.

Organizational Segmentation — particularly District and Role Cluster — also contribute meaningfully to the model’s predictive capacity. District accounts for 7.25% and Role Cluster for 7.05% of the total importance, indicating that the geographical and functional positioning of employees plays a non-negligible role in shaping turnover trends. Store Type, although less prominent with 2.01%, still adds relevant insight. These findings suggest that organizational structure remains a relevant factor in long-term turnover forecasting, complementing individual and economic dimensions.

To better reflect their collective relevance, we aggregated the importance scores of all individual categories under each parent variable. After this consolidation, variables such as District (7.25%) and Role Cluster (7.05%) emerged among the top contributors — proving that, although their individual categories may hold limited discriminative power, the overall role of these segmentation variables in the model is not negligible.

Altogether, the model’s predictive strength is concentrated in a relatively small subset of variables: just six variables explain over 83% of the total importance, revealing a more balanced — yet still focused — distribution than observed in the one-year model, where only three variables accounted for a similar share. This broader reliance on multiple moderately influential variables enhances the model's robustness and generalization capacity under the greater uncertainty of long-term forecasting, providing a solid foundation for scenario-based workforce planning.

5.3 Final Remarks

This chapter presented a detailed analysis of two predictive models developed to estimate voluntary employee turnover: one focused on a one-year horizon and the other on a five-year horizon. The one-year model, built using individual-level data, demonstrated good predictive performance with an AUC-ROC of 93% and a Brier Score of 0.046, confirming its ability to differentiate between employees at risk of turnover and those likely to stay. However, the model's average prediction error (MAE = 10.36 pp) represents a significant margin when compared to the actual turnover rate of 11.9%. Despite this, the model performed considerably better when applied to aggregated data at the district × store type × role cluster level, where the MAE decreased to 4.42 pp and the Brier Score to 0.004. Similarly, the inclusion of *personas* led to a modest improvement, reducing the MAE to 9.17 pp and the Brier Score to 0.021.

The five-year model, developed for long-term workforce planning, exhibited greater prediction uncertainty due to the broader forecast horizon and use of aggregated data. It achieved a global MAE of 15.97 pp and a Brier Score of 0.062, with a slight underestimation bias (−1.2 pp). Nonetheless, its calibration remained solid and useful for scenario simulation. At the district × store type × role cluster level, the MAE improved to 6.63 pp and the Brier Score to 0.008. When *personas* were included, the MAE decreased to 14.09 pp and the Brier Score to 0.042.

Both models provide valuable insights for workforce planning, though each serves a different purpose and aligns with distinct decision-making horizons. The one-year model is particularly effective for short-term tactical decisions — such as retention strategies and recruitment planning — by focusing on a small set of key drivers like compensation, seniority and age. In contrast, the five-year model draws on a broader and more diversified set of variables, including macroeconomic indicators such as minimum wage and unemployment rates, making it better suited for long-term planning and scenario-based simulations.

The models also highlight the importance of several key variables, particularly compensation, seniority and age, in predicting turnover. These variables consistently appear as top drivers in both models, confirming their central role in turnover behavior. While both models demonstrated good calibration, with Brier Scores well below the baseline of a random classifier, their strengths lie primarily in aggregate-level forecasting, rather than in highly granular individual predictions.

Overall, the results demonstrate the applicability of predictive analytics for workforce planning, offering a solid foundation for both short-term tactical decisions and long-term strategic planning. Beyond statistical accuracy, these models enable HR teams to proactively anticipate turnover dynamics and better align workforce supply with organizational needs.

The next chapter builds upon these findings through a critical reflection, assessing the methodological, structural and strategic implications of the models and identifying opportunities for further development.

6. Conclusion

This chapter concludes the study by revisiting the project's overarching objectives and summarizing its key contributions. The central aim was to develop predictive models capable of estimating voluntary employee turnover in the retail sector over both short- and long-term horizons, spanning up to five years. By applying people analytics techniques and leveraging decision tree-based algorithms — particularly XGBoost — the study demonstrated that turnover can be effectively forecasted using a combination of intrinsic and extrinsic employee characteristics. Beyond predictive accuracy, the project also highlighted the value of integrating these predictions into a broader workforce planning framework, reinforcing the strategic potential of data-driven HR practices.

The remainder of this chapter is structured in two parts: Section 6.1 presents a critical assessment of the model outcomes, linking them to strategic HR decision-making needs, while Section 6.2 proposes concrete directions for future development.

6.1 Critical Analysis Strategic Assessment and Critical Reflection

Before presenting a critical assessment of the models developed in this dissertation, it is essential to revisit the broader context, the strategic motivation of the work and the research questions that guided this investigation. The project set out to explore how predictive modelling and People Analytics could be applied to address a persistent challenge in the retail sector: high voluntary turnover. The objective was twofold — to forecast employee exits over one-year and five-year horizons and to use these insights as inputs for workforce planning at both tactical and strategic levels. In doing so, the project contributed not only by developing predictive models, but also by demonstrating how such tools can guide HR decision-making aligned with different planning horizons.

This section builds on those contributions by examining the models' performance, identifying key limitations and assessing their broader utility for workforce planning. The analysis considers both methodological aspects — such as variable structure and model calibration — and strategic implications for organizational decision-making. This critical analysis reflects both on the achievement of the initial objectives and on the answers provided to RQ1 – "What are the main individual and contextual factors that influence voluntary employee turnover in the retail sector?" – and RQ2 – "How can predictive analytics be used to forecast voluntary turnover and support strategic workforce planning in high-turnover environments?".

The results of both predictive models provide relevant insights into their applicability, while also revealing limitations that warrant careful consideration. The one-year model demonstrates high discriminatory power and strong calibration, although its absolute error at the individual level remains non-negligible. This limitation, however, is substantially mitigated when applied to aggregated forecasts — especially when segmented by District, Store Type and Role Cluster — aligning better with the granularity required for workforce planning decisions. The relatively modest performance gains observed from introducing *personas* raise questions about the behavioral coherence of these groupings. If not reflective of distinct turnover patterns, such segmentation may introduce noise rather than improve predictive accuracy.

The five-year model, while inherently less precise due to its extended forecast horizon and the absence of individual-level data, still delivers calibrated predictions at the group level. However, the slight performance gain from *personas* — along with an inversion in prediction bias — suggests that the current *persona* structure may not be well suited for long-term forecasting.

Alongside these insights, two structural limitations may be contributing to reduced predictive performance. First, the strong collinearity between age and seniority could hinder the model's ability to isolate early-career turnover patterns. Second, variables such as Performance and Potential exhibit very low predictive importance, likely due to their skewed distributions. Although included in the *persona* framework due to their strategic relevance, these attributes added limited discriminative power and may have weakened the segmentation logic. These aspects were not explicitly tested in this study but represent avenues for future refinement.

Despite these challenges, both models achieved good probability calibration, confirming their reliability for use in scenario analysis and workforce optimization. While the Mean Absolute Error remains high at the individual level, prediction errors tend to cancel out in aggregate — as reflected by consistently lower BIAS — making the models well suited for strategic planning at higher levels of aggregation.

From a strategic standpoint, the models' distinct roles reflect their complementary value. The one-year model supports short-term, actionable planning, grounded in interpretable variables like compensation and tenure. In contrast, the five-year model captures broader trends by incorporating external indicators such as minimum wage and unemployment, supporting long-term scenario simulation. Together, they enable a multi-horizon planning approach that aligns immediate actions with longer-term strategic foresight.

Although not yet confirmed as actively deployed in a real-time decision-making environment, the models have already been integrated into a workforce optimization model within the organization. This ensures they are readily available for practical use and have the capacity to support workforce planning through scenario-based assessments.

Building on these insights, it is now possible to directly address the research questions that guided this dissertation. In response to RQ1, the analysis confirmed the central role of compensation, seniority and broader macroeconomic factors in influencing voluntary turnover. For RQ2, the development and evaluation of two machine learning models demonstrated that predictive analytics can effectively anticipate turnover trends and inform workforce planning strategies across different timeframes.

Beyond this analytical perspective, three key contributions emerge from the work. First, it introduces a predictive framework tailored to the retail sector's specific challenges. Second, it demonstrates how predictive outputs can be incorporated into workforce planning tools to support both operational and strategic decisions. Third, it provides a reflective assessment of methodological and structural limitations, offering clear directions for improvement and future development.

Ultimately, this work advances the shift from reactive to proactive HR management. By embedding predictive analytics into workforce planning, organizations can better anticipate future workforce needs and align hiring, mobility and retention strategies accordingly. This positions HR teams to make more informed decisions, while strengthening organizational resilience in dynamic environments like retail.

The next section outlines a set of short- and long-term recommendations to enhance the models and extend their impact.

6.2 Future Research and Development

Given the dynamic nature of workforce behaviours and organizational needs, the continuous refinement of predictive models is essential. Rather than proposing a single list of future improvements, this section organizes recommendations into two distinct time horizons — short-term and medium/long-term — to better reflect their complexity, feasibility and expected impact.

Short-term developments focus on optimizing the current framework, enhancing input structures and improving model interpretability and alignment with decision-making needs. These are actions that can be pursued using available data and tools, with relatively low implementation risk.

In contrast, medium to long-term developments aim to extend the capabilities of the model by integrating advanced forecasting methods, exploring alternative modeling techniques and enabling richer scenario simulations. These directions involve a higher level of uncertainty or require structural changes in data collection but offer the potential for significant strategic value over time.

6.2.1 Short-Term Improvements

Personas Revision

A key opportunity for improvement lies in how *personas* are defined. In the current model, *personas* are pre-structured based on a set of manually selected variables such as contract type, workload type, academic qualifications, performance, potential, age and tenure. While these variables were chosen for their strategic relevance and practical availability, the way in which they are segmented — particularly the numerical ones — may not effectively capture real behavioral differences in turnover.

For instance, arbitrary discretization of continuous variables like age or tenure (e.g., creating bins such as "under 25," "25–35," "35–45") may group together employees who do not share similar patterns of turnover, while separating those who do. This may result in poorly defined *personas* that dilute the predictive power of the model.

To address this, it is recommended to explore unsupervised clustering techniques (e.g., K-means, DBSCAN, hierarchical clustering), which can consider both categorical and numerical variables in their raw form. These approaches can automatically identify naturally occurring patterns in the data — including optimal groupings for continuous variables — resulting in *personas* that are more cohesive in terms of actual turnover behavior. By allowing the data to guide both the selection of variables and the way they are grouped, the resulting *personas* may better reflect the true heterogeneity of the workforce, increasing the model's ability to generalize and enhancing the effectiveness of the optimization framework.

Improvement of Performance and Potential Evaluations

The current use of *Performance* and *Potential* variables is significantly limited by the way these attributes are distributed in the dataset. As shown in the figures, both variables are heavily skewed:

- Performance: More than 64% of employees receive a score of 3, followed by 26% at level 4. The remaining levels (1, 2 and 5) are heavily underrepresented (Figure 12).
- Potential: Approximately 80% of employees are rated as level 1, with only 16% in level 2 and less than 5% in level 3 (Figure 13).

This extreme concentration around one or two values greatly reduces the discriminative power of these variables. For example, if nearly all employees are assigned a potential score of 1, the model cannot effectively differentiate turnover patterns across potential levels — limiting its ability to extract meaningful insights from this feature.

In the short term, normalization or rescaling techniques such as z-score transformation or quantile scaling are not suitable in this case, as the data are ordinal, categorical and highly imbalanced.

Instead, alternative preprocessing methods could be explored, such as smoothing or weighting mechanisms to slightly increase the influence of underrepresented classes without artificially distorting the dataset's structure.

However, the most impactful improvement lies in redefining how these variables are evaluated. Collaborating with HR to revise the current evaluation framework is recommended, with the goal of producing more informative and balanced distributions. Potential steps include:

- Introducing more granular rating systems (e.g., 1 to 10 scale);
- Standardizing calibration criteria to ensure consistent evaluations across teams and managers;
- Providing clear guidelines and benchmarks to reduce rating inflation or clustering around the median;
- Combining subjective assessments with objective metrics such as goal achievement, skills development or mobility history.

By improving both the scale and the consistency of these evaluations, these features could become significantly more valuable for turnover prediction. In addition to enhancing model performance, such improvements could foster greater fairness and transparency in employee assessment practices — bringing added value to both the predictive system and HR decision-making as a whole.

Analysis of Collinearity Between Age and Seniority

During the exploratory analysis of the dataset, a strong correlation was observed between the variables *age* and *seniority*. This relationship raised early concerns about multicollinearity and was later supported by the model's performance across different employee segments: while the model showed high accuracy in predicting turnover by *age*, it struggled to do so by *seniority* — particularly underestimating turnover in early-career groups (e.g., 1–3 years and 3–6 years of tenure). Given that these two variables are structurally related, such inconsistency suggests that the model may be failing to disentangle their individual effects.

This behavior may indicate that *age* is absorbing most of the explanatory power of *seniority*, leading to a redundancy that compromises interpretability and potentially limits predictive performance in segments where age and tenure do not align perfectly. To address this issue, a formal multicollinearity analysis should be conducted — starting with correlation matrices and variance inflation factor (VIF) calculations.

If high multicollinearity is confirmed, several strategies can be considered:

- Excluding one of the variables to simplify the model;

- Applying dimensionality reduction techniques (e.g., principal component analysis) to capture shared variance;
- Creating engineered features that express their interaction more effectively (e.g., age-to-tenure ratio or tenure adjusted for age group).

This analysis should not be restricted to age and seniority. Similar checks should be extended to other potentially correlated feature pairs. The insights from these analyses can guide further model refinement, helping to improve both stability and interpretability by ensuring that each variable adds unique, non-redundant value to the model's predictions.

Strengthening Local Interpretability with SHAP

While the global feature importance analysis provides valuable insights into the general behaviour of the model, local interpretability — understanding why the model made a specific prediction for a specific group — is equally important, especially for operational decision-making at the regional level.
To achieve this, it is recommended to implement SHAP (SHapley Additive exPlanations) at the group level — particularly for combinations such as district x store type x role cluster x *persona*. SHAP values can help explain how each feature contributes (positively or negatively) to the predicted turnover for a specific group, offering a more nuanced and transparent view than global feature importance alone.

For instance, in one region, high turnover might be driven by low wages, while in another it might be more affected by the prevalence of fixed-term contracts. By highlighting these local dynamics, SHAP explanations empower regional HR managers to design targeted retention strategies and trust the model's outputs, ultimately increasing the impact and adoption of data-driven decisions.

Future Adjustment for Systematic Bias

Despite the strong average performance, both the one-year and five-year models display systematic bias in certain segments — consistently overestimating or underestimating turnover. These patterns can be addressed by implementing a lightweight bias correction mechanism based on historical errors.

Such a mechanism could involve calculating the average error for each segment over time (e.g., by district x store type x role cluster x *persona*) and then applying a correction factor to future predictions for those groups. This should be done cautiously, as overly aggressive adjustments could lead to overfitting or instability. However, if applied conservatively and monitored regularly, it could improve alignment between predicted and actual turnover, especially for recurring edge cases.

Additionally, this feedback loop could be embedded into the model's lifecycle — triggering minor updates or alerts when bias exceeds predefined thresholds.

More Frequent Model Updates

Currently, the model is updated on an annual basis, following the yearly data refresh. However, labour dynamics and external factors can shift more rapidly, especially during periods of organizational change, economic volatility or policy reform.

It is therefore suggested to evaluate the feasibility of semi-annual updates, particularly for the one-year model, which is more sensitive to recent trends. These updates would not necessarily require retraining the entire model — in some cases, incremental learning approaches or fast recalibration of certain components could suffice.

By increasing the refresh rate, the models could respond more quickly to changing conditions and avoid the lag that comes with outdated training data. This is particularly relevant in fast-moving environments such as retail, where turnover patterns can shift significantly in a matter of months.

6.2.2 Medium/Long-Term Improvements

Integration of Internal Mobility Dynamics

At present, the turnover prediction model operates on an annual basis, meaning that it only considers individuals or *personas* who remain in the company until the end of the year to predict their turnover for the following year. Consequently, employees who enter and leave within the same year are excluded from the model, as they do not meet the criteria for the turnover status at the year-end. This limitation means that any intra-year dynamics — such as seasonal fluctuations in turnover, especially common in retail environments — are not being captured.

The absence of seasonal turnover, which can be significantly higher in certain periods (such as summer or holiday seasons), represents a gap in the model's predictive capability. While the current model focuses solely on predicting turnover based on the status at year-end, incorporating turnover during the year, including entries and exits within the same period, could offer a more comprehensive view of workforce dynamics. Such an improvement would make the model more reflective of real-world employee movement, particularly in industries like retail where seasonal turnover is a common trend.

Expansion of External Data Sources

While the five-year model already includes macroeconomic variables such as minimum wage and unemployment rate, its predictive potential could be enhanced by incorporating additional external data sources. These may include region-specific labour demand indicators, inflation-adjusted cost of living indexes, real estate prices or even commuting and transportation accessibility metrics.

By broadening the contextual scope, the model can better capture variations in turnover drivers across regions and time periods. For example, a district with limited public transportation and rising living costs might exhibit higher turnover even if internal compensation remains stable. Integrating these external datasets would also support more nuanced scenario analyses and enhance the realism of long-term workforce planning.

Ethical and Fairness Considerations

As predictive models increasingly support strategic HR decisions, it becomes important to ensure they do not unintentionally reinforce inequities. Although the current models do not include sensitive attributes such as gender or ethnicity, potential proxy bias may still exist through correlated variables like age or contract type. Future developments should include fairness evaluations to promote ethical and equitable use of predictive analytics in workforce planning.

Benchmarking with Alternative Architectures

While XGBoost has proven to be a robust and interpretable algorithm, the continuous evolution of predictive modelling in HR analytics opens the door to exploring alternative approaches. In the medium to long term, it may be worthwhile to benchmark the current model against both more advanced architectures and simpler ones, in order to assess whether different modelling strategies offer greater stability or adaptability in uncertain long-term forecasting scenarios. The priority should remain on balancing predictive performance with interpretability and robustness.

Incorporating Exit Interview and Qualitative Data

One of the limitations of the current model is its reliance on structured, quantitative variables. However, qualitative signals — such as exit interview notes, employee engagement surveys or written performance reviews — may contain valuable insights about the underlying reasons for turnover.

Natural Language Processing (NLP) techniques could be used to extract sentiment, thematic patterns or behavioural cues from these unstructured sources. These insights could then be incorporated as additional features or used to validate and contextualize model predictions. While the integration of such data requires careful ethical and privacy considerations, it offers a powerful avenue to complement numerical models with human-centric understanding.

Operational Feedback Loops and Continuous Learning

A long-term goal for the modelling system is to embed a real-time feedback mechanism into the HR decision-making process. Currently, predictions are made and interpreted externally from the outcomes they seek to influence. By closing this loop — for instance, by recording how well the model's predictions aligned with reality after workforce planning actions were implemented — the model could learn continuously from its environment.

This would allow for adaptive recalibration, faster detection of concept drift and a more proactive approach to workforce planning. While such systems require investment in infrastructure and governance, they represent the next step in making predictive analytics not just a forecasting tool, but a core component of strategic HR operations.

REFERENCES

Al-Suraihi, W. A., Samikon, S. A., Al-Suraihi, A.-H. A. & Ibrahim, I. (2021). Employee Turnover: Causes, Importance and Retention Strategies. *European Journal of Business and Management Research*, *6*(3), 1–10. https://doi.org/10.24018/EJBMR.2021.6.3.893

Bar-Gil, O., Ron, T. & Czerniak, O. (2024). AI for the people? Embedding AI ethics in HR and people analytics projects. *Technology in Society*, *77*. https://doi.org/10.1016/j.techsoc.2024.102527

Barrick, M. R. & Zimmerman, R. D. (2005). *Reducing Voluntary, Avoidable Turnover Through Selection*. https://doi.org/10.1037/0021-9010.90.1.159

Bass, B. M. (1990). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, *18*(3), 19–31. https://doi.org/10.1016/0090-2616(90)90061-S

Benson, G. S. (2006). Employee development, commitment and intention to turnover: a test of 'employability' policies in action. *Human Resource Management Journal*, *16*(2), 173–192. https://doi.org/10.1111/J.1748-8583.2006.00011.X

Cappelli, P. (2008). Talent management for the twenty-first century. *Harvard Business Review*, *86*, 74–81. https://www.researchgate.net/publication/5440660

Chen, G., Tang, Y. & Su, Y. (2021). The Effect of Perceived Over-Qualification on Turnover Intention From a Cognition Perspective. *Frontiers in Psychology*, *12*, 699715. https://doi.org/10.3389/FPSYG.2021.699715/BIBTEX

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4

Chow, C. W., Haddad, K. & Singh, G. (2007). Human Resource Management, Job Satisfaction, Morale, Optimism, and Turnover. *International Journal of Hospitality & Tourism Administration*, *8*(2), 73–88. https://doi.org/10.1300/J149V08N02_04

Coombs, M. A. (2024). Strategies to Lower Employee Turnover in the Retail Industry. *Walden Dissertations and Doctoral Studies*. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=17326&context=dissertations

de Oliveira, J. M., Zylka, M. P., Gloor, P. A. & Joshi, T. (2019). Mirror, Mirror on the Wall, Who Is Leaving of Them All: Predictions for Employee Turnover with Gated Recurrent Neural Networks. *Studies on*

Entrepreneurship, Structural Change and Industrial Dynamics, 43–59. https://doi.org/10.1007/978-3-030-17238-1_2

Fukui, S., Wu, W., Greenfield, J., Salyers, M. P., Morse, G., Garabrant, J., Bass, E., Kyere, E. & Dell, N. (2023). Applying Machine Learning to Human Resources Data: Predicting Job Turnover among Community Mental Health Center Employees. *Pmc.Ncbi.Nlm.Nih.GovS Fukui, W Wu, J Greenfield, MP Salyers, G Morse, J Garabrant, E Bass, E Kyere, N DellThe Journal of Mental Health Policy and Economics, 2023•pmc.Ncbi.Nlm.Nih.Gov*, *26*(2), 63–76. https://pmc.ncbi.nlm.nih.gov/articles/PMC10424701/

Gamba, D., Yu, Y., Yuan, Y., Schoenebeck, G. & Romero, D. M. (2024). Exit Ripple Effects: Understanding the Disruption of Socialization Networks Following Employee Departures. *WWW 2024 - Proceedings of the ACM Web Conference*, 211–222. https://doi.org/10.1145/3589334.3645634

Giermindl, L. M., Strich, F., Christ, O., Leicht-Deobald, U., Redzepi, A., Mikalef, P., Popovic, A., Lundström, J. E. & Conboy, K. (2022). The dark sides of people analytics: reviewing the perils for organisations and employees. *European Journal of Information Systems*, *31*(3), 410–435. https://doi.org/10.1080/0960085X.2021.1927213

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G. & Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. *Machine Learning*, *106*(9–10), 1469–1495. https://doi.org/10.1007/S10994-017-5642-8/FIGURES/6

Grissom, J. A., Nicholson-Crotty, J. & Keiser, L. (2012). Does My Boss's Gender Matter? Explaining Job Satisfaction and Employee Turnover in the Public Sector. *Journal of Public Administration Research and Theory*, *22*(4), 649–673. https://doi.org/10.1093/JOPART/MUS004

Hall, O. P. (2021). Managing employee turnover: machine learning to the rescue. *International Journal of Data Science*, *6*(1), 57. https://doi.org/10.1504/IJDS.2021.117472

Heidemann, A., Hülter, S. M. & Tekieli, M. (2024). Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency. *International Journal of Human Resource Management*, *35*(14), 2343–2366. https://doi.org/10.1080/09585192.2024.2335515;REQUESTEDJOURNAL:JOURNAL:RIJH20;WGROUP:STRING:PUBLICATION

Jaillet, P., Loke, G. & Sim, M. (2022). Strategic workforce planning under uncertainty. *Operations Research*, *70*(2), 1042–1065. https://doi.org/10.1287/OPRE.2021.2183

Ju, B. & Li, J. (2019). Exploring the impact of training, job tenure, and education-job and skills-job matches on employee turnover intention. *European Journal of Training and Development*, *43*(3–4), 214–231. https://doi.org/10.1108/EJTD-05-2018-0045/FULL/HTML

Judge, T. A. & Kammeyer-Mueller, J. D. (2012). Job attitudes. *Annual Review of Psychology*, *63*(Volume 63, 2012), 341–367. https://doi.org/10.1146/ANNUREV-PSYCH-120710-100511/CITE/REFWORKS

Juvitayapun, T. (2021). Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods. *KST 2021 - 2021 13th International Conference Knowledge and Smart Technology*, 181–185. https://doi.org/10.1109/KST51265.2021.9415794

Kim, E. G. & Kim, H. (2021). A meta-analysis on predictors of turnover intention of hospital nurses in South Korea (2000–2020). *Nursing Open*, *8*(5), 2406–2418. https://doi.org/10.1002/NOP2.872

Kim, H. & Stoner, M. (2008). Burnout and Turnover Intention Among Social Workers: Effects of Role Stress, Job Autonomy and Social Support. *Administration in Social Work*, *32*(3), 5–25. https://doi.org/10.1080/03643100801922357

Koo, B., Yu, J., Chua, B. L., Lee, S. & Han, H. (2020). 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 and Tourism*, *21*(4), 371–401. https://doi.org/10.1080/1528008X.2019.1663572

Kovvuri, S. R. & Dommeti, L. D. S. (2022). *Employee Turnover Prediction-A Comparative Study of Supervised Machine Learning Models*. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1679511

Lee, G. J. & Rwigema, H. (2005). Mobley revisited: Dynamism in the process of employee turnover. *International Journal of Human Resource Management*, *16*(9), 1671–1690. https://doi.org/10.1080/09585190500239333

Lee, T. & Mitchell, T. (1994). An Alternative Approach: The Unfolding Model of Voluntary Employee Turnover. *The Academy of Management Review*, *19*(1), 51–89. https://doi.org/https://doi.org/10.2307/258835

Lee, T. W. (1988). How job dissatisfaction leads to employee turnover. *Journal of Business and Psychology*, *2*(3), 263–271. https://doi.org/10.1007/BF01014043

Levenson, A. (2018). Using workforce analytics to improve strategy execution. *Human Resource Management*, *57*(3), 685–700. https://doi.org/10.1002/HRM.21850

Li, Z. & Fox, E. (2023). *Prediction and optimization of employee turnover intentions in enterprises based on unbalanced data*. https://doi.org/10.1371/journal.pone.0290086

Majam, T. & Jarbandhan, D. B. (2022). Data driven human resource management in the Fourth Industrial Revolution (4IR). *Africa's Public Service Delivery & Performance Review*, *10*(1). https://doi.org/10.4102/APSDPR.V10I1.588

Marler, J. H. & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. *The International Journal of Human Resource Management*, *28*(1), 3–26. https://doi.org/10.1080/09585192.2016.1244699

Marler, J., Management, J. B. H. R. & 2017, undefined. (2017). An evidence-based review of HR Analytics. *Taylor & FrancisJH Marler, JW BoudreauThe International Journal of Human Resource Management, 2017•Taylor & Francis*, *28*(1), 3–26. https://doi.org/10.1080/09585192.2016.1244699

Minbaeva, D. B. (2018). Building credible human capital analytics for organizational competitive advantage. *Human Resource Management*, *57*(3), 701–713. https://doi.org/10.1002/HRM.21848

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, *62*(2), 237–240. https://doi.org/10.1037/0021-9010.62.2.237

Morrell, K., Loan-Clarke, J., Arnold, J. & Wilkinson, A. (2008). Mapping the decision to quit: A refinement and test of the unfolding model of voluntary turnover. *Applied Psychology*, *57*(1), 128–150. https://doi.org/10.1111/J.1464-0597.2007.00286.X

Ng, T. W. H. & Feldman, D. C. (2009). Re-examining the relationship between age and voluntary turnover. *Journal of Vocational Behavior*, *74*(3), 283–294. https://doi.org/10.1016/J.JVB.2009.01.004

Olubiyi, O., Smiley, G., Luckel, H. & Melaragno, R. (2019). A qualitative case study of employee turnover in retail business. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01796

Orujaliyev, R. (2024). The Effect of Non-Financial Incentives on Employee Engagement and Employee Retention. *European Journal of Humanities and Social Sciences*, 3–10. https://doi.org/10.29013/EJHSS-24-2-3-10

Park, J., Feng, Y. & Jeong, S. P. (2024). Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques. *Scientific Reports 2024 14:1*, *14*(1), 1–14. https://doi.org/10.1038/s41598-023-50593-4

Polzer, J. T. (2022). The rise of people analytics and the future of organizational research. *Research in Organizational Behavior*, *42*. https://doi.org/10.1016/j.riob.2023.100181

Punnoose, R. & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9). https://doi.org/10.14569/IJARAI.2016.050904

Rahaman, M. A. & Bari, M. H. (2024). Predictive Analytics for Strategic Workforce Planning: A Cross-Industry Perspective from Energy and Telecommunications. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4983349

Rathore, R. & Rathore, S. P. S. (2024). Machine Learning Applications in Human Resource Management: Predicting Employee Turnover and Performance. *International Journal for Global Academic & Scientific Research*, *3*(2), 48–59. https://doi.org/10.55938/IJGASR.V3I2.77

Solomon, C., Mohankumar, D. & Sivanandam, C. (2024). Employee Attrition Analysis using Machine Learning. *2nd International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2024*. https://doi.org/10.1109/IC-ETITE58242.2024.10493429

Tessema, M. T., Tesfom, G., Faircloth, M. A., Tesfagiorgis, M. & Teckle, P. (2021). The "Great Resignation": Causes, Consequences, and Creative HR Management Strategies. *Journal of Human Resource and Sustainability Studies*, *10*, 161–178. https://doi.org/10.4236/jhrss.2022.101011

Tian, H., Iqbal, S., Akhtar, S., Qalati, S. A., Anwar, F. & Khan, M. A. S. (2020). The Impact of Transformational Leadership on Employee Retention: Mediation and Moderation Through Organizational Citizenship Behavior and Communication. *Frontiers in Psychology*, *11*, 513537. https://doi.org/10.3389/FPSYG.2020.00314/BIBTEX

Wallace, J. & Gaylor, K. (2012). A study of the dysfunctional and functional aspects of voluntary employee turnover. *SAM Advanced Management Journal*. https://search.proquest.com/openview/330099feb2610537a13c99819d612709

Wöcke, A. & Heymann, M. (2012). Impact of demographic variables on voluntary labour turnover in South Africa. *International Journal of Human Resource Management*, *23*(16), 3479–3494. https://doi.org/10.1080/09585192.2011.639028

Work Institute. (2024). *Employee Turnover Insights and Trends in 2023*. https://info.workinstitute.com/2023-retention-report

Wynen, J. & Op De Beeck, S. (2014). The impact of the financial and economic crisis on turnover intention in the U.S. federal government. *Public Personnel Management*, *43*(4), 565–585. https://doi.org/https://doi.org/10.1177/0091026014537043

Yoon, S. W. (2023). Introduction to Special Issue: People Analytics. *Human Resource Development Review*, *23*(1), 3–10. https://doi.org/10.1177/15344843231217534

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B. & Zhu, X. (2019). Employee Turnover Prediction with Machine Learning: A Reliable Approach. *Advances in Intelligent Systems and Computing*, *869*, 737–758. https://doi.org/10.1007/978-3-030-01057-7_56

APPENDIX

Appendix A – Supporting Data and Extended Analysis

Figure 36 demonstrates the relationship between marital status and turnover rate, revealing noticeable differences between single and married employees. While single employees make up the majority of the workforce at 66.49%, married employees account for 33.51%. Despite this, turnover is significantly higher among single employees, at 13.22%, compared to just 4.51% for married employees.

This pattern suggests that single employees are more inclined to leave the company than their married colleagues. Several factors may contribute to this disparity. Single employees often have fewer obligations tying them to a specific job, allowing for greater mobility and flexibility to seek new opportunities. Additionally, they may be at an earlier stage in their careers, where job transitions are more frequent as they explore career advancements, salary improvements or different professional experiences.

In contrast, married employees tend to show greater job stability, likely influenced by financial responsibilities, family commitments and a stronger preference for long-term security. Their professional choices may be guided by the need for stability, reducing the likelihood of voluntary turnover.

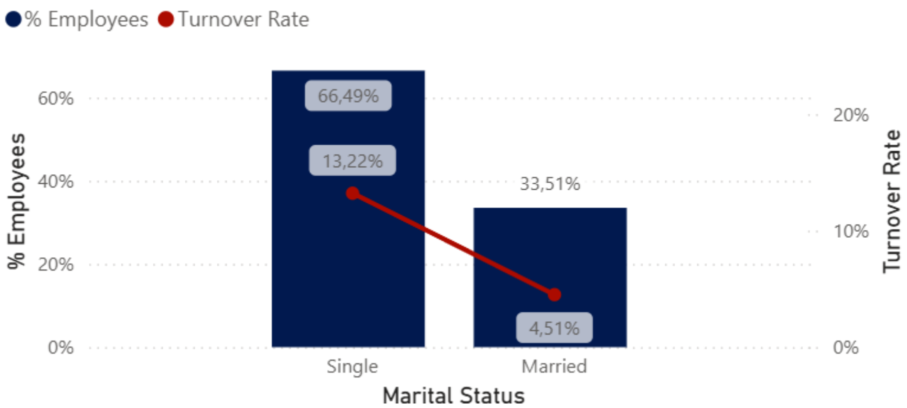

Figure 36 - Workforce Distribution and Turnover Rate by Marital Status

Figure 37 highlights the distribution of employees across different store types based on their seniority. This visualization complements Figure 4 by providing additional insights into how workforce composition varies across store formats. The key takeaway is that convenience stores are overwhelmingly populated by employees with lower seniority, while hypermarkets concentrate the majority of employees with over 10 years of experience. Supermarkets, on the other hand, present a more balanced mix of all seniority levels.

Notably, there are almost no employees with more than 10 years of seniority in convenience stores, reinforcing the idea that these locations are primarily staffed by newer employees. In contrast, hypermarkets have the highest proportion of long-tenured employees, suggesting that they offer greater long-term career stability or progression opportunities. Supermarkets serve as an intermediary category, accommodating both newer and more experienced employees without a strong dominance of any particular seniority group.

This distribution may indicate that convenience stores serve as an entry point for employees who later transition to other store types, while hypermarkets retain a more experienced workforce, possibly due to better career growth opportunities, enhanced job security or differences in store operations that influence long-term retention.

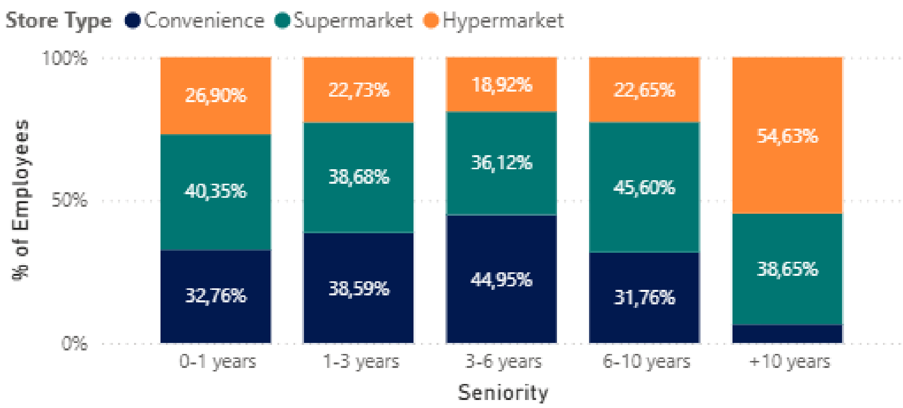

Figure 37 - Store Type Distribution by Seniority

Figure 38 highlights a clear upward trend in average employee age as seniority increases. Employees in the early years of tenure tend to be younger, while those with longer tenure are significantly older on average. The progression is gradual, showing a steady increase in age as employees gain more years in the company.

This pattern suggests that workforce aging is naturally linked to tenure, with younger employees either transitioning to higher seniority levels over time or leaving the company early. These findings reinforce the relationship observed in Figure 6 and Figure 7, where seniority and age were separately analyzed. By directly linking the two variables, Figure 38 confirms that older employees tend to hold higher-tenure positions, providing further validation of the trends previously discussed.

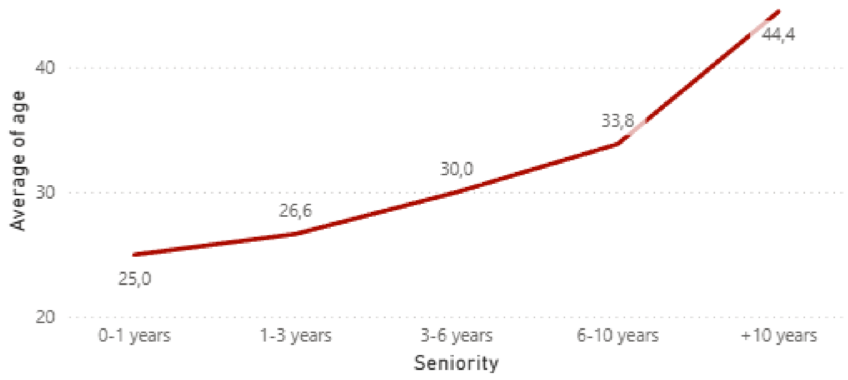

Figure 38 - Age Distribution by Seniority

Figure 39 illustrates the relationship between age and contract type, showing a clear trend where fixed-term contracts are predominantly assigned to younger employees, while permanent contracts become the norm as employees grow older. Most employees under 25 years old are on fixed-term contracts, with this being especially evident in the ≤ 20 years group, where over 95% of employees hold temporary positions. As age increases, the share of permanent contracts rises progressively, surpassing fixed-term contracts around the 25-35 years category. By the 35-50 years age group, permanent contracts become the overwhelming majority and among employees over 50, almost all have permanent employment.

This distribution aligns with the turnover patterns observed in Figure 8, where fixed-term employees exhibit a significantly higher voluntary turnover rate. Given that younger employees are more likely to hold fixed-term contracts, the higher turnover among this group may be more reflective of age and career stage mobility rather than just contract type alone. Employees in early career phases may be more likely to seek new opportunities, change jobs or transition into full-time roles, contributing to the higher turnover rate observed in fixed-term contracts.

Thus, rather than fixed-term contracts being the direct cause of higher turnover, this graph suggests that the tendency of younger employees to have temporary contracts – and their higher mobility in the job market—may be key factors in explaining the observed turnover differences.

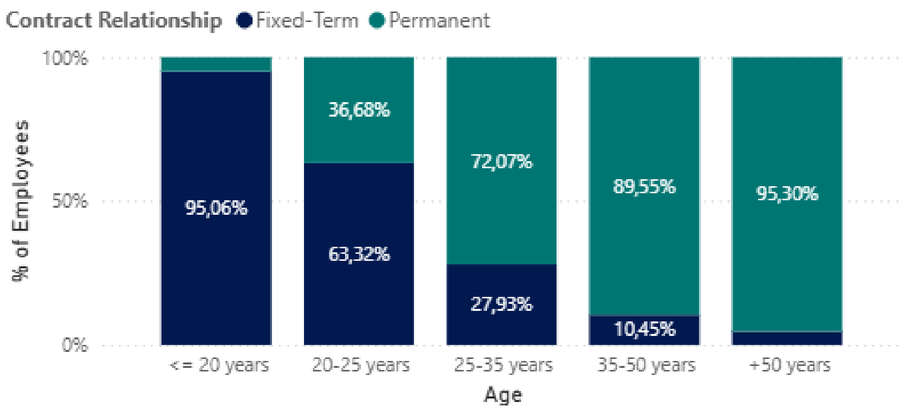

Figure 39 - Contract Type Distribution by Age

Figure 40 establishes the connection between workload type and contract type, reinforcing the patterns observed in Figure 8 and Figure 11, which highlighted the impact of both contract stability and workload type on voluntary turnover. The chart reveals that full-time employees overwhelmingly hold permanent contracts (82.44%), whereas in part-time roles, there is a more even distribution, with 58.13% on fixed-term contracts and 41.87% on permanent contracts.

Given that fixed-term and part-time employees both exhibit higher turnover rates, as seen in Figure 8 and Figure 11, this visualization provides additional evidence that workload type and contract stability are closely linked and may be acting together as key factors influencing voluntary turnover. Since part-time employees are more likely than full-time employees to have fixed-term contracts, their higher turnover may not be due solely to their working hours but also to the fact that they hold less stable contractual agreements, making voluntary exits more frequent.

This insight helps explain why employees with part-time and fixed-term contracts leave more often—not only due to their workload type or contract type individually, but because these two factors often overlap. This reinforces a pattern of higher mobility and job transitions among less stable employment categories.

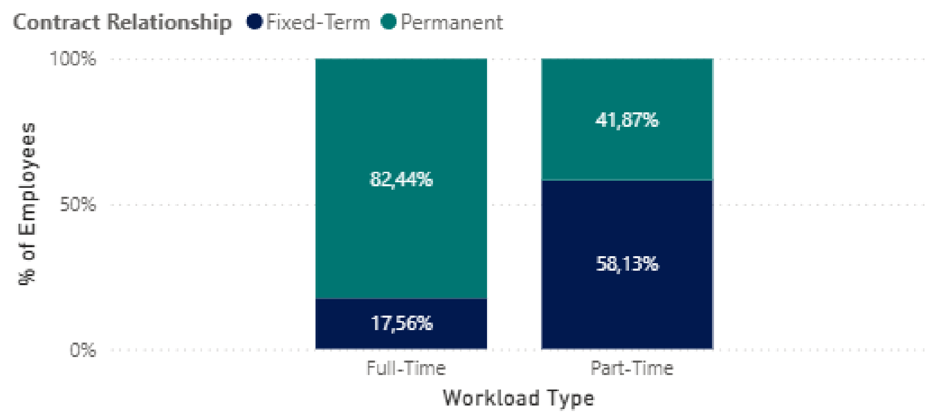

Figure 40 - Contract Type Distribution by Workload Type

Figure 41 presents the distribution of employees across different role clusters and their corresponding average standardized wage. A clear distinction emerges between roles with a higher proportion of employees and those with higher standardized wages.

The Cashier Operator and Flow Operator positions account for nearly half of the workforce, making them the most prevalent roles. However, these positions, along with most other frontline operational functions, maintain a standardized wage level of 15, indicating limited salary progression despite their prevalence.

In contrast, roles that require specialized skills or managerial responsibilities — such as *Main Supervisor*, *Section Coordinator* and those classified under the *Other* category — exhibit significantly higher wages, with standardized values reaching 29-32. The *Other* category is particularly noteworthy, displaying one of the highest average wages despite representing a small fraction of the workforce. This suggests that it comprises highly specialized or senior positions that, while less common, command significantly higher compensation due to their strategic importance or the expertise required.

This deeper breakdown of salary distribution across roles builds upon the insights from Figure 9, where the overall relationship between salary levels and turnover rates was introduced. A more detailed examination of the *Other* category and its wage distribution is provided in in Figure 42, which focuses on specific roles within this classification.

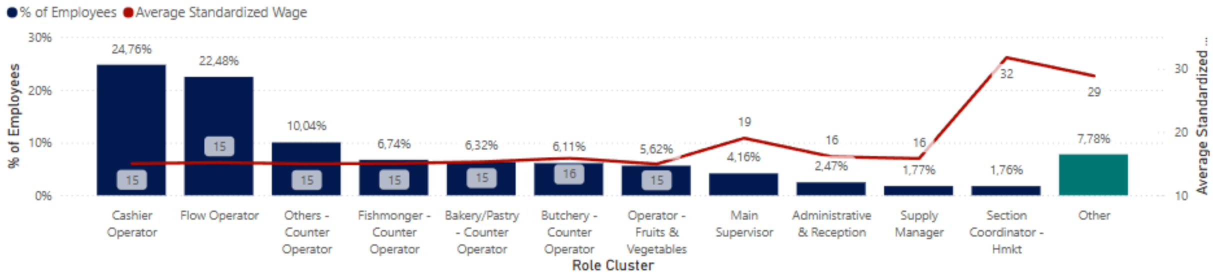

Figure 41 - Workforce Distribution and Standardized Wage by Role Cluster

Figure 42 highlights a diverse mix of role clusters, encompassing both high-responsibility leadership positions and more specialized or transitional roles with fewer employees. The key observation is the clear disparity in standardized wages, with managerial positions standing out as the highest-paid roles, while most operational or support functions remain at lower wage levels.

A particularly notable case is the *Store Director – Hmkt* (Hypermarket), which exhibits a standardized wage of 111, significantly surpassing all other roles. This reinforces the hierarchical wage structure, where leadership roles overseeing larger and more complex store formats command the highest salaries. Similarly, the *Store Director – Smkt* (Supermarket) and *Area Coordinator - Hmkt* (Hypermarket) also demonstrate elevated wages relative to their workforce proportion, confirming the premium placed on senior managerial positions.

Additionally, it is important to highlight that the highest-ranking position within convenience stores, the Cvn (Convenience) *Store Manager*, has a significantly lower standardized wage (38) compared to its counterparts in supermarkets and hypermarkets. This suggests that store format plays a key role in defining salary structures, with larger stores demanding greater managerial complexity, broader responsibilities and higher compensation levels. The stark contrast between convenience store leadership and supermarket/hypermarket directors further reinforces the progressive wage differentiation based on store size and operational scale.

In contrast, roles such as *Operator 360* and *Sales Operator* maintain both low representation and low standardized wages, indicating that these positions are either entry-level, temporary or require fewer specialized skills. Meanwhile, technical or operational roles such as *Quality Control Operator*, *Customer Support Service* and *Maintenance Technician* occupy a middle ground, with wages slightly above the base level but still well below managerial roles.

A particularly interesting observation is the presence of *Area Coordinator* roles across different store types (Hypermarket, Supermarket and Convenience). These positions exhibit varying wage levels, with hypermarket coordinators earning significantly more than their counterparts in smaller store formats. This suggests that store size and complexity influence compensation, even within the same hierarchical level.

Overall, this graph reinforces the dual nature of role clusters within the company: senior leadership roles are scarce but highly compensated, while operational, support and transitional positions tend to have lower wages and representation. The significant wage gap between store directors and lower-level roles reflects the company's structured career progression, where greater managerial responsibility translates into higher salaries.

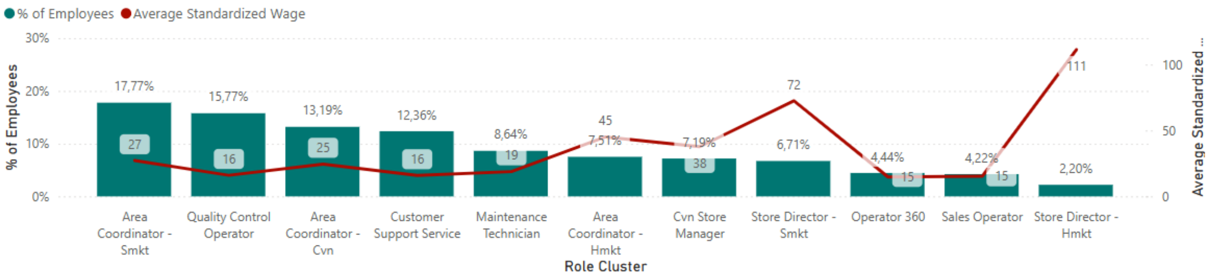

Figure 42 - Workforce Distribution and Standardized Wage by Role Cluster (Others)