NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

Employee Turnover Predictive Model

A model to predict who is more likely to leave a company

Nuno Rafael de Almeida Durães

Project Work

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Employee Turnover Predictive Model

A model to predict who is more likely to leave a company

by

Nuno Rafael de Almeida Durães

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Professor Roberto Henriques, PhD

July, 2024

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 Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15th July, 2024

DEDICATION

Gostaria de dedicar este estudo, primeiramente aos meus pais, por me terem possibilitado perseguir esta nova etapa da minha vida, à minha cara-metade e melhor amiga, e por último às pessoas que me permitiram ter os dados para concluir este marco. Todos os que me suportaram e ajudaram desde o inicio, que serviram como autentica fonte de inspiração e força para fechar mais um grande capítulo.

ABSTRACT

This master's thesis explores the predictive modeling of employee turnover within a company in the Retail & Food Industry, using a data-driven approach to identify potential leavers and understand the dynamics affecting their decisions. Employing machine learning techniques such as Logistic Regression, Random Forest, and Neural Networks, the study focuses on optimizing the prediction of employee turnover through sophisticated model selection and hyperparameter tuning. The research began with an extensive data preparation phase, which involved cleaning, normalization, and transformation of the dataset to ensure robustness and relevancy for model training. This process also included the application of SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance within the dataset, ensuring that the predictive performance was not biased towards the majority class. Key features influencing turnover, such as job satisfaction, management styles, compensation packages, and career progression opportunities, were identified and engineered to enhance the predictive performance of the models. Several models were evaluated, with the ensemble approach integrating Random Forest, Gradient Boosting, and Neural Networks showing the most promising results. This ensemble model, optimized for high precision in predicting 'Active' status without overfitting, achieved remarkable accuracy (95.1% to 95.8%), precision for label 0 (non-leavers) up to 92.5%, and an ROC-AUC score demonstrating excellent classification capabilities (up to 0.983). The refined models significantly outperformed initial predictions, highlighting the effectiveness of the feature selection and machine learning techniques employed. The findings suggest that the integrated approach can effectively predict employee turnover, providing HR departments with a valuable tool for strategic human resource planning. This predictive capability enables proactive interventions tailored to mitigate turnover and enhance employee retention strategies. In conclusion, this thesis not only demonstrates the applicability of advanced analytical techniques to real-world HR challenges but also lays the groundwork for future research. It suggests exploring further cross-industry applications, integration of additional data sources, and employing alternative modeling techniques to expand the model's robustness and adaptability across different organizational contexts.

KEYWORDS

Employee Termination; Employee Turnover; Predictive Modelling; Machine Learning; Human Resources Analytics; Employee Retention

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

Statement of Integrity i
Dedication ii
Abstract iii
Keywords iv
List of Figures vii
List of Tables viii
List of Abbreviations and Acronyms ix
1. Introduction 1
2. Literature review 3
2.1. Objectives & Research Methods 3
2.2. Related Work 6
2.2.1. Understanding and Addressing Employee Turnover 7
2.2.2. Employee Turnover Prediction Models 7
2.2.2.1. Machine Learning Techniques and Predictive Frameworks 8
2.2.2.2. Application in Specific Contexts and Industries 8
2.2.2.3. Features Across Studies 8
2.2.2.4. Feature Selection 9
2.2.2.5. Comparative Analysis and Synthesis of Model Performance 10
2.2.2.6. Innovations and Future Directions 11
2.3. Conclusion 11
3. Methodology 13
3.1. Data Understanding 14
3.1.1. Initial Dataset 14
3.1.2. Target Variable 18
3.1.3. Data Types 19
3.1.4. Descriptive Analysis 19
3.1.5. Exploratory Data Analysis (EDA) 21
3.1.5.1. Univariate Analysis 21
3.1.5.2. Bivariate Analysis – Correlation and Dependency 25
3.1.5.3. Preliminary Findings 30
3.1.5.4. Addressing Outliers 31
3.1.5.5. Concluding EDA 32
3.2. Data Preparation 33

3.2.1. Feature Engineering 33
3.2.1.1. Normalization 33
3.2.1.2. Encoding 33
3.2.2. Class Imbalance 34
3.2.2.1. Initial Examination of Class Imbalances 34
3.2.2.2. Synthetic Minority Over-sampling Technique and Tomek Links 34
3.2.2.3. Evaluation of Rebalanced Dataset 34
3.2.2.4. Results and Implications 34
3.2.3. Feature Selection and Dimensionality Reduction 35
3.2.3.1. Recursive Feature Elimination (RFE) 35
3.2.4. Concluding Data Preparation 36
3.3. Modelling 36
3.3.1. Model Selection 36
3.3.2. Hyperparameter Tuning 38
3.3.3. Selection Methodology 38
3.4. Evaluation Process 39
3.5. Conclusion 39
4. Results and discussion 41
4.1. Results 41
4.1.1. Results Analysis 41
4.1.1.1. Models Performance 41
4.1.1.2. Feature Importance 47
4.1.1.3. Permutation Importance 48
4.1.1.4. Partial Dependence Plots 49
4.1.1.5. Sensitivity Analysis 51
4.1.1.6. Interaction Effect Analysis on Top 5 Important Features 55
4.1.1.7. Probability of the Active Employees 57
4.2. Discussion and Findings 59
4.3. Conclusions 59
4.3.1. Limitations 61
4.3.2. Future Research 62
5. Conclusions and future work 64
Bibliographical References 67

LIST OF FIGURES

Figure 1 - Methodology diagram 13
Figure 2 - Employee Distribution by Class 19
Figure 3 - Metric Features Histogram 22
Figure 4 - Categorical Features Histograms 24
Figure 5 - Boolean Features Histograms 25
Figure 6 - Correlation between Metric features. 26
Figure 7 - Metric features box plot for outlier detection – part 1 31
Figure 8 - Metric features box plot for outlier detection – part 2 32
Figure 9 - Modelling Process Methodology Scheme 38
Figure 10 - Best Model Top 5 Features by Importance 47
Figure 11 - Permutation Feature Importance on mean importance value 48
Figure 12 - Partial Dependence Analysis on the Top 5 features 50
Figure 13 - Sensitivity Analysis on feature “NumberOfPositionsChanges” 52
Figure 14 - Sensitivity Analysis on feature “Monthly Remuneration” 52
Figure 15 - Sensitivity Analysis on feature “Seniority” 53
Figure 16 - Sensitivity Analysis on feature “YearsLastTLMeeting” 53
Figure 17 - Sensitivity Analysis on feature “NumberSalaryIncreases” 54
Figure 18 - Interaction Effect Analysis on Top 5 Important Features 56
Figure 19 - Distribution of Termination Probabilities for Active Employees 58

LIST OF TABLES

Table 1 - Final list of papers 4
Table 2 - Table of the key features across studies 9
Table 3 - List of Initial Features 16
Table 4 - Average values by Class 20
Table 5 - Metric features correlation with Target feature 27
Table 6 - Chi Squared test between Categorical features – part 1 27
Table 7 - Chi Squared test between Categorical features – part 2 28
Table 8 - Chi Squared test between Categorical features – part 3 29
Table 9 - Chi Squared test between Categorical features – part 4 29
Table 10 - Chi Squared test between Categorical features and Target features 30
Table 11 - Initial Model Application Performance Summary 42
Table 12 - Refined Models Performance Summary 44
Table 13 - Ensemble Models Performance Metrics Summary 45

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AUC	Area Under Curve
DT	Decision Tree
EDA	Exploratory Data Analysis
GBC	Gradient Boosting Classifier
HR	Human Resources
HRIS	Human Resources Information System
KNN	K-Nearest Neighbors
LR	Logistic Regression
MLP	Multi-Layer Perceptron
RFC	Random Forest Classifier
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
TL	Team Lead
SDG	Sustainable Development Goals
ML	Machine Learning
HRIS	Human Resources Information System
SMOTE	Synthetic Minority Over-sampling Technique
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
EDA	Exploratory Data Analysis

1. INTRODUCTION

Intricate and multifaceted organizations are dynamic environments where diverse factors interact and influence each other. Among the foundational pillars of an organization, its people stand out as a core element vital for success. Employees are one of the most essential assets for a company, contributing with intellect, knowledge, and expertise that are essential for an organization and its development. Their capabilities and experience are crucial in developing an organization from its core, being the central function of a company in addressing the different challenges in its environment. Assessing organizational efficiency gravitates toward a critical focus on the workforce. In this context, evaluating employee turnover is a pivotal tool for evaluation. This metric provides valuable insights into how it impacts various points of organizational performance, including sales, return on equity, customer service quality, profitability, and others. Satisfaction is a key point for employee turnover. High employee satisfaction is reflected in higher productivity, where people tend to feel committed to achieving higher results, contributing to higher organizational success (Mazzetti & Schaufeli, 2022; Schneider & Pulakos, 2022).

Employee turnover, defined as voluntary departures from an organization, has garnered significant attention due to its demonstrated negative impact on organizational dynamics (Han, 2020). The repercussions extend beyond mere departures, translating into tangible costs for the company. High turnover rates increase costs for the organization, stemming from recruitment and replacement expenses, training, and the development of new hires to align them with the company's culture (Alsheref et al., 2022). Consequently, one of the core objectives of Human Resources departments is to comprehend the reasons for attrition and formulate effective strategies to mitigate its impact. At the organizational level, departing employees have valuable experience and connections that are lost when they leave (Setiawan et al., 2020). Given its potential to influence and disrupt organizational stability, employee turnover has become a central study area for numerous Human Resources and People Departments. The primary objective is to discern the significant factors influencing employee exits and implement effective solutions to mitigate high turnover rates. This strategic approach aims to sustain employee satisfaction and foster continuous growth of the organization.

Within People Analytics, there is a notable research gap concerning the timing of employee attrition. While existing research has explored the factors contributing to attrition and has concentrated on developing predictive attrition models based on IBM or Kaggle datasets, so artificial datasets, there is a limited understanding of the application on real company datasets, where the data represent a living organization with real factors being considered in the data imputation. Moreover, a significant gap in leveraging advanced techniques, such as artificial neural networks, in building the model represents a noteworthy opportunity. This reality aspect is crucial for organizations to implement timely intervention strategies in the real world, facing the actual industry and the specifications and uniqueness of the specifics of

the organization as one, as it is essential to understand if using more advanced algorithms would proportionate better model results. To address this research gap, this study is centered on the following key research questions:

- How can we enhance attrition prediction accuracy by employing artificial neural networks compared to conventional classifiers?
- What are the primary indicators and patterns that precede employee departures in business workplaces?
- Can we accurately predict employee termination and termination probability using real company data?

To achieve these objectives, the study will analyze historical attrition data from one company, emphasizing the aspects of attrition patterns, and subsequently develop a predictive model capable of pointing out what employees have a higher turnover risk and identifying potential attrition events. This comprehensive approach aims to deepen our understanding of attrition patterns and significantly enhance predictive capabilities within People Analytics.

2. LITERATURE REVIEW

This literature review on employee turnover within the realm of people analytics adopts a structured approach to examine the impact of turnover and the methodologies for predicting it. The review selectively focuses on significant studies from reputable journals and conferences by setting clear objectives aligned with research questions. Through meticulous analysis, it delves into predictive models, their efficacy, identified predictors, and feature selection techniques used in the selected papers. This process uncovers common themes and trends and highlights research gaps, synthesizing insights to provide a comprehensive understanding of predicting employee turnover intentions.

2.1. OBJECTIVES & RESEARCH METHODS

The primary objective was to provide a comprehensive overview of the existing knowledge related to the topic in this study. We analyzed the datasets and techniques and explored the results as the principal predictors observed. In more detail, our objectives are:

1. Identify key features and patterns associated with employee leaves.

Identify commonly reported features as main predictors and patterns that precede employee departures in various business environments or the different produced models. The rationale is to deepen the understanding of which variables are most predictive of employee turnover and how these variables have been utilized in existing models. This knowledge is crucial for both developing new predictive models and refining existing ones.

2. Explore the best methods for predicting employee leaves in real-world organizations.

Investigate the application of predictive models in real organizations using authentic and representative data from actual employees and company operations. This includes approaches that estimate the probability of employee departure. This aspect of turnover modeling is particularly challenging and requires an in-depth examination of how different models have been structured and validated in previous research.

To present the most important and relevant information for the literature review in this study, we follow a 3-stage approach to select the papers. In the 1 stage, we present the databases used and the searching rationale; in the second stage, we narrow down papers explicitly related to the topic that presented specific models, techniques, and results in predicting

employee turnover; and in the third stage, we selected the Human Resources and People related papers to found a common understandings on the related topic.

In the first stage, the papers search was done between September and December 2023 using the different databases: Google Scholar, Scopus and ResearchGate. We decided to use the following strings to search for related papers and works, because we believe they best represent the topic in search related to employee turnover:

- "employee" AND "turnover"
- "predicting" AND "employee" AND "turnover"
- "employee" AND "turnover" AND "prediction"
- "employee" AND "churn" AND "prediction"

The number of papers we obtained in the first stage was large since all papers related to the topic were included. We first analyzed the paper abstract to understand its relation to the problem, narrow the number of papers, and have a higher paper value for the study we want to develop. In the third stage, we evaluated the quality of the paper using “Scimago Journal & Country Rank”. Only papers from Q1 and Q2 were selected to ensure the selection of high-quality papers. The criteria for selection included the use of machine learning and artificial intelligence techniques in predicting employee turnover, with an emphasis on papers published within the last five years to ensure the relevance and currency of the data. Using this approach, we found 18 papers, displayed in the list below, that were used as sources and inspiration.

The 18 papers are presented below as follows in the Table 1 below.

Table 1 - Final list of papers

Authors	Year	Title	Journal/Source
Akasheh, M. A., Malik, E. F., Hujran, O., & Zaki, N.	2024	A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review	Expert Systems with Applications
Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A.	2021	Employee Attrition Prediction Using Deep Neural Networks	Computers

Alsheref, F. K., Fattoh, I. E., & M. Ead, W.	2022	Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms	Computational Intelligence and Neuroscience
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
Chung, D., Yun, J., Lee, J., & Jeon, Y.	n.d.	Predictive model of employee attrition based on stacking ensemble learning	N/A
El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M.	2020	Predicting employee attrition using tree-based models	International Journal of Organizational Analysis
Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E.	2020	Predicting Employee Attrition Using Machine Learning Techniques	Computers
Han, J. W.	2022	A review of antecedents of employee turnover in the hospitality industry on individual, team and organizational levels	International Hospitality Review
Kang, I. G., Croft, B., & Bichelmeyer, B. A.	2021	Predictors of Turnover Intention in U.S. Federal Government Workforce: Machine Learning Evidence That Perceived Comprehensive HR Practices Predict Turnover Intention	Public Personnel Management
Mohammed, A. Q.	2019	HR ANALYTICS: A MODERN TOOL IN HR FOR PREDICTIVE DECISION MAKING	JOURNAL OF MANAGEMENT
Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B.	2023	Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data	Benchmarking: An International Journal
Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S.	2021	An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection	Mathematics

Naz, K., Siddiqui, I. F., Koo, J., Khan, M. A., & Qureshi, N. M. F.	2022	Predictive Modelling of Employee Churn Analysis for IoT-Enabled Software Industry	Applied Sciences
Tian, X., Pavur, R., Han, H., & Zhang, L.	2023	A machine learning-based human resources recruitment system for business process management: Using LSA, BERT and SVM	Business Process Management Journal
Wang, X., & Zhi, J.	2021	A machine learning-based analytical framework for employee turnover prediction	Journal of Management Analytics
Wild Ali, A. B.	2021	Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm	Wireless Personal Communications
Mazzetti & Schaufeli, 2022	2020	The impact of engaging leadership on employee engagement and team effectiveness	PLOS ONE
Schneider & Pulakos, 2022	2020	Expanding the I-O psychology mindset to organizational success	Industrial and Organizational Psychology, Cambridge Core

2.2. RELATED WORK

The evolution of employee turnover literature has seen various contributions, focusing on understanding the implications for organizations and the root causes of attrition. With the rise of data mining and predictive models as solutions across different business domains, integrating these approaches into turnover analysis and prediction has become imperative. Machine Learning (ML), a versatile tool applied in diverse fields such as chemistry, bioinformatics, medicine, biology, finance, and HR, has become integral to this integration (Al Akasheha et al., 2023).

Driven by this central focus and need, the analysis of turnover data and the development of turnover predictive models have gained traction. The goal is to unravel the factors contributing to turnover and predict which employees are more likely to leave the company and when they intend to do it. In this context, supervised learning, a category of ML, becomes particularly relevant. This approach involves tasks like classification and regression, performed

using various algorithms such as artificial neural networks (ANN), logistic regression (LR), support vector machines (SVM), random forests (RF), and decision trees (DT). Supervised learning relies on labelled datasets and training algorithms based on known outputs (employees who leave) for each known input. The trained model can then predict outcomes for unseen data, identifying employees at risk of departure and the key contributors to turnover. This empowers companies to take proactive measures, developing strategies to reduce turnover and retain valuable employees.

2.2.1. Understanding and Addressing Employee Turnover

Employee turnover remains a pivotal challenge in organizational management, particularly in the hospitality industry, where its multifaceted nature demands a nuanced understanding of its antecedents and effective retention strategies. This review synthesizes findings from two comprehensive studies, blending their insights to offer a more holistic understanding of turnover.

The antecedents of employee turnover can be broadly categorized into individual, team, and organizational levels. At the individual level, job satisfaction and work-life balance are critical (Han, 2022; Al-Suraihi et al., 2021). Employees' personal experiences and fulfilment in the job role significantly influence their decision to stay or leave. The team level introduces another dimension, where the dynamics of team relationships and the quality of interactions among team members become influential (Han, 2022). Positive team dynamics can enhance job satisfaction and reduce turnover intentions.

At the organizational level, broader factors like compensation, managerial support, and the overall work environment play a crucial role (Al-Suraihi et al., 2021). These elements reflect the organization's commitment to its workforce, impacting employee perceptions and overall satisfaction. Managerial styles and management systems are key factors (Al-Suraihi et al., 2021). The management approach to handling employee concerns, fostering a supportive environment, and involving employees in decision-making processes can significantly impact turnover rates.

2.2.2. Employee Turnover Prediction Models

Predicting employee turnover is critical to human resources management, especially in industries like hospitality, pharmaceuticals, and the public sector. Recent advancements in machine learning (ML) and artificial intelligence (AI) have opened new avenues for predicting and managing employee turnover. This review synthesizes findings from various studies, each contributing unique insights into the development of predictive models using ML and AI techniques.

2.2.2.1. Machine Learning Techniques and Predictive Frameworks

A systematic literature review by Akasheh et al. (2023) highlights a decade of research on ML techniques in predicting employee turnover. The study provides an exhaustive analysis of various ML algorithms and their applications in the domain. It emphasizes the importance of high-quality empirical studies and identifies critical factors and knowledge gaps in this area. Similarly, Wang and Zhi (2021) discuss a machine learning-based analytical framework, addressing challenges like influential human factors and the selection of candidate models. They propose a streamlined approach to feature engineering and ensemble learning for robust prediction models.

Tian et al. (2023) introduce a novel approach using LSA, BERT, and SVM to enhance recruitment processes, indicating the potential of these techniques in employee selection and turnover prediction. Alsheref et al. (2022) focus on an ensemble model based on ML algorithms for predicting employee churn, emphasizing autotuning techniques to enhance predictive capabilities. El-Rayes et al. (2020) investigate using tree-based models for predicting employee attrition, highlighting the influence of firm cultural and management attributes.

2.2.2.2. Application in Specific Contexts and Industries

The study by Wild Ali (2021) demonstrates the effectiveness of the Random Forest Classifier and Intensive Optimized PCA in predicting employee turnover, using the ORACLE ERP dataset as a case study. Kang et al. (2021) employ machine learning to identify predictors of turnover intention in the U.S. Federal Government workforce, indicating the effectiveness of comprehensive HR practices in managing turnover.

Mozaffari et al. (2023) presents a model combining machine learning with qualitative data analysis for predicting employee attrition in the pharmaceutical industry. This approach provides a holistic understanding of turnover factors in a specific industry context. Najafi-Zangeneh et al. (2021) introduce an improved framework emphasizing feature selection for attrition prediction, highlighting the importance of the pre-processing stage in ML models.

2.2.2.3. Features Across Studies

In predicting employee turnover, selecting features (predictors) plays a crucial role in the model's performance. The studies reviewed provide insights into the standard features used across various models and highlight the importance of feature selection techniques in enhancing predictive accuracy. Several predictors emerge consistently across different studies, underlining their significance in turnover prediction. To summarize the previous description, we added the Table 2 with the key features identified in the different studies.

Table 2 - Table of the key features across studies

Factor	Description	References
Job Satisfaction	Job satisfaction is often a primary predictor of turnover intention. Factors contributing to job satisfaction include work environment, role clarity, and job security.	Kang et al. (2021), El-Rayes et al. (2020)
Organizational Factors	Factors like company culture, managerial support, and compensation are crucial. Organizational policies and practices also play a significant role.	El-Rayes et al. (2020), Kang et al. (2021)
Employee Engagement and Loyalty	Employee engagement and loyalty are key predictors. This includes involvement in decisions and alignment with organizational goals.	Kang et al. (2021)
Personal Attributes	Personal attributes such as age, tenure, and educational background influence turnover.	Fallucchi et al. (2020)
Work-Life Balance	The balance between professional and personal life can impact an employee's decision to stay with or leave an organization.	Akasheh et al. (2023)

2.2.2.4. Feature Selection

Effective feature selection is vital for improving model performance. Several studies reviewed employ advanced feature selection techniques:

- Intensive Optimized PCA: Wild Ali (2021) uses this technique for dimensionality reduction, enhancing the Random Forest Classifier's effectiveness in identifying turnover predictors.
- Filter-Based Methods: Naz et al. (2022) apply filter-based methods for feature selection, improving the performance of machine learning algorithms in predicting employee churn.
- "Max-Out" Algorithm: Najafi-Zangeneh et al. (2021) introduce this algorithm in the pre-processing stage, emphasizing its role in enhancing logistic regression model performance for employee attrition prediction.

The choice and optimization of features are critical for the success of turnover prediction models. Advanced feature selection techniques like optimized PCA and filter-based methods help identify the most relevant predictors, thereby improving the accuracy and reliability of the predictions. These techniques manage the complexity of the data and avoid overfitting, ensuring that the models capture the actual patterns and relationships in the data.

2.2.2.5. Comparative Analysis and Synthesis of Model Performance

In evaluating the performance of employee turnover prediction models, it is essential to consider the variety of algorithms used, the complexity of the data, and the specific application context. Comparative analysis reveals significant insights into how different models perform and interrelate, offering a nuanced understanding of their capabilities. Akasheh et al. (2023) conducted an extensive systematic review covering a wide range of machine learning algorithms, providing a thorough overview of their effectiveness across various contexts. This review did not single out the best model but underscored the diversity in model performance based on application scenarios. In parallel, Wang and Zhi (2021) emphasized the significance of ensemble methods in their machine learning-based analytical framework, advocating for a meticulous approach to model selection tailored to specific needs.

Advanced techniques such as deep learning, as demonstrated by Tian et al. (2023) through the use of LSA, BERT, and SVM in recruitment processes, showcase the potential of these models in handling complex language data. This is particularly evident in BERT's capability to understand intricate human language patterns, though a direct performance comparison with other models was not provided. The sophistication of ensemble models is further highlighted by Alsheref et al. (2022), who combined various machine learning algorithms with autotuning for hyperparameter optimization, suggesting enhanced predictive capabilities. Tree-based models, including decision trees, random forests, and gradient-boosted trees, were effectively utilized by El-Rayes et al. (2020), who provided valuable insights into their applicability for predicting employee attrition based on organizational culture and management attributes. Similarly, Wild Ali (2021) demonstrated the efficacy of combining the Random Forest Classifier with Intensive Optimized PCA, achieving notable accuracy in predictions using the ORACLE ERP dataset. The role of machine learning in identifying predictors of turnover intention was explored by Kang et al. (2021), who provided a comprehensive list of predictors without focusing extensively on model performance metrics. In contrast, Fallucchi et al. (2020) highlighted the Gaussian Naïve Bayes classifier's superior recall rate and low false negative rate, indicating its high potential for accurate employee attrition predictions.

Among the highest reported accuracies is the 98% achieved by Naz et al. (2022) using a decision tree algorithm in the IoT-enabled software industry. Chung et al. (2022) also demonstrated superior performance with their stacking ensemble learning model, achieving high accuracy, F1-score, AUC, precision, and recall, thus outperforming individual models. Deep neural networks utilized by Al-Darraji et al. (2021) further emphasize the effectiveness of deep learning techniques in attrition prediction. Comparing these models reveals a trend towards higher accuracy and robustness in advanced techniques like deep learning and ensemble methods. Notable models include the stacking ensemble model by Chung et al. (2022) and the decision tree algorithm by Naz et al. (2022), which stand out for their high accuracy rates. The success of the Gaussian Naïve Bayes classifier in Fallucchi et al. (2020) and

the Random Forest Classifier in Wild Ali (2021) highlights the importance of selecting algorithms that align with data characteristics and prediction goals. In synthesizing the performance of various employee turnover prediction models, accuracy and precision emerge as critical metrics. High accuracy rates reported by Naz et al. (2022) and Chung et al. (2022) underscore these models' effectiveness in correctly identifying turnover and retention cases. Recall and F1 scores, as seen in Fallucchi et al. (2020) and Chung et al. (2022), indicate balanced performance between precision and recall, crucial for minimizing false positives and negatives.

Ensemble techniques and deep learning approaches, as evidenced by Alsheref et al. (2022) and Al-Darraji et al. (2021), generally enhance performance by capturing complex data patterns. Tailored combinations, such as Wild Ali's (2021) use of Random Forest Classifier with Intensive Optimized PCA, highlight the significance of customized approaches based on data characteristics. Feature selection methods like Intensive Optimized PCA (Wild Ali, 2021) and filter-based techniques (Naz et al., 2022) further improve model performance by focusing on relevant predictors and reducing overfitting risks. In conclusion, the effectiveness of employee turnover prediction models is highly contingent on the specific context and dataset nature, with advanced machine learning techniques demonstrating superior performance and robustness. The synthesis of various studies provides a comprehensive understanding of these models' capabilities, emphasizing the importance of contextually appropriate model selection and feature optimization for achieving the best predictive outcomes.

2.2.2.6. Innovations and Future Directions

Fallucchi et al. (2020) explore the Gaussian Naïve Bayes classifier for predicting employee attrition, emphasizing its applicability in HR. Naz et al. (2022) integrate IoT with advanced ML techniques for employee turnover prediction in the software industry, achieving high accuracy with the decision tree algorithm.

Chung et al. (2022) proposes a stacking ensemble learning model for attrition prediction, achieving superior performance compared to single-model approaches. Mohammed (2019) discusses the role of HR analytics in predictive decision-making, categorizing it into descriptive, predictive, and optimization analytics. Finally, Al-Darraji et al. (2021) explore the potential of deep neural networks in predicting employee attrition, highlighting the effectiveness of deep learning techniques.

2.3. CONCLUSION

Standard features such as job satisfaction, organizational factors, employee engagement, personal attributes, and work-life balance are critical predictors of employee turnover.

Advanced feature selection techniques, as demonstrated in the reviewed papers, play a pivotal role in enhancing the models' ability to predict turnover accurately. Future research and model development should continue refining feature selection methods to improve predictive accuracy in employee turnover models.

While there is no one-size-fits-all model for predicting employee turnover, advanced techniques like deep learning and ensemble methods generally demonstrate higher accuracy and robustness. The choice of the model should be tailored to the specific context, data characteristics, and the aspects of employee turnover that an organization aims to predict. Future research may benefit from focusing on these advanced techniques, especially in complex and large-scale datasets. The studies reviewed demonstrate a broad spectrum of approaches and methodologies employed in employee turnover prediction. The research points towards an increasingly sophisticated and nuanced understanding of turnover prediction, from traditional ML algorithms to advanced AI techniques like deep learning and ensemble models. This literature review underscores the significance of these developments in enhancing predictive accuracy and providing actionable insights for organizations across various industries.

The performance of employee turnover prediction models varies significantly across different studies, influenced by the choice of algorithms, advanced techniques like ensemble methods and deep learning, and effective feature selection methods. The trend towards higher accuracy and balanced precision and recall in recent studies indicates the evolving sophistication of these models. Future research should continue to explore and refine these advanced techniques, especially in the context of complex and diverse datasets, to further enhance the predictive capabilities in employee turnover.

3. METHODOLOGY

In this section, we apply data-driven methodologies to explore employee turnover, utilizing Python and Microsoft Excel for data manipulation and analysis. Python, an open-source programming language, is renowned for its versatility and powerful libraries in data science, such as Pandas (pandas-dev, 2024) and NumPy (Harris et al., 2020). Microsoft Excel, a widely used spreadsheet software, offers robust features for data organization and preliminary analysis (Microsoft Corporation, 2024). Our investigation tackles a main analytical challenge: classification to identify potential leavers. We also use Keras (Chollet, 2015), the neural network package, to develop a predictive model for enhanced results.

For classification, we integrated Logistic Regression, Random Forest, and Neural Networks on feature selection, employing a diverse number of models, including KNN Classifier and Gradient Boosting, among others, and leading to an ensemble of the top three performers, to train the final predictive model.

This strategic approach ensures a comprehensive analysis tailored to the nuanced dynamics of employee turnover. Figure 1 represents the visual representation of the methodology in a diagram.

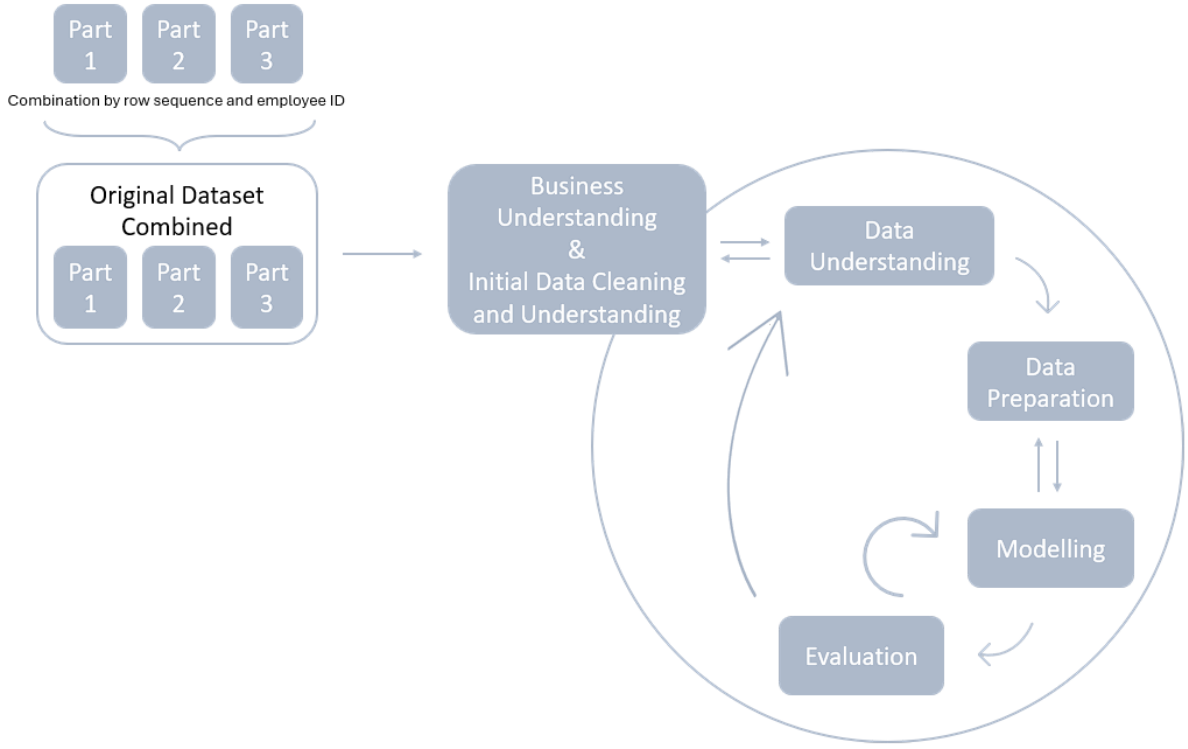


Figure 1 - Methodology diagram

3.1. DATA UNDERSTANDING

3.1.1. Initial Dataset

The procurement of the dataset for this study was facilitated through a collaborative agreement with a company in the Retail & Food Industry, where the importance of maintaining confidentiality and ethical stewardship of data was always a priority. Spanning from 2017 through the end of 2023, the company provided three different datasets, each representing a unique spectrum of employee-related metrics, which needed prior consolidation into a final, singular, and harmonized dataset to work on. The different datasets had different lines per employee, corresponding to the changes made for each employee during their time in the company. To consolidate, we considered the last position of each set and joined it into a single frame with the respective information. For some column aggregation, we had to input the position information before the last. The consolidated dataset, to be prepared for the predictive turnover model, involved a series of transformations and cleanings to ensure the data's integrity, readability, and relevance. This intricate process was not merely about data amalgamation. It represented a critical foundation for our subsequent analytical endeavors. By transforming these datasets into a unified format, we ensured that each employee's profile was encapsulated within a comprehensive entry line. The following steps were considered before being imported into a Python notebook:

- Tracking Career Progression: Determined the number of positions each individual had held during their tenure, creating the NumberOfPositions feature.
- Tracking Changes in the Employee Files: By creating the NumberOfPositionsChanges feature, we could track how many changes in the employee occur, such as changes in contract type, education level, employment status, nationality, or other factors.
- Converting Dates to Numeric Values: Transformed dates into numerical values relative to the termination date.
- Target Feature Creation: Introduced the 'Terminated' feature derived from the employment status, serving as the target variable.
- Column Elimination: Removed columns related to comments, as they pertained specifically to employment status.
- Seniority Calculation: Computed seniority from the hire date, assuming 31st December 2023 for those not terminated.
- Data Filtering: Excluded employees with hire or termination dates in 2024.
- Age Calculation: Calculated age from birth dates, excluding entries with inconsistent or under 18 years of age.
- City and Zip Code Removal: Eliminated the 'City' column due to over 1300 unique values. The 'Zip Code' column underwent a similar process.
- Date-Related Columns Removal: Deleted columns like 'Employment Status: Date', 'Job Information: Date', and 'Compensation: Date' due to constant updates in the HRIS platform without meaningful changes.

- Start Date Deletion: Removed 'Start Date' due to inconsistencies and misalignment with 'Hire Date'.
- Employment Status Refinement: Created a new column in employment status, eliminating values for employees with a single position who had already terminated.
- Nationality Column Deletion: Removed the 'Nationality' column due to 120 unique values, not standardized, meaning that some were country abbreviations, others country short names, and other US states abbreviations.
- Location Standardization: Derived 'Location Corrected', 'Office_FC', and 'OfficeCountry' columns using an auxiliary table.
- Department Simplification: Streamlined the 'Department' column by matching validated department values from an auxiliary table.
- Employment Type Extraction: Extracted employment type information using an auxiliary table.
- Job Title Deletion: Removed 'Job Title' due to over seven hundred unique values.
- Job Level Category Elimination: Excluded 'Job Level Category' due to inconsistency in categories across employees.
- Binary Transformation: Converted 'Yes' or 'No' columns to 1s and 0s.
- Working Hours Extraction: Extracted working hours from employment type, assuming 20 hours for part-timers, hourly, and work-study, and forty for contractors and full-time employees.
- Monthly Pay Rate Calculation: Calculated monthly pay rate based on working hours, pay rate, and currency exchange rates.
- Pay Type Correction: Corrected 'Pay Type' to hourly for pay rates below 50 and labeled as monthly or salary.
- Inconsistency Handling: Removed entries with values under seven hundred for monthly contractors and salary at 40 hours.
- Annual Program Normalization: Normalized 'Annual Program' values using an auxiliary table.
- Equity Feature Creation: Introduced a Boolean feature indicating the presence or absence of equity.
- Vesting Start Date Elimination: Removed the 'Vesting Start Date' due to limited availability among employees.
- Equity Amount Imputation: Imputed zero for null equity amounts.
- Column Deletions: Deleted 'Contract Type', 'Length', and 'Veteran Status' due to over 80% missing data.
- Last Team Lead Meeting Duration Calculation: Extracted the number of years since the last team lead meeting, assuming 31st December 2023 for active employees and termination dates for those who left.
- Data Exclusion: Disregarded employees starting after the end of 2023.
- Inconsistency Checking:

- Negative or null seniority values.
- Null seniority where termination date is absent.
- Ages above one hundred years, with a minimum acceptable age set at 75.
- Ages below 18 were deleted.
- Negative years since the last team lead meeting were removed.
- Employees with negative seniority or below 30 days were excluded for lack of coherence and insignificance in the predictive model.
- This meticulous data cleaning process ensures a robust foundation for subsequent analyses in the development of a predictive turnover model.

The final dataset offers a longitudinal snapshot of the workforce, encompassing both the tenure of active employees and the departure of those no longer with the company. With a total of 27 columns and 8.213 employees recorded, the dataset provides a robust foundation for our analysis, with 1,297 individuals with an active status and 6.916 having exited the organization within the observed period.

These steps were crucial for several reasons. Firstly, it facilitated a more uniform analysis, ensuring our models could interpret and learn from these variables effectively. Secondly, it allowed us to introduce standardization and comparability across the dataset, which was essential for accurately identifying trends and correlations.

Furthermore, in the pursuit of a dataset that truly reflected the multifaceted nature of employee dynamics, we also created new groupings. These were not arbitrary categorizations but were instead informed by a deep understanding of the dataset's initial structure and the specific needs of our research. By reclassifying and aggregating certain variables, we not only refined the dataset's granularity but also enhanced its analytical value, paving the way for more insightful and impactful findings.

The 27 initial columns are as follows in the Table 3.

Table 3 - List of Initial Features

Field	Description
Employee#	Unique identifier assigned to each employee in the dataset.
Termination	Date when the employee left the company, indicating the end of the employment period.
Seniority	The duration of the employee's tenure within the company, calculated in days.

NumberOfPositionsChanges	The total number of changes in position, company, contract type, country, location, or any other changes in the employee file during their time at the company.
NumberOfPositions	The total number of changes in job title during the employee time in the company
Age	The age of the employee at the time of data analysis or termination, excluding those below 18 years.
Gender	The gender of the employee, typically categorized as male, female, or other.
MaritalStatus	The marital status of the employee, indicating whether they are single, married, divorced, etc.
HighestEducationLevel	The highest level of education attained by the employee, such as high school, bachelor's, master's, etc.
LivingCountry	The country where the employee resides.
EmploymentStatus	The current status of employment, indicating whether the employee is active or terminated.
EmploymentType	The type of employment, specifying if the employee is full-time, part-time, hourly, contractor, or a work-student.
Location	The specific location where the employee works.
Office_FC	The office code or identifier associated with the employee's workplace.
OfficeCountry	The country where the employee's office is located.
Division	The division or segment within the company to which the employee belongs.
Department	The department or functional area where the employee is assigned.
JobLevel	The hierarchical level or rank of the employee within the organization.
TeamLead	The name or identifier of the employee's team lead.
WeeklyWorkingHours	The number of hours the employee is scheduled to work on a weekly basis.
PaySchedule	The frequency at which the employee receives payment, such as monthly or hourly.
PayType	The method by which the employee is compensated, whether salary, hourly wage, or other.
MonthlyRemuneration	The amount of remuneration the employee receives on a monthly basis.
NumberSalaryIncreases	The total number of salaries increases the employee experienced during their tenure.
AnnualEquityProgram	A binary indicator denoting whether the employee is enrolled in an annual equity program (Yes/No).

EquityPlan	The specific equity plan in which the employee is participating, if applicable.
TotalEquity	The total value or quantity of equity granted to the employee.
YearsLastTLMeeting	The number of years since the employee's last 1:1 meeting with their team lead, calculated until Dec 31, 2023, for active employees.

This dataset was rigorously anonymized in alignment with the General Data Protection Regulation (GDPR), which is pivotal for our research's integrity and ethical consideration. The company's endorsement of this study signifies a mutual recognition of the value and potential insights to be obtained from this analysis.

Central to our investigation are the features "Terminations" and "Seniority," selected for their relevance to our dual objectives of classification and regression. These features serve as focal points for our predictive modelling and embody the intricate dynamics of employee turnover and tenure within the sector. The dataset's composition reveals a significant imbalance between active and terminated employees, posing unique challenges and opportunities for deepening our understanding of turnover phenomena. Such disparities necessitate tailored analytical approaches, underscoring workforce dynamics' complexity and nuanced nature in the Retail & Food Industry.

This expanded dataset description offers a more detailed view of the scope, structure, and strategic importance of the data underpinning our study. Through this, we aim to unravel the multifaceted aspects of employee turnover and tenure, contributing valuable insights to the domain of human resources management and organizational studies.

In the next chapter, we dive into the data pre-processing phase of our methodologies to ensure the final data is ready for the model assessment. All the transformations mentioned below were done using Python in a Google Colab notebook, acknowledging that some of the points discussed below were somehow assured with the initial dataset preparation.

3.1.2. Target Variable

A pivotal part of our preprocessing involved the strategic extraction and transformation of critical features, most notably "Termination Date" and "Hire Date," from which we derived two essential variables: "Termination" and "Seniority", our future target features for the predictive model. The former, a Boolean indicator, was engineered to signify an employee's current status with the company—either active or terminated, and produced using the termination date, where if an employee had a considered termination date in the last position in the company, it was considered terminated, otherwise, active. The latter, a metric of years, quantitatively captured the duration of an employee's tenure, calculated by subtracting the last recognized date in the company (31/12/2023 for active employees) with the start date,

offering a nuanced lens through which to examine patterns of longevity and departure within the organization. Beyond these transformations, we embarked on a comprehensive normalization effort, targeting assessing the correct data types in our data frame.

3.1.3. Data Types

The data preprocessing phase involved meticulously refining data types, and categorizing features into metric, categorical, and Boolean subgroups, denoted as metric_features, categorical_features, and boolean_features, respectively. The initial data types were assigned appropriately, obviating the need for further transformations. However, the first step of cleaning and transformations in Excel already addressed essential corrections, particularly in converting dates to numerical representations, including calculating metric features. Throughout the dataset transformation process, pivotal adjustments were implemented to manage changes in data types. Specifically, the "Converting Dates to Numeric Values" step ensured a standardized format for temporal analysis by transforming date values into numerical representations relative to the termination date. Additionally, introducing the 'Terminated' feature in the "Target Feature Creation" step resulted in creating a binary variable, influencing the data type of this column. Furthermore, the "Binary Transformation" step converted 'Yes' or 'No' columns to numerical values (1s and 0s), impacting data types and enhancing compatibility with machine learning algorithms. This meticulous consideration of data types and the strategic conversion process contributed significantly to the overall robustness and coherence of the predictive turnover model, generating actionable features that enriched the dataset's analytical depth.

3.1.4. Descriptive Analysis

Starting the results chapter of our study, we first delve into some basic descriptive statistics of the data. This step lays the groundwork for our analysis and helps clarify the results that will follow, such as the performance of various models and the importance of different features. The dataset includes 8,213 employees, categorized as either 'Active' or 'Terminated.'. In the Figure 2, we detail the distribution of these classes.

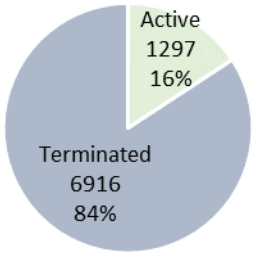


Figure 2 - Employee Distribution by Class

We note the initial class imbalance, which we addressed in detail in the methodology chapter using the 'tomekSMOTE' technique. This was an important step to prevent the models from being biased toward the more prevalent class.

Moving on, we have pinpointed vital features instrumental in understanding employee turnover. These features were chosen based on domain-specific knowledge and an appreciation of how certain factors may sway an employee's decision to stay with or leave the company. The Table 4 below highlights the average values for these features by class.

Table 4 - Average values by Class

Class	Seniority	Number of Positions	Number of Positions Changes	Number of Salary Increases	Age	Weekly Working Hours	Monthly Remuneration	Total Equity
Active	2.30	2.12	2.13	2.29	36.05	32.69	3264.27	3360.18
Terminated	0.89	1.34	2.47	0.81	31.36	27.80	2105.41	749.99

When we look at these averages, we observe some apparent differences between active and terminated employees. Active employees tend to have higher seniority, receive more salary increases, and work more hours weekly. They also have a greater monthly remuneration, number of positions and total equity, which aligns with the notion that financial stability and rewards are linked to employee retention. On the other hand, on average, terminated employees have had more position changes, representing the number of changes they suffered in the employment file. This finding is intriguing as it contradicts the idea that varied experience within the company promotes retention. It suggests that other factors might be at play, such as job satisfaction, career progression, or the nature of their roles, which could influence turnover. The contrasts highlighted here will be explored in depth as we progress, ensuring we fully understand their implications for employee retention and turnover. The descriptive statistics delineate a compelling narrative about the workforce dynamics at play. Averaging seniority of 2.30 years, active employees appear to exhibit a longer tenure compared to their terminated counterparts, who average 0.89 years, hinting at a possible trend where longevity may correlate with continued employment.

Interestingly, the average number of positions is higher for terminated employees, which may suggest that frequent internal transfers or promotions do not necessarily lead to higher retention and could, in fact, be a precursor to turnover. This is a critical insight, as it challenges the conventional wisdom that internal mobility is always a positive retention factor. The

number of Salary Increases for active employees is notably higher on average, suggesting that regular recognition in the form of pay raises may play a pivotal role in an employee's decision to remain with the company.

Additionally, age seems to be a differentiating factor, with active employees being older on average. This may indicate a workforce that values experience and, perhaps, a company culture that supports career longevity. Weekly working hours are also higher for active employees, potentially reflecting a more significant commitment or a more demanding workload for those who stay. In stark contrast, the financial indicators — monthly remuneration and total equity — are significantly lower for terminated employees, signaling that financial incentives could be a decisive factor in retention strategies.

The data presented paints a picture of a workforce where tenure, financial recognition, and possibly workload is intertwined with employee retention. However, the unexpected pattern of higher positions among those who have left the company warrants a deeper investigation. This could reveal structural issues within the job roles or point to the necessity for more nuanced career progression paths. Each statistic opens a new avenue for exploration and serves as a building block towards a more thorough understanding of employee turnover.

3.1.5. Exploratory Data Analysis (EDA)

With our approach to data types firmly established, we seamlessly transitioned into the exploratory data analysis (EDA) phase, a pivotal stage in uncovering insights and patterns within our dataset. During EDA, we delved into statistical and visual methods to better understand the data's underlying structure. This involved scrutinizing distributions, identifying potential outliers, and evaluating the balance of our target variable, 'Terminated.' The analysis extended to correlation assessments, allowing us to discern relationships between various features and uncover potential predictors of employee turnover. We gained valuable insights into the data's characteristics through visualization techniques such as histograms, scatter plots, and correlation matrices. This phase facilitated the cleaning of outliers and addressing imbalances and informed subsequent steps in feature engineering and model selection. The exploratory data analysis, thus, played a crucial role in shaping the trajectory of our predictive turnover model by providing a comprehensive understanding of the dataset's nuances.

3.1.5.1. Univariate Analysis

We started our analysis by approaching a univariate analysis of all the data types in the data frame. We used histogram plotting to make an initial assessment of the different data.

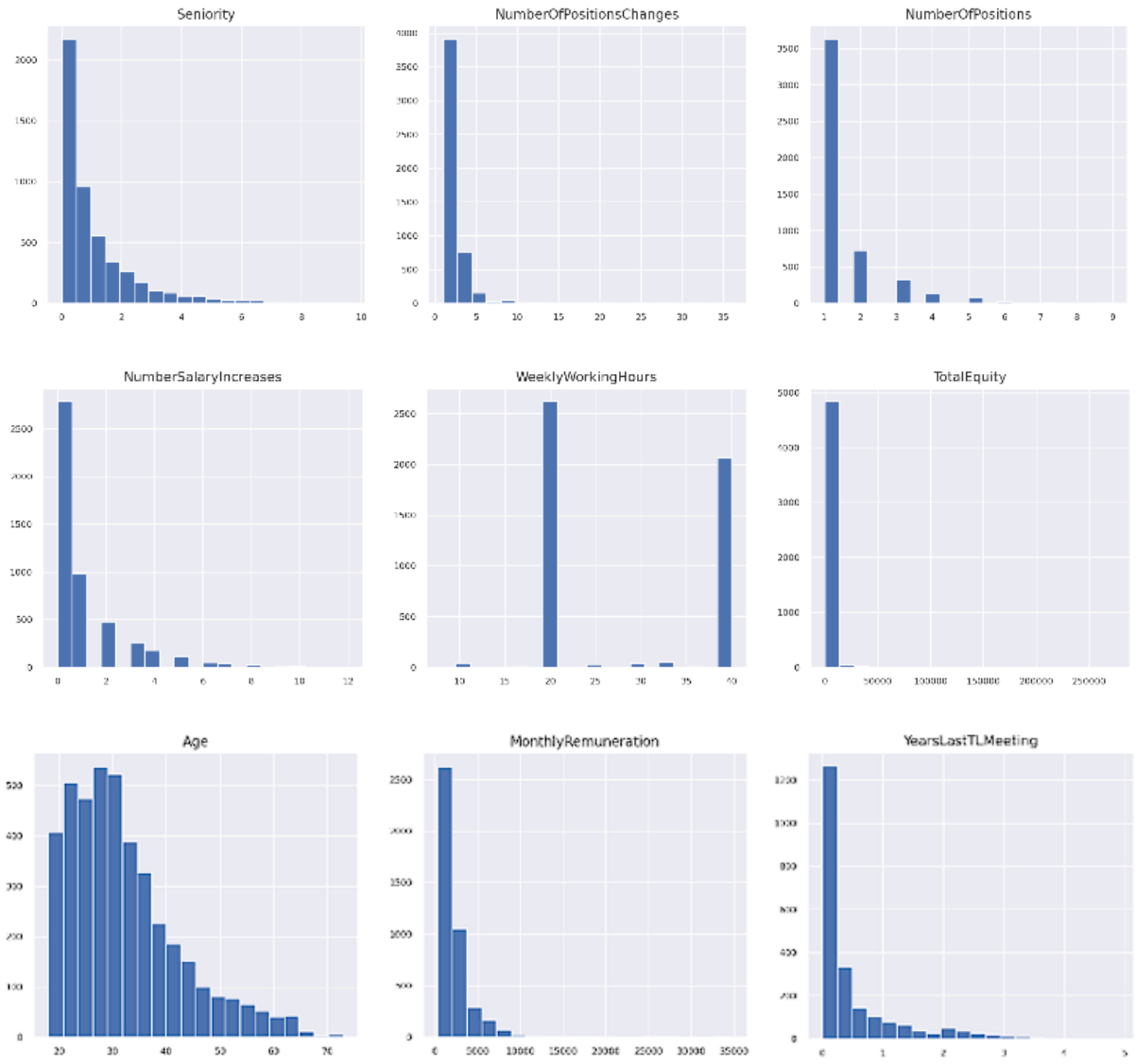
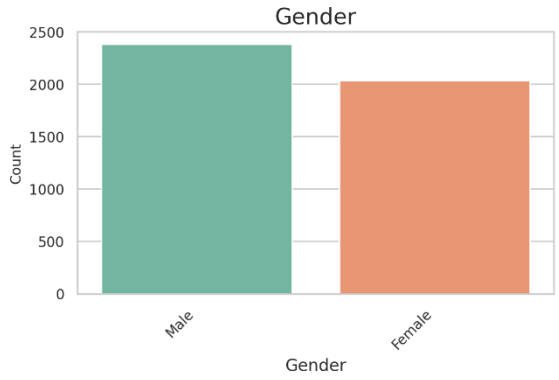
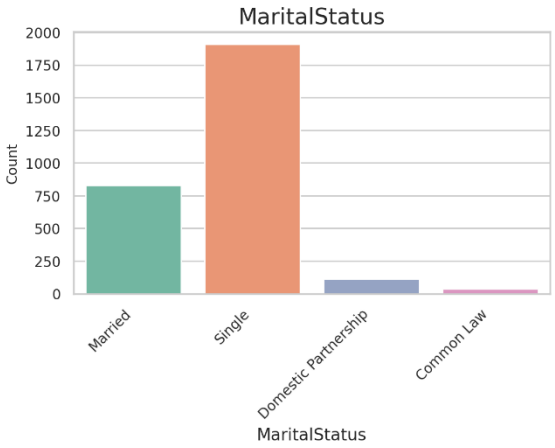
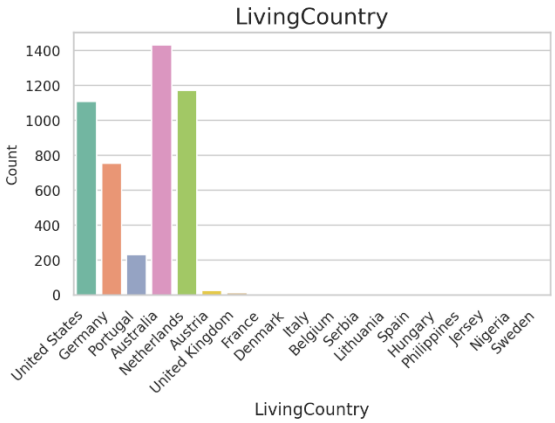


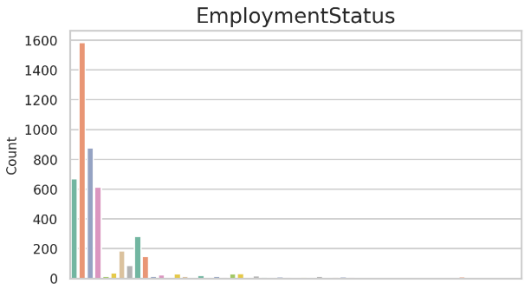
Figure 3 - Metric Features Histogram

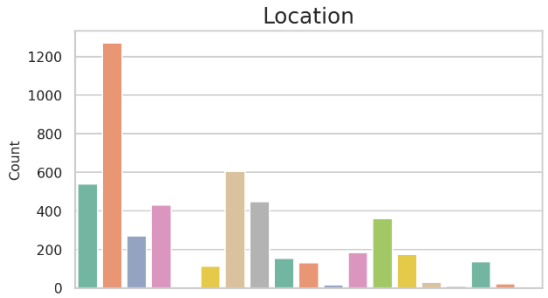
We can extrapolate from the Figure 3 three major points from these histograms. That there are no negative values, that the data might have some outliers, that a deeper analysis into outliers is required, and that most of the metric features are skewed. The data skewness will be approached further in the chapter with the required transformations described.

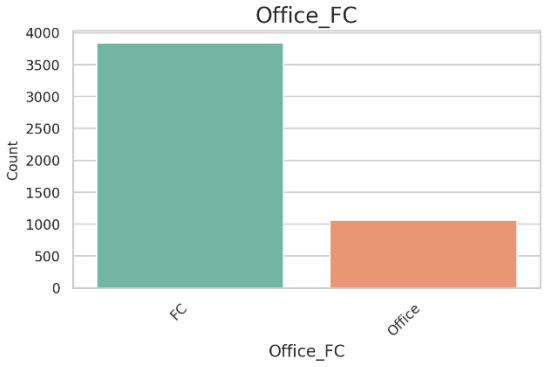

Gender
Count
2500
2000
1500
1000
500
0
Male
Female
Gender

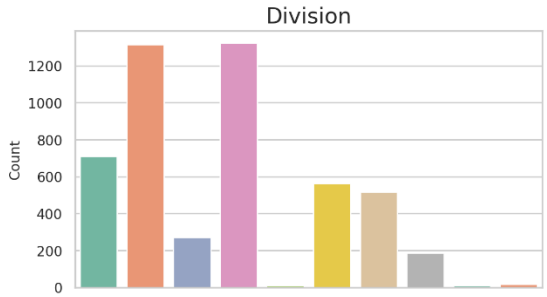

MaritalStatus
Count
2000
1750
1500
1250
1000
750
500
250
0
Married
Single
Domestic Partnership
Common Law
MaritalStatus

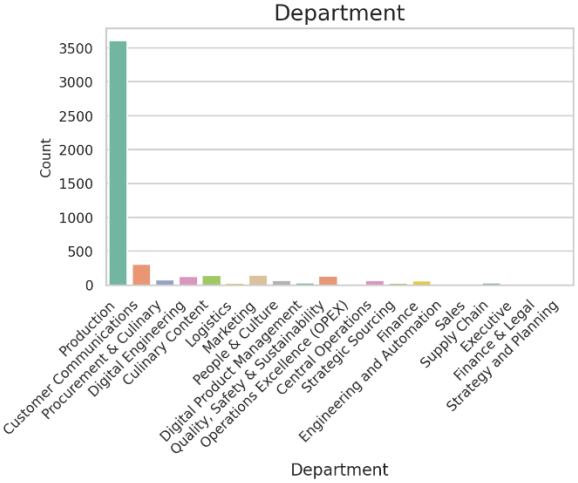

LivingCountry
Count
1400
1200
1000
800
600
400
200
0
United States
Germany
Portugal
Australia
Netherlands
Austria
United Kingdom
France
Denmark
Italy
Belgium
Serbia
Lithuania
Spain
Hungary
Philippines
Jersey
Nigeria
Sweden
LivingCountry


EmploymentStatus
Count
1600
1400
1200
1000
800
600
400
200
0


Location
Count
1200
1000
800
600
400
200
0


Office_FC
Count
4000
3500
3000
2500
2000
1500
1000
500
0
FC
Office
Office_FC


Division
Count
1200
1000
800
600
400
200
0


Department
Count
3500
3000
2500
2000
1500
1000
500
0
Production
Customer Communications
Procurement & Culinary
Digital Engineering
Culinary Content
Logistics
Marketing
People & Culture
Digital Product Management
Quality, Safety & Sustainability
Operations Excellence (OPEX)
Central Operations
Strategic Sourcing
Finance
Engineering and Automation
Sales
Supply Chain
Executive
Finance & Legal
Strategy and Planning
Department

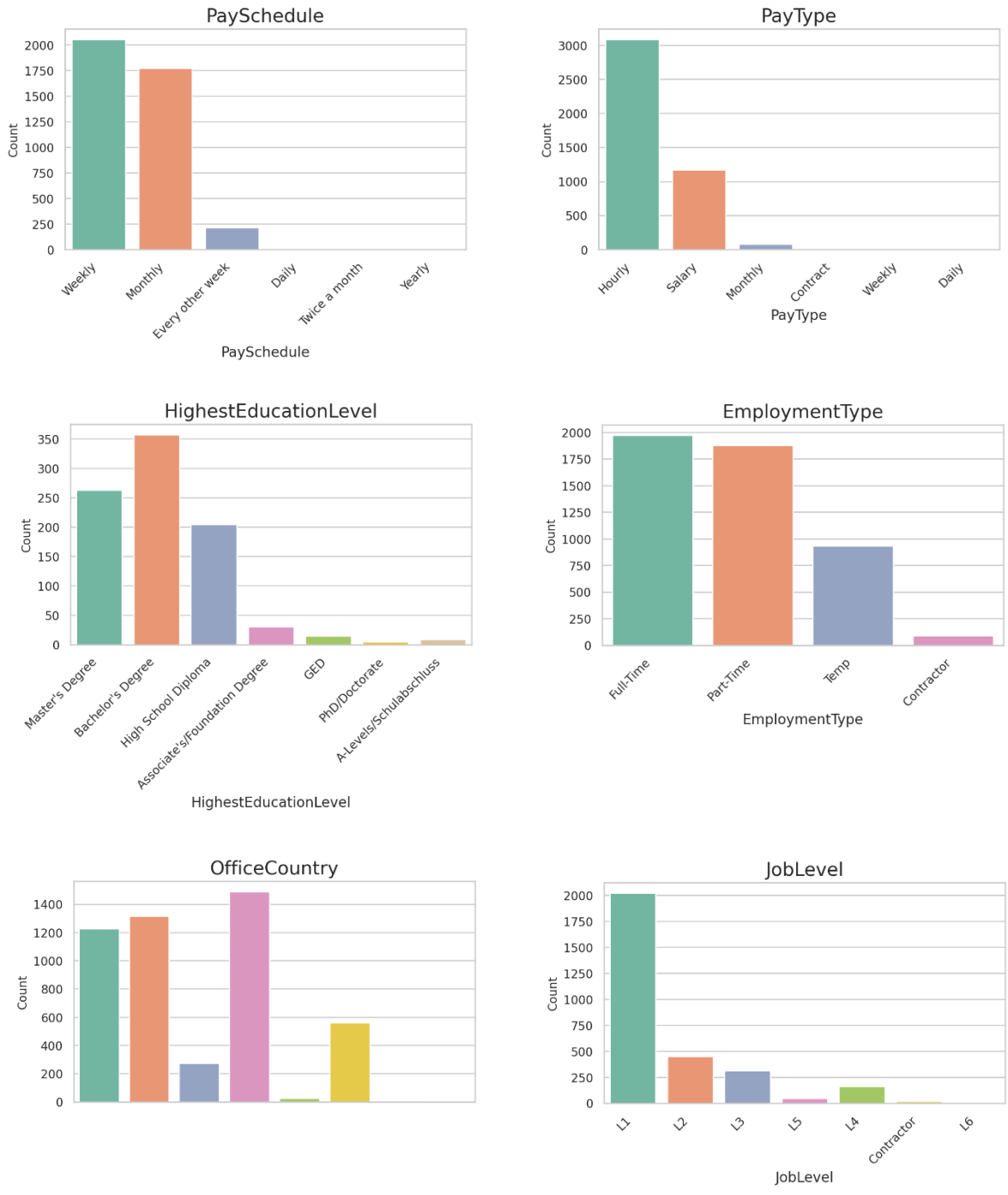


Figure 4 - Categorical Features Histograms

The “Division”, “EmploymentStatus”, “Location” and “OfficeCountry” charts do not display the x axis information as they contain sensible information about the company that cannot be presented.

By analyzing the categorical histogram plot, represented in Figure 4, we can understand the number of unique variables that one single feature can have.

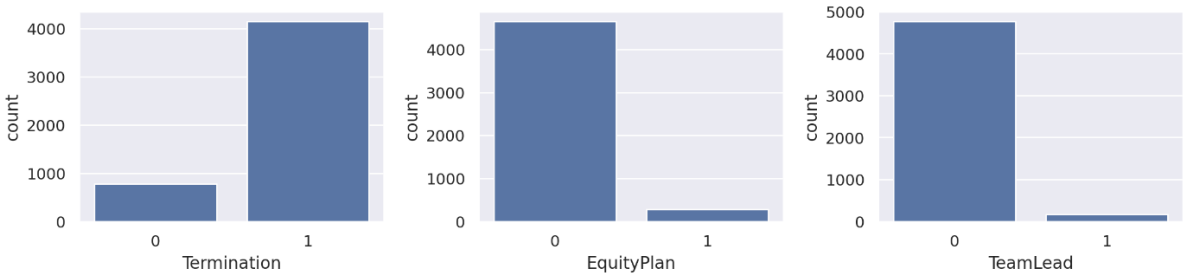


Figure 5 - Boolean Features Histograms

We can understand the imbalance on termination, which is the target feature, from the Boolean features on Figure 5. On the other two features, we can understand an imbalance, but it needs to be addressed together with the target feature to understand if a correlated imbalance exists inside the feature that requires a transformation of the features.

3.1.5.2. Bivariate Analysis – Correlation and Dependency

Bivariate Analysis focuses on analyzing two features together. This component was separated into two sections. The first is the correlation between features from the same datatype. Second, the correlation with the target feature. This approach was taken to all the different data types present in the dataset to follow an organized analysis.

Correlation analysis between metric features was performed using the Pearson Correlation Coefficient, with a threshold above |0.8|. The image below is a heatmap of correlation providing insights into relationships among numerical features and potential multicollinearity.

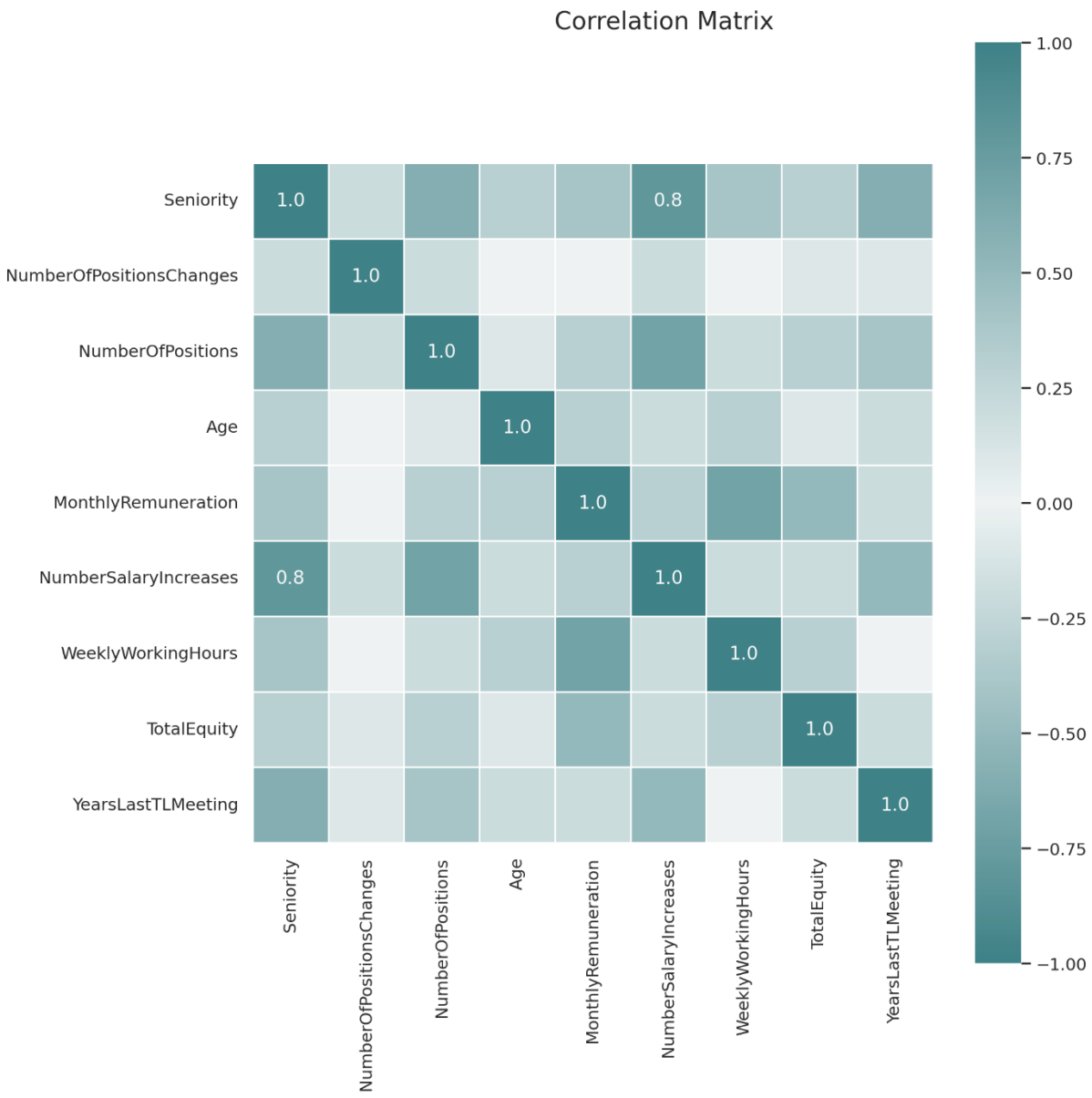


Figure 6 - Correlation between Metric features.

From Figure 6, we can see some variables with high correlation, with dark colouration, but not reaching the threshold defined of |0.8|, to be considerably high to be removed, unless one feature, the Number of Salary Increases. This feature will not be removed as it is considered to add more value to modelling rather than to reduce it since the only high relation is with seniority, an important metric. All the metric features were maintained during this step.

Because termination assumes a Boolean type, we could test the direct correlation from the metric features with the target, where the outcome was the respective in the Table 5 below.

Table 5 - Metric features correlation with Target feature

Feature	Correlation
Seniority	-0.389225
NumberOfPositions	0.27415
Age	-0.168638
MonthlyRemuneration	-0.247135
NumberSalaryIncreases	-0.175622
WeeklyWorkingHours	-0.177738
TotalEquity	-0.136126
YearsLastTLMeeting	-0.388274

The correlation is not imposed in one unique feature, but we can see that all contribute similarly, with some variations, to the representation of the outcome variable, through the Table 5. After this, an ANOVA test, with a p-value of 0.05, was executed to understand if there existed a specific feature less likely to be significant with the outcome. No feature was identified in this test.

For the correlation between Boolean features and correlation with the target feature, it was used the Point Biserial analysis, where no Boolean features were identified as having a high correlation value, both between them and with the target feature. No Boolean columns were dropped from the data frame.

Regarding the categorical features correlation, the chi squared test was executed for both scenarios as shared before, presented in Tables 6, 7, 8, 9 and 10.

Table 6 - Chi Squared test between Categorical features – part 1

	AnnualEquityProgram	Department	Division	EmploymentStatus
AnnualEquityProgram		2.16E-252	5.75E-60	5.23E-157
Department	2.16E-252		0	0
Division	5.75E-60	0		0
EmploymentStatus	5.23E-157	0	0	
EmploymentType	3.00E-66	1.75E-289	0	0
Gender	1.60E-02	5.28E-39	3.18E-35	1.01E-27
HighestEducationLevel	8.01E-09	1.59E-16	1.74E-39	3.42E-31

JobLevel	3.17E-286	0	0	0
LivingCountry	2.02E-24	0	0	0
Location	1.04E-160	0	0	0
MaritalStatus	7.22E-05	3.43E-03	1.26E-16	NaN
OfficeCountry	2.31E-47	0	0	0
Office_FC	4.09E-140	0	0	0
PaySchedule	1.11E-97	1.50E-295	0	0
PayType	7.30E-155	0	0	0

Table 7 - Chi Squared test between Categorical features – part 2

	EmploymentType	Gender	HighestEducationLevel	JobLevel
AnnualEquityProgram	3.00E-66	1.60E-02	8.01E-09	3.17E-286
Department	1.75E-289	5.28E-39	1.59E-16	0
Division	0	3.18E-35	1.74E-39	0
EmploymentStatus	0	1.01E-27	3.42E-31	0
EmploymentType		9.50E-36	7.42E-04	0
Gender	9.50E-36		NaN	4.88E-05
HighestEducationLevel	7.42E-04	NaN		2.89E-35
JobLevel	0	4.88E-05	2.89E-35	
LivingCountry	0	3.61E-31	9.41E-24	0
Location	0	7.82E-46	3.85E-46	0
MaritalStatus	3.77E-03	2.76E-04	2.08E-04	7.14E-08
OfficeCountry	0	8.72E-37	2.85E-45	0
Office_FC	2.40E-221	1.36E-14	2.54E-33	0
PaySchedule	1.35E-299	6.84E-07	2.13E-39	1.17E-252
PayType	0	6.90E-06	9.70E-27	0

Table 8 - Chi Squared test between Categorical features – part 3

	LivingCountry	Location	MaritalStatus	OfficeCountry
AnnualEquityProgram	2.02E-24	1.04E-160	7.22E-05	2.31E-47
Department	0	0	3.43E-03	0
Division	0	0	1.26E-16	0
EmploymentStatus	0	0	NaN	0
EmploymentType	0	0	3.77E-03	0
Gender	3.61E-31	7.82E-46	2.76E-04	8.72E-37
HighestEducationLevel	9.41E-24	3.85E-46	2.08E-04	2.85E-45
JobLevel	0	0	7.14E-08	0
LivingCountry		0	5.43E-12	0
Location	0		5.83E-25	0
MaritalStatus	5.43E-12	5.83E-25		6.57E-19
OfficeCountry	0	0	6.57E-19	
Office_FC	0	0	NaN	0
PaySchedule	0	0	NaN	0
PayType	1.44E-296	0	NaN	0

Table 9 - Chi Squared test between Categorical features – part 4

	Office_FC	PaySchedule	PayType
AnnualEquityProgram	4.09E-140	1.11E-97	7.30E-155
Department	0	1.50E-295	0
Division	0	0	0
EmploymentStatus	0	0	0
EmploymentType	2.40E-221	1.35E-299	0
Gender	1.36E-14	6.84E-07	6.90E-06
HighestEducationLevel	2.54E-33	2.13E-39	9.70E-27

JobLevel	0	1.17E-252	0
LivingCountry	0	0	1.44E-296
Location	0	0	0
MaritalStatus	NaN	NaN	NaN
OfficeCountry	0	0	0
Office_FC		4.93E-236	0
PaySchedule	4.93E-236		0
PayType	0	0	

Table 10 - Chi Squared test between Categorical features and Target features

Feature	P-value
Gender	5.37E-09
MaritalStatus	1.37E-12
HighestEducationLevel	6.17E-01
LivingCountry	1.03E-52
EmploymentStatus	4.57E-55
EmploymentType	1.70E-51
Location	3.06E-81
Office_FC	2.65E-22
OfficeCountry	2.33E-49
Division	1.31E-95
Department	4.96E-30
JobLevel	2.82E-22
PaySchedule	6.36E-02
PayType	1.39E-30
AnnualEquityProgram	3.83E-23

3.1.5.3. Preliminary Findings

EDA emphasized the importance of certain features in predicting turnover. It also highlighted the need for feature transformations and the potential value of interaction terms between Age and Seniority.

This pragmatic EDA offered a comprehensive view of the dataset, its feature characteristics, and their potential impact on turnover. It served as a foundation for subsequent feature engineering and modelling phases.

3.1.5.4. Addressing Outliers

Within our dataset, some metric features exhibited a significant presence of outliers, Figure 7 and 8, posing potential challenges to model accuracy and training efficacy. We employed a limit imputation technique to mitigate these effects, strategically setting thresholds for various features to curtail outlier influence. This methodological adjustment was essential for maintaining the integrity and reliability of our predictive models, ensuring that extreme values did not skew the data's central tendencies. This approach optimized the dataset for analysis and aligned with best practices in data preprocessing, enhancing the overall robustness of our modelling efforts.

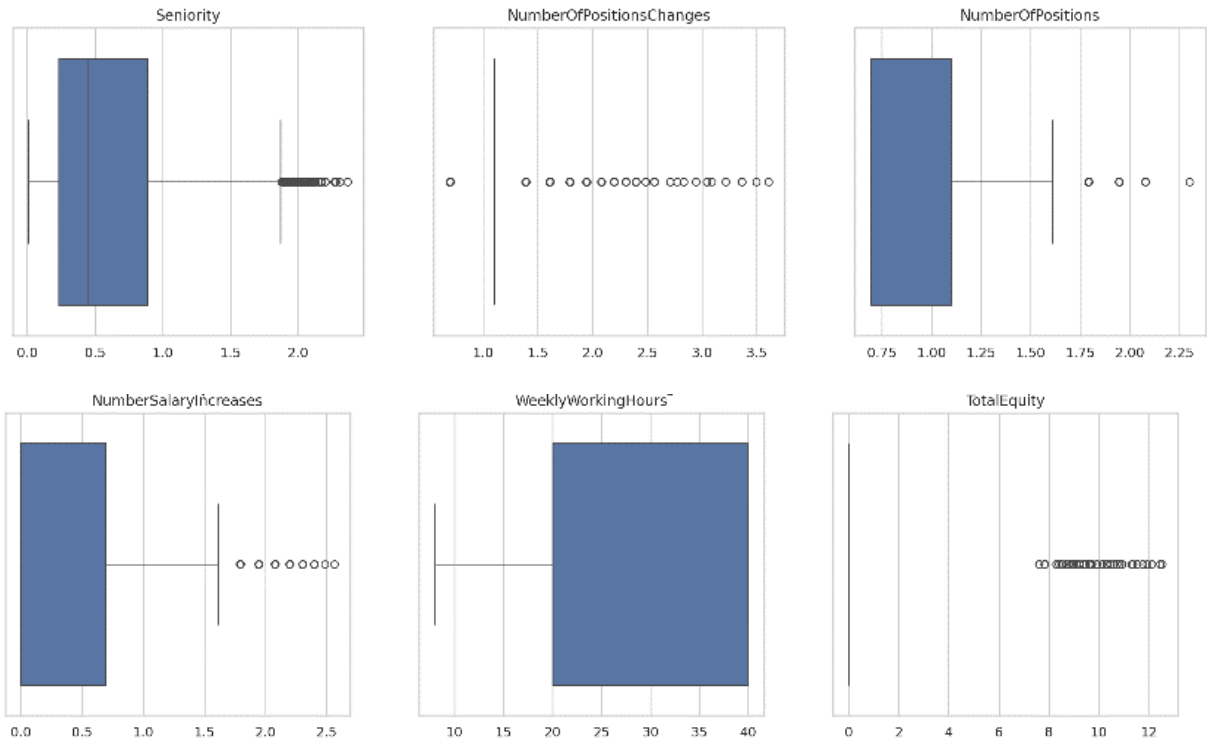


Figure 7 - Metric features box plot for outlier detection – part 1

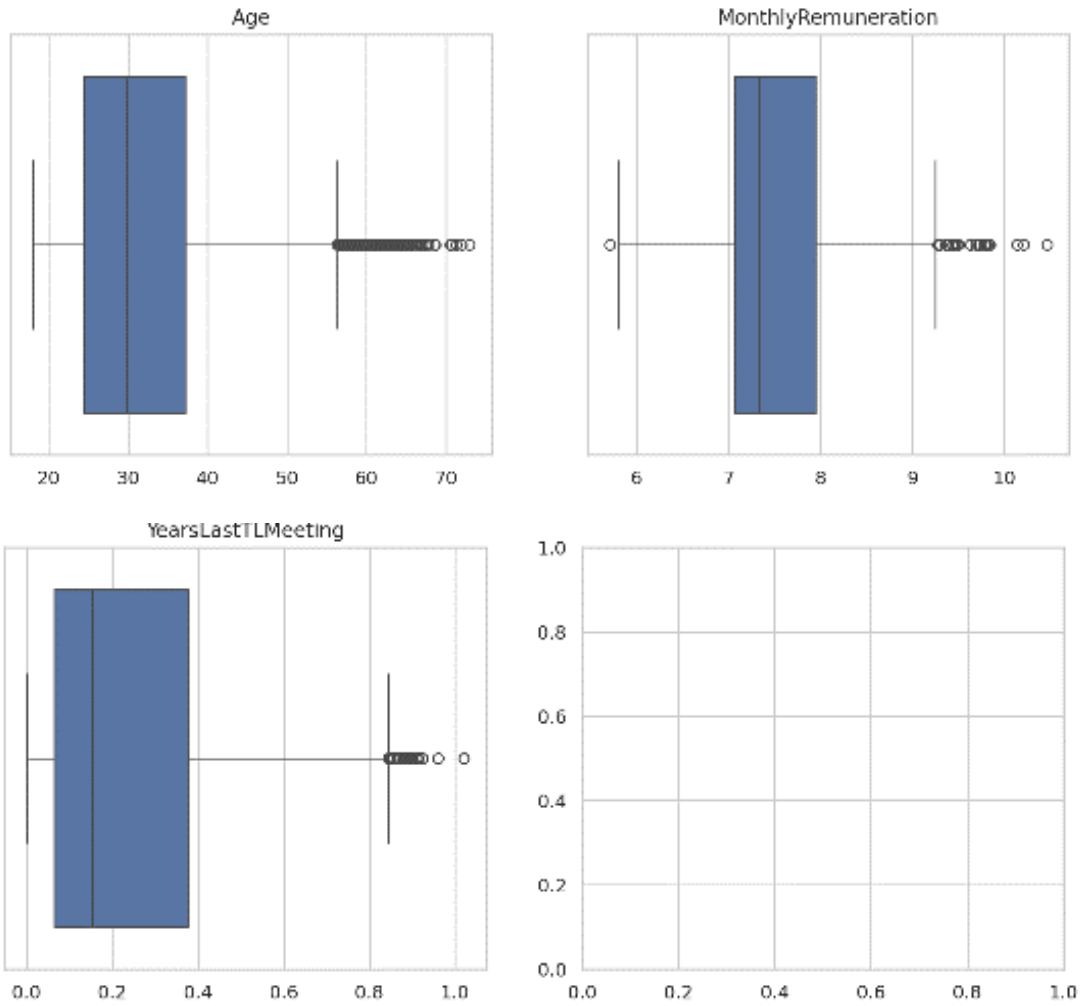


Figure 8 - Metric features box plot for outlier detection – part 2

3.1.5.5. Concluding EDA

The completion of the EDA phase marks a crucial milestone in preparing our dataset for predictive modelling. After meticulous testing for outliers and confirming the absence of duplicate values, we focused on addressing missing data, a common challenge in real-world datasets. While no duplicates required attention, missing values prompted thoughtful transformations.

We opted for the mode imputation method for categorical and Boolean features, strategically filling in missing values with the mode of each respective feature. This approach ensures that the imputed values align with the prevailing trends in these categorical aspects. Conversely, the K-Nearest Neighbors (KNN) imputation technique was applied for metric features, leveraging the collective information from 30 neighboring data points for accurate and contextually relevant imputation.

As we conclude the EDA chapter, it is essential to underscore that these tailored imputation strategies addressed missing values and set the stage for subsequent analyses and model development. The refined dataset is now a more complete and reliable foundation, ready to be leveraged for feature engineering steps.

3.2. DATA PREPARATION

In the following section, we present further steps, such as data cleaning and other transformations, apart from those explored before. Some of the transformations were already described before the specific analysis was presented.

3.2.1. Feature Engineering

3.2.1.1. Normalization

We employed a Min-Max scaling approach to standardize and prepare the metric features for advanced modelling; leveraging the MinMaxScaler from the scikit-learn library, the metric features in the dataset were individually scaled to a uniform range. The original metric features were then removed from the original dataset to avoid redundancy. Finally, the scaled features were seamlessly merged back into the dataset, ensuring that the standardized metrics now coexist harmoniously with other categorical and Boolean features. This scaling process enhances the comparability and interpretability of metric features, contributing to the overall optimization of the dataset for subsequent modelling endeavors.

3.2.1.2. Encoding

In the pursuit of handling categorical features effectively for modelling purposes, a robust encoding strategy was employed. Utilizing the pd.get_dummies function, the original dataset, traindf1_dtypes_outlier_corr_incon, underwent categorical feature encoding. Specifically, dummy variables were created for each categorical feature, expanding the dataset horizontally. The drop_first=True parameter was utilized to prevent multicollinearity by excluding the first category in each feature.

To maintain consistency and coherence, working with the dataset post-encoding is crucial. The categorical features in their encoded form were identified by comparing the columns before and after encoding. The encoded columns and original columns lists were generated to facilitate this comparison.

The resulting categorical features encoded list represents the refined set of categorical features, crucial for subsequent modeling steps. This comprehensive encoding approach

ensures that categorical variables are appropriately represented in a format compatible with machine learning algorithms, fostering the development of accurate and insightful models.

3.2.2. Class Imbalance

In predictive modeling, addressing class imbalance is imperative to ensure the model's ability to generalize across diverse outcomes. This chapter delves into the comprehensive strategy employed to manage class imbalances, presenting a multifaceted approach that encompasses initial exploration, synthetic oversampling, and a final assessment of the rebalanced dataset.

3.2.2.1. Initial Examination of Class Imbalances

To commence, a meticulous examination of class proportions within each feature was undertaken. The code systematically calculated the overall proportions for each category of the features, shedding light on the distribution of classes. Additionally, the proportions of the 'Termination' outcome concerning each feature category were analyzed, providing nuanced insights into the intricate relationships between these features and the target variable.

3.2.2.2. Synthetic Minority Over-sampling Technique and Tomek Links

Recognizing the need for rebalancing, the Synthetic Minority Over-sampling Technique (SMOTE) in conjunction with Tomek links (SMOTE Tomek) was employed. This method not only oversamples the minority class but also addresses potential issues arising from Tomek links, effectively enhancing the overall balance of the dataset. The implementation involved initializing SMOTE Tomek with carefully selected parameters, specifically setting k_neighbors to 4 for optimal performance.

3.2.2.3. Evaluation of Rebalanced Dataset

Following the application of SMOTE Tomek, an evaluation of the class distribution was conducted. The class distribution after the rebalancing process was analyzed, affirming the successful mitigation of class imbalances.

3.2.2.4. Results and Implications

The culmination of these steps underscores the commitment to creating a balanced and equitable dataset for subsequent model development. By systematically addressing class

imbalances, this chapter sets the stage for building robust, unbiased predictive models that can make accurate predictions across diverse outcomes.

The next chapter will delve into model development and evaluation specifics, leveraging the refined dataset to create a predictive turnover model.

3.2.3. Feature Selection and Dimensionality Reduction

Feature selection is a critical preprocessing step in constructing effective predictive models. It involves identifying and retaining key features that contribute most significantly to the model's predictive power, thereby enhancing interpretability and generalization capabilities. One prominent technique employed for this purpose is Recursive Feature Elimination (RFE).

3.2.3.1. Recursive Feature Elimination (RFE)

RFE is an iterative feature selection method designed to optimize model performance by removing less influential features systematically. It operates by fitting the model multiple times, each time eliminating the least significant feature until the desired number of features is reached. RFE leverages the model's inherent capacity to discern feature importance, facilitating the identification of a refined subset that maximizes predictive accuracy.

Recognizing the versatility of RFE, this study applied the method across three distinct models—Logistic Regression, Random Forest Classifier, and Neural Networks. Each model provided a unique perspective on feature importance and selection criteria. By incorporating RFE into multiple models, we aimed to glean comprehensive insights into the relative importance of features and assess their impact on predictive model performance. The decision to apply RFE to multiple models stems from the desire for a holistic evaluation of feature importance. Each model's inherent strengths and characteristics contribute to a nuanced understanding of which features are consistently crucial across diverse modeling approaches. This comprehensive evaluation not only enriches our understanding of the dataset but also paves the way for the development of robust and interpretable predictive models.

In conclusion, the adoption of Recursive Feature Elimination across Logistic Regression, Random Forest Classifier, and Neural Networks signifies a strategic approach to feature selection. RFE's iterative nature, coupled with its model-agnostic adaptability, ensures a nuanced examination of feature importance. This chapter lays the groundwork for subsequent model development and underscores the commitment to optimizing the feature space for improved predictive modeling outcomes.

This process of feature selection and dimensionality reduction was pivotal in refining the dataset, emphasizing the most pertinent predictors, and enhancing computational efficiency. By systematically evaluating and reducing the feature set, the model was poised to deliver more precise and interpretable predictions regarding employee turnover, devoid of embellishments.

3.2.4. Concluding Data Preparation

In conclusion, the data preprocessing stage was a testament to our methodological diligence and analytical foresight. By carefully setting the target variables and undergoing a rigorous data cleaning, normalization, and transformation process, we prepared a robust foundation for our predictive models. This preparation was instrumental in enabling us to address our research questions with the utmost precision, ensuring that our exploration of classification and regression problems related to employee turnover and tenure was grounded in a dataset of unparalleled quality and depth.

3.3. MODELLING

The features identified through Recursive Feature Elimination (RFE) using Logistic Regression, Random Forest Classifier, and Neural Networks have undergone a meticulous selection process, where less significant features were systematically eliminated. This strategic feature extraction phase sets the stage for the subsequent model selection and development. By applying the selected features to each respective model—Logistic Regression, Random Forest Classifier, and Neural Networks—we aim to discern the impact of these features on model performance. This comprehensive evaluation will provide insights into the relative importance of features across diverse modeling approaches, guiding us toward the identification of the most effective model for our predictive turnover analysis. The synergy between feature extraction and model evaluation is pivotal in ensuring that the chosen model not only accommodates the refined feature subset but also maximizes predictive accuracy and interpretability. The study explores a range of classification models to identify the most effective ones in predicting the desired outcomes.

3.3.1. Model Selection

Our approach to model selection is deliberate and purposeful, incorporating a comprehensive suite of classification models, each chosen for its unique strengths and relevance to the predictive task at hand. The models include K-Nearest Neighbors (KNN), Logistic Regression,

Decision Trees, Neural Networks (Multi-layer Perceptron), Random Forest Classifier and Gradient Boosting Classifier. This diverse set ensures a robust exploration of various algorithmic approaches, and we are considering these models as the more traditional ones that we will use.

This study employed advanced machine learning techniques using a neural network model built with the TensorFlow library, particularly leveraging its high-level Keras API. TensorFlow's Keras API is renowned for its comprehensive capabilities in constructing, training, and evaluating models, offering seamless integration of functionalities that are essential for deep learning workflows. The model architecture was constructed using the Sequential class, which facilitates the creation of models where each layer is sequentially connected without branching, each receiving input from one tensor and sending output to another. The layers include Dense layers, which are fully connected neural network layers, employed here to map inputs to outputs with specific activation functions. Dropout layers were strategically placed to randomly nullify a fraction of the input units at each update during the training phase to mitigate overfitting- a common issue in neural networks. The model configuration was finalized using the compile method, which sets the optimization algorithm (adam), loss function (binary_crossentropy), and metrics (accuracy). This configuration underscores the model's readiness to handle binary classification tasks effectively. Additionally, the data preprocessing steps involved converting all input data to float32 format, ensuring compatibility and optimal performance during training. Collectively, these elements underscore the advanced nature of our neural network model, distinguishing it from traditional statistical models by its ability to learn complex patterns through deep learning techniques, thus enhancing predictive accuracy and model robustness. This model will be represented as a TF (TensorFlow) model, as it uses the Keras package, only for sake of representing a different model.

Following this, hyperparameter tuning was performed to optimize the model configurations. This iterative process hones the models to improve performance. Finally, an ensemble model is created, combining the top three high-performing models, post-hyper parameterization, into a unified model that leverages the strengths of its components to deliver a sophisticated predictive tool, likely for the purpose of forecasting employee turnover. The hole process can be seen in Figure 9 below.

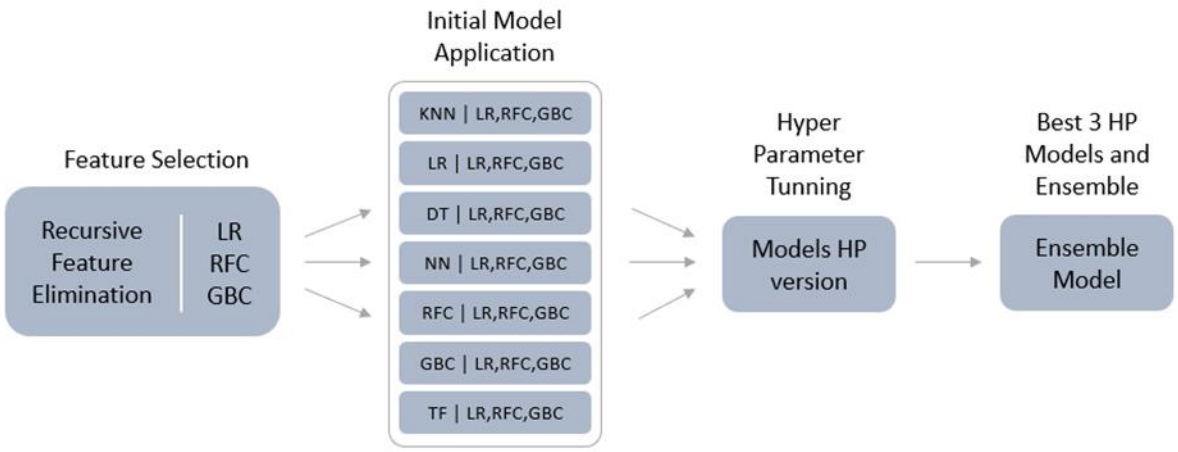


Figure 9 - Modelling Process Methodology Scheme

3.3.2. Hyperparameter Tuning

To enhance the predictive capabilities of our traditional considered models, we engage in a meticulous hyperparameter tuning process. This phase is crucial in extracting optimal performance from models demonstrating promising initial results. To optimize the parameters of the different models, a randomized search approach was employed using the RandomizedSearchCV estimator from the scikit-learn library. This technique involves a probabilistic search over the parameter space, where a fixed number of parameter settings is sampled from specified distributions. The model enhancement was done using a 5-fold cross-validation to evaluate the effectiveness of each parameter setting. This fine-tuning ensures that our models are optimized for the specific nuances of our dataset, contributing to their overall accuracy and predictive power.

3.3.3. Selection Methodology

Our model selection process is a meticulous and iterative journey to identify the most effective predictive models for employee turnover and seniority. The approach begins with applying each classification model to subsets generated through Recursive Feature Elimination (RFE), a process that systematically extracts the most informative features. This step allows us to evaluate the models' performance across different feature sets, ensuring adaptability to varying complexities within the dataset.

The subsequent analysis involves scrutinizing the models' outcomes and selecting the best-performing model based on a holistic assessment of scores, including but not limited to accuracy, precision, and recall. This initial selection phase serves as a precursor to

hyperparameter tuning, where we fine-tune the chosen models to optimize their performance.

Hyperparameter tuning is executed with precision, leveraging techniques such as RandomizedSearchCV to explore a range of parameter configurations. This thorough exploration ensures that the models are calibrated to the unique characteristics of the dataset, enhancing their predictive capabilities.

After this iterative process for each classification model, we identify the best-performing models that have undergone hyperparameter optimization. These models serve as the foundational components for our ensemble model. The ensemble is constructed by leveraging the strengths of the top three individual models, combining their predictions in a synergistic manner. This approach aims to capitalize on the diverse strengths of each model, fostering a more robust and resilient predictive model for employee turnover and seniority.

The ensemble model undergoes a final round of evaluation and, if necessary, further fine-tuning to ensure its efficacy. The culmination of this process results in a predictive framework that draws on the strengths of individual models while mitigating potential weaknesses, thereby enhancing the overall reliability and accuracy of our predictive model.

3.4. EVALUATION PROCESS

The evaluation process ensures a consistent and thorough assessment of each model's performance. We adhere to stringent evaluation criteria to enable a fair and robust performance comparison across the model's outcomes. Key steps involve model instantiation, comprehensive training on the dataset, prediction of outcomes on the validation, and subsequent metric-based comparisons. Evaluation metrics include the Confusion Matrix, Accuracy, ROC-AUC Score, Precision, Recall, and the ROC Curve. The culmination of these evaluations results in final predictions for the test set, providing a foundational basis for further in-depth analysis. To deepen our understanding of the final predictions and the model created, we will analyze the 5 most important features, their level of predictability, their permutation importance, how they relate to each other, and procure the reasonability based on people analytics understanding so the best advices can result in the outcome of this study, for the company and for the field of Human Resources Management.

3.5. CONCLUSION

This section used a comprehensive framework for predicting employee turnover and seniority, enabling data preprocessing, and ensuring the dataset's readiness for advanced analytics. The process encompassed the aggregation of diverse datasets, career progression tracking, and conversion of features. Target features were strategically created, the data types refined, and

irrelevant columns eliminated, laying the groundwork for exploration and the consideration given to maintain consistency and compatibility. The handling of inconsistent data, from outliers to correlation nuances, showcased a meticulous commitment to data integrity, as the utilization of SMOTE Tomek for balancing the target variable exemplified a nuanced approach to ensure model fairness and accuracy. Our methodology encompassed model selection, feature selection, and evaluation processes across various classification models. Through systematic hyperparameter tuning and model-specific methods, we established a robust foundation for our research. This chapter's structured approach ensures the reliability and rigor of our subsequent analyses, enabling us to gain valuable insights into the factors influencing employee turnover and seniority prediction, preparing the analysis on the different results generated by the strategy applied.

4. RESULTS AND DISCUSSION

In the following section, we will analyze the results obtained with the development and application of the different models to the dataset, comprehended from the different steps and methods described in the previous section, and the approach to answering the business questions defined in this work. Various models were applied to the dataset in an orderly manner. As explained before, first, we selected the features based on three different methods, then produced the base models and evaluated their performance; we selected the best and applied the respective hyper parameterization to enhance their predictability power, and in the end, we produced an ensemble model with the best two models already hyper parameterized. This methodology created a classification model that identifies which employees are more likely to leave the company based on the company characteristics and specific environments. We intend to present the findings of our research by developing a model that can have a good performance in predicting both classes, leavers and not leavers, and be a good tool for the company that provided the data, so in the future, can work on the critical points and with that reduce the turnover, and act proactively regarding their leavers.

We will follow a methodological approach to the produced results and findings as they represent the best tools to evaluate the model's predictability and applicability to the context in which it is desired.

4.1. RESULTS

4.1.1. Results Analysis

4.1.1.1. Models Performance

We commence the results and discussion section by analysing the overall performance of all employed models. The data originated from a real company and underwent various transformations until a final dataset suitable for model testing and performance evaluation was achieved. This process involved iterative trial and error, exploring different approaches to enhance results. Notably, the dataset suffered from class imbalance, with a predominance of leavers, potentially leading to overfitting of the validation and test sets. Consequently, the precision of class 0 (non-leavers) was carefully considered when assessing model performance to ensure the model performance was well determined and addressed based on the classification of both classes and not specifically the overall accuracy or precision of the model. By that motive, the consideration weight attributed to the precision of class 0 was higher than the remaining performance elements.

It is imperative to recall the developmental approach employed for the models. Initially, each model was constructed using different features identified through the feature selection

method. Subsequently, we determined the optimal subset of features yielding the best results and refined the model through hyper parameterization. Finally, we constructed an ensemble model comprising the top three models from the approaches, with hyper parameterization if indeed the models were improved.

In our study's Initial Model Application phase, we have methodically evaluated various predictive models using a suite of performance metrics. As demonstrated in Table 11, we have compared models including KNN, LR, DT, MLP, RFC, GBC, and Tensor, each subjected to different feature selection methods: Logistic Regression (LR), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC). The efficacy of each model was measured in terms of accuracy, precision (overall and for label 0), recall, F1 score, and ROC-AUC, alongside the confusion matrix to understand the distribution of true and false predictions. Below is a summary table with all the performance scores for the different models.

Table 11 - Initial Model Application Performance Summary

Model	Feature Selection Method	Accuracy	Precision	Precision Label 0	Recall	F1 score	ROC-AUC	Confusion Matrix
KNN	LR	0.873	0.947	0.585	0.898	0.922	0.900	[[193, 67], [137, 1205]]
KNN	RFC	0.860	0.947	0.551	0.882	0.914	0.899	[[194, 66], [158, 1184]]
KNN	GBC	0.857	0.949	0.543	0.876	0.911	0.893	[[197, 63], [166, 1176]]
LR	LR	0.875	0.951	0.588	0.896	0.923	0.909	[[198, 62], [139, 1203]]
LR	RFC	0.878	0.955	0.594	0.896	0.925	0.911	[[203, 57], [139, 1203]]
LR	GBC	0.880	0.957	0.597	0.896	0.926	0.912	[[206, 54], [139, 1203]]
DT	LR	0.932	0.960	0.789	0.959	0.959	0.876	[[206, 54], [55, 1287]]
DT	RFC	0.939	0.961	0.825	0.967	0.964	0.882	[[207, 53], [44, 1298]]
DT	GBC	0.945	0.967	0.833	0.968	0.967	0.897	[[215, 45], [43, 1299]]
MLP	LR	0.864	0.957	0.556	0.877	0.915	0.912	[[207, 53], [165, 1177]]
MLP	RFC	0.871	0.959	0.574	0.885	0.920	0.915	[[209, 51], [155, 1187]]
MLP	GBC	0.874	0.957	0.582	0.890	0.922	0.917	[[206, 54], [148, 1194]]
RFC	LR	0.951	0.966	0.866	0.975	0.971	0.981	[[214, 46], [33, 1309]]
RFC	RFC	0.950	0.972	0.841	0.969	0.970	0.983	[[222, 38], [42, 1300]]
RFC	GBC	0.948	0.970	0.837	0.968	0.969	0.982	[[220, 40], [43, 1299]]
GBC	LR	0.940	0.971	0.795	0.958	0.964	0.977	[[221, 39], [57, 1285]]
GBC	RFC	0.944	0.971	0.815	0.963	0.967	0.978	[[221, 39], [50, 1292]]
GBC	GBC	0.943	0.969	0.811	0.962	0.966	0.978	[[219, 41], [51, 1291]]
Tensor	LR	0.924	0.949	0.788	0.962	0.955	0.000	[[190, 70], [51, 1291]]
Tensor	RFC	0.932	0.946	0.842	0.974	0.960	0.000	[[186, 74], [35, 1307]]
Tensor	GBC	0.937	0.957	0.824	0.968	0.963	0.000	[[202, 58], [43, 1299]]

Upon careful analysis of the performance metrics from our initial model application phase, as depicted in the table, it is clear that the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) models demonstrate superior accuracy over other models, including Decision Trees (DT). When utilizing logistic regression feature selection method, the RFC model

achieves an outstanding accuracy of 0.951, with exceptional precision for label 0 at 0.866. Similarly, the GBC model also performs admirably, showcasing an accuracy of 0.944 with the RFC feature selection method and a precision on label 0 of 0.815.

The precision for label 0 is particularly noteworthy, as it indicates the model's adeptness at correctly predicting the negative class, which, in this context, corresponds to the employees who continue their tenure with the company, and due to the imbalance of the dataset, it is considered one of the critical points of this analysis. This metric is crucial in a business setting where falsely identifying an employee as likely to leave could lead to unnecessary intervention costs. Although the DT model displayed high accuracy (0.945) with the GBC feature selection methods, the balance between all metrics guides our choice for further exploration. The RFC model has high accuracy and maintains robustness across other performance indicators, evidenced by the F1 score and ROC-AUC, which are among the highest. The ROC-AUC values for the RFC model, exceeding 0.98, indicate a strong discriminatory ability, whether distinguishing between the employees who stay or leave. The F1 score, which harmonizes the precision and recall, suggests that the RFC model maintains a balanced sensitivity and specificity, which is crucial for making informed decisions in workforce management. Therefore, while the DT model shows promise, the consistent and holistic performance of the RFC and GBC models, especially when paired with the RFC feature selection method, positions them as the leading candidates for hyperparameter tuning and a powerful ensemble model. These models provide a comprehensive understanding of the dataset and present a reliable foundation for developing a turnover prediction tool that can effectively inform human resource strategies.

The high F1 scores accompanying the RFC and GBC models, particularly when paired with the RFC and LR feature selection method, underscore their ability to achieve a harmonious balance between precision and recall. This balance is crucial, as it ensures that the model is accurate overall and effective in identifying true positives without inflating the number of false positives or false negatives. The consistency across different metrics with these models is not accidental but a reflection of their sophisticated algorithms, which can handle the dataset's complexity and uncover subtle turnover patterns. The confusion matrices further corroborate this, revealing fewer false negatives. This is vital in a turnover context where the cost of mistakenly predicting that an employee will stay when they are likely to leave can be high. It's also worth noting the strength of the RFC model, which exhibits high precision for label 0 and strong ROC-AUC values comparable to the GBC model. This suggests that RFC is equally adept at distinguishing between the classes, a desirable trait for any predictive model.

Moreover, the Tensor model shows a strong performance with precision figures comparable to the leading models, suggesting that deep learning approaches have potential in this domain. However, the slight edge in accuracy and the balance of metrics provided by the RFC and GBC models cement their status as the preferred models for hyperparameter tuning.

The objective drives the decision to focus on these models for hyperparameter tuning to refine an already strong predictive performance to achieve an even higher level of accuracy and reliability. In pursuing a robust turnover prediction model, these results provide a clear direction for the next steps in model optimization and validation.

In our analysis, we adjusted the parameters of each model to improve performance, focusing on the precision of class 0 predictions. After this optimization, we retested each model, and the results are presented in Table 12. These refined models set the stage for constructing the final ensemble model.

Table 12 - Refined Models Performance Summary

Model	Accuracy	Precision	Precision Label 0	Recall	F1 score	ROC-AUC	Confusion Matrix
KNN_HP	0.892	0.936	0.667	0.935	0.936	0.802	[[174, 86], [87, 1255]]
LR_HP	0.876	0.954	0.591	0.896	0.924	0.899	[[202, 58], [140, 1202]]
DT_HP	0.937	0.961	0.812	0.964	0.962	0.890	[[207, 53], [48, 1294]]
MLP_HP	0.920	0.950	0.760	0.955	0.952	0.934	[[193, 67], [61, 1281]]
RFC_HP	0.953	0.972	0.851	0.971	0.972	0.985	[[223, 37], [39, 1303]]
GBC_HP	0.955	0.965	0.898	0.982	0.973	0.981	[[212, 48], [24, 1318]]

After the hyperparameter optimization (HP), the models exhibit a refined performance, with notable improvements in accuracy and precision for class 0. The results, as depicted in the attached figure, highlight the enhanced capabilities of each model to predict the 'Active' employee status, which is crucial due to the initial class imbalance.

The Decision Tree (DT_HP), Multi-Layer Perceptron (MLP_HP), Random Forest Classifier (RFC_HP), and Gradient Boosting Classifier (GBC_HP) models show a marked increase in precision for label 0 compared to their K-Nearest Neighbours (KNN_HP) and Logistic Regression (LR_HP) counterparts. Specifically, the GBC_HP model stands out with the highest precision for label 0 at 0.898, indicating its exceptional ability to correctly identify 'Active' status employees, which is vital for avoiding unnecessary retention strategies.

The recall metric, which measures the model's ability to find all relevant instances of class 0, is remarkably high across all models, with the GBC_HP model achieving the highest recall of 0.982. This suggests that the GBC_HP model will most likely identify all 'Active' employees after hyperparameter tuning, with very few slipping through the net.

F1 scores, which balance precision and recall, also increased, with the RFC_HP and GBC_HP models showcasing the top scores of 0.972 and 0.973, respectively. These scores reflect a

robust model performance, especially when the costs of false negatives and false positives are high.

The ROC-AUC values for all models are strong, particularly for the RFC_HP and GBC_HP models, the only models scoring above 0.98. This demonstrates their ability to distinguish between the 'Active' and 'Terminated' classes. High ROC-AUC values indicate the models' overall effectiveness and confirm that the chosen threshold separates the two classes well.

Lastly, examining the confusion matrices reveals that the RFC_HP and GBC_HP models predict 'Active' status with high precision and maintain low false positives and false negatives. For instance, the GBC_HP model's confusion matrix of [[212, 48], [24, 1318]] indicates a robust predictive power, successfully capturing the majority of true 'Active' and 'Terminated' cases while keeping incorrect predictions to a minimum.

In conclusion, the GBC_HP and RFC_HP models, following hyperparameter optimization, are shown to be the most proficient in terms of precision for label 0, recall, F1 score, and ROC-AUC. Despite the initial class imbalance, their strong performance across these metrics demonstrates their suitability for the task of turnover prediction, fulfilling the objectives of precision and reliability in a predictive model for this master thesis.

As we approach the culmination of our methodology to construct an optimal predictive model, our focus is on accurately identifying employees at risk of termination. After a thorough process of hyperparameter optimization, we have narrowed down our selection to the three most promising models—the Gradient Boosting Classifier with Hyperparameter tuning (GBC_HP), the Random Forest Classifier with Hyperparameter tuning (RFC_HP), and the Tensor model. Their demonstrated predictive strength informed this selection, and prior analyses yielded substantial results. These models have shown their adeptness in general performance metrics and have proven especially capable in predicting the crucial class 0, despite the challenges posed by class imbalance. Our next step is to evaluate these models' final performance to establish the most effective approach to predicting employee turnover. In Table 13, we have a summary of the performance metrics for the produced ensemble models.

Table 13 - Ensemble Models Performance Metrics Summary

Model	Feature Selection Method	Accuracy	Precision	Precision Label 0	Recall	F1 score	ROC-AUC	Confusion Matrix
Ens_RF_GBC_Tensor	LR	0.951	0.961	0.892	0.981	0.971	0.980	[[207, 53], [25, 1317]]
Ens_RF_GBC_Tensor	RFC	0.956	0.964	0.909	0.984	0.974	0.982	[[211, 49], [21, 1321]]
Ens_RF_GBC_Tensor	GBC	0.958	0.963	0.925	0.987	0.975	0.983	[[209, 51], [17, 1325]]

The ensemble models, which integrate the predictive capabilities of the GBC_HP, RFC_HP, and Tensor models, present a nuanced improvement in performance metrics, as depicted in the attached image. The focus on precision for label 0 is particularly salient given the initial class imbalance, and it's a critical factor in the model's effectiveness in practical HR scenarios.

The ensemble model using the Logistic Regression (LR) feature selection method, Ens_RF_GBC_Tensor_LR, shows a solid accuracy of 0.951. Its precision for label 0 is commendable at 0.892, indicating a high rate of correct predictions for 'Active' employees. The ensemble's ability to discern true 'Active' from 'Terminated' employees is further reflected in an ROC-AUC of 0.980, a testament to its excellent classification ability. Moving to the ensemble with the Random Forest Classifier (RFC) feature selection method, Ens_RF_GBC_Tensor_RFC, there is a slight increase in accuracy to 0.956. The precision for label 0 increases to 0.909, which is a significant improvement and suggests that this ensemble is particularly effective in accurately identifying employees who are not at risk of turnover. Additionally, with an F1 score of 0.974, this model demonstrates a strong balance between precision and recall, crucial for maintaining reliability across varying thresholds.

The ensemble with the Gradient Boosting Classifier (GBC) feature selection method, Ens_RF_GBC_Tensor_GBC, maintains this trend with an accuracy of 0.958 and the highest precision for label 0 at 0.925. This ensemble not only excels in precisely predicting the 'Active' class but also boasts the highest ROC-AUC of 0.983, indicating its superior ability to differentiate between the two classes over a variety of threshold settings. Across all ensemble models, the recall is consistently high, demonstrating the ensembles' ability to correctly identify the majority of 'Active' cases. The confusion matrices reinforce this, showing a low number of false negatives and false positives, which is crucial in minimizing the risk of unintended consequences from incorrect predictions.

In summary, the Ens_RF_GBC_Tensor_GBC ensemble emerge as the front runner, considered upfront as the “best model”, showing a slight edge due to its precision for label 0 and ROC-AUC. This ensemble, through their sophisticated integration of multiple models and feature selection techniques, have successfully addressed the challenge of class imbalance and have proven to be robust in their predictive performance. This positions it as potentially valuable tools in the strategic management of employee retention, offering insightful and actionable data to HR professionals.

The following analysis will be central to the model chosen as the Best Model Ens_RF_GBC_Tensor_RFC, as it has proven to be the best model with the best predictability, accuracy, and performance. We will dive deeper in the analysis, by evaluating the features importance, permutation importance and other analysis, relevant for this study.

4.1.1.2. Feature Importance

In the ensemble model composed of RandomForestClassifier, GradientBoostingClassifier, and a Keras neural network, our feature importance analysis is drawn from the tree-based components, which intrinsically provide this measure. The Keras model, despite its strengths, lacks a built-in feature importance capability, so the tree-based classifiers will inform our insights here.

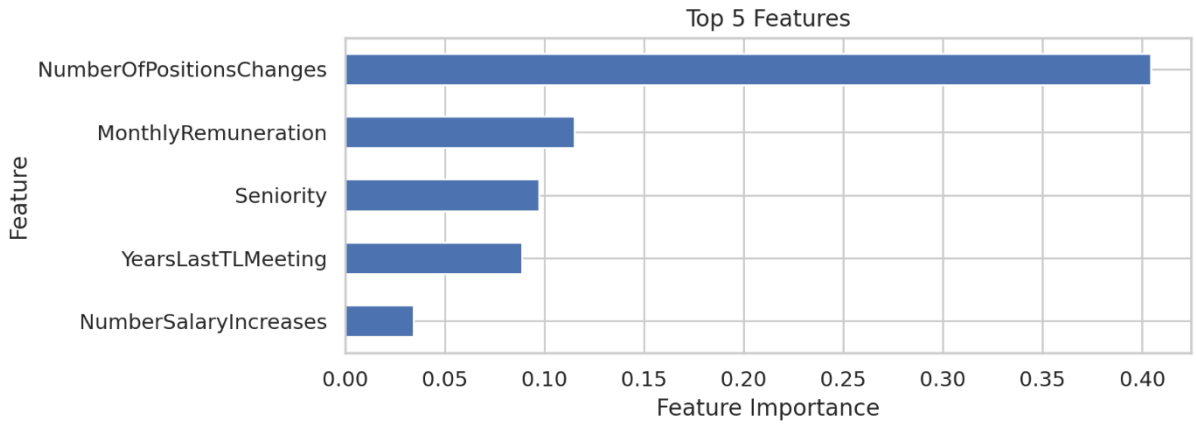


Figure 10 - Best Model Top 5 Features by Importance

In the Figure 10, the chart presented illustrates the top five features by importance, as identified by the ensemble model. The 'NumberOfPositionsChanges' emerges as the most influential factor, indicating that the frequency of position changes within the company significantly predicts employee turnover. This could suggest that employees who transition through roles more frequently are at higher risk of leaving, perhaps reflecting issues with job fitness or career progression. The importance placed on 'MonthlyRemuneration' highlights the role of financial compensation in retention, which aligns with the well-documented correlation between pay levels and job satisfaction. 'Seniority' is the next crucial feature, underscoring the intuitive understanding that the longer employees stay with an organization, the less likely they are to leave. Interestingly, 'YearsLastTLMeeting' is identified as a significant predictor. This could indicate that regular interaction with leadership is key to employee engagement and retention, supporting the notion that managerial support and recognition play into an employee's decision to stay. Lastly, 'NumberSalaryIncreases' rounds out the top five, pointing to the impact of financial recognition on employee loyalty. Frequent salary increases may serve as a barometer for career development, with stagnation potentially driving turnover. These findings from the tree-based models within our ensemble resonate with established theories in people analytics. They affirm that both tangible rewards, such as remuneration and role stability, and intangible factors, like leadership engagement, are instrumental in predicting employee turnover. Applying these insights in the real world can

inform targeted strategies for employee retention, guiding HR interventions to address the root causes of turnover effectively. This analysis validates the chosen model's predictive power and provides actionable intelligence for managing the workforce.

4.1.1.3. Permutation Importance

The results for permutation importance provide the mean importance value along with its standard deviation for each feature. These results can be interpreted to understand the significance of each feature in predicting employee turnover. The mean importance value indicates how crucial each feature is, with higher values representing more important features. The standard deviation indicates the variability or uncertainty in the importance estimate, with larger values suggesting greater uncertainty. Permutation importance analysis is a powerful tool in people analytics for interpreting predictive models. It provides insights into the stability of feature importance across the validation set by calculating the decrease in model accuracy when each feature's information is randomly shuffled. This approach helps identify which features are consistently important and which ones have more variability in their importance. Figure 11 showcases the top ten features with their respective importance scores and variability. This visualization helps to easily identify the most critical factors in the model and assess their reliability.

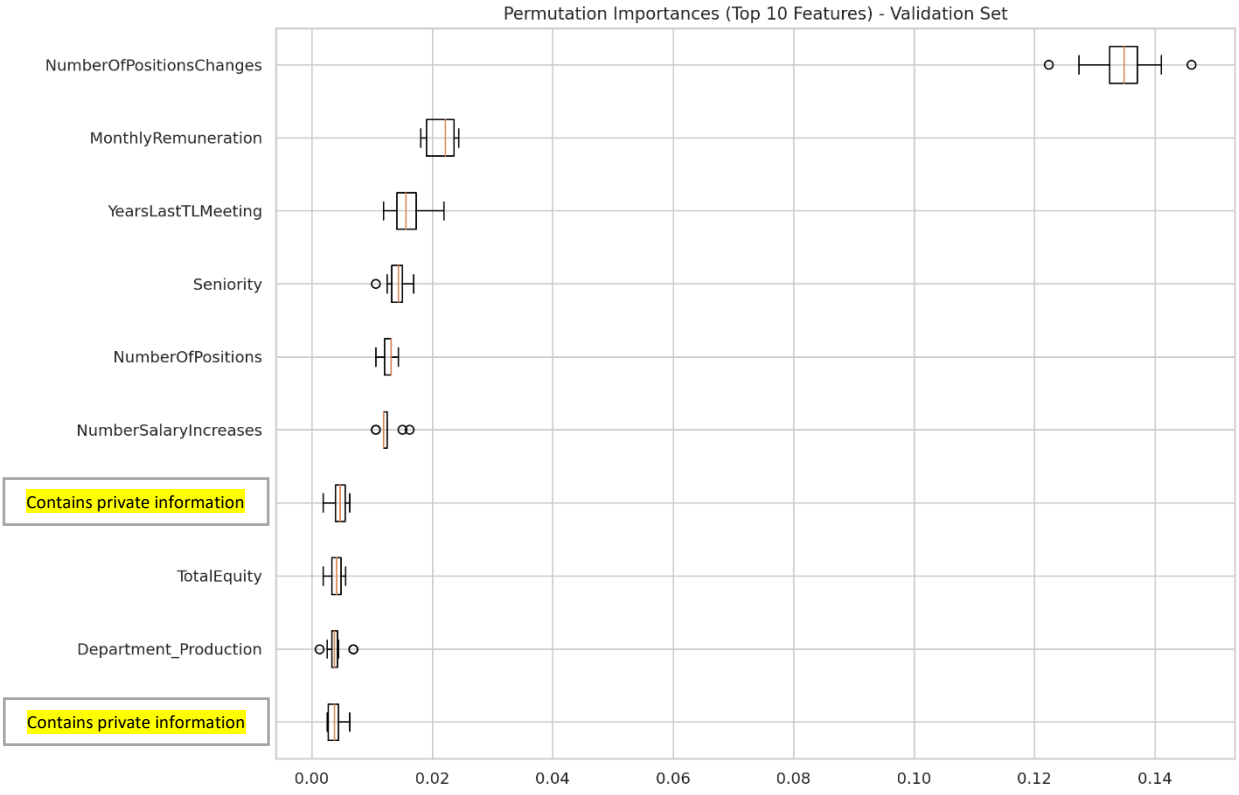


Figure 11 - Permutation Feature Importance on mean importance value.

The chart presents the permutation importances for the top 10 features from a validation set, aimed at identifying key factors influencing employee turnover. 'NumberOfPositionsChanges' stands out as the most significant feature, suggesting that the number of times an employee changes roles within the company is a strong indicator of turnover. This observation aligns with organizational behaviour research, which highlights that frequent job changes could indicate dissatisfaction or a quest for new challenges. The 'MonthlyRemuneration' feature also shows notable importance, confirming the established connection between pay and turnover intentions. This suggests that ensuring competitive compensation is crucial for retaining talent, as financial considerations play a significant role in career choices. The 'YearsLastTLMeeting' is another feature with considerable importance, reinforcing the importance of management interaction in employee retention. Consistent and impactful meetings with team leaders could be essential in nurturing loyalty and a sense of belonging. The 'NumberOfPositions' feature, indicating the variety of job titles an employee has held, shows a solid presence with relatively low variability. This reflects its consistent relevance in predicting employee departures. Interestingly, the 'NumberSalaryIncreases' feature is also a strong predictor of retention, with relatively lower variability, indicating that how often employees receive pay raises—a sign of recognition and value—can significantly influence their likelihood to stay. 'Seniority' is another prominent feature, supporting the idea that employees with longer tenure are less prone to leaving, assuming other factors do not intervene. However, unlike the initial text provided, features such as 'EmploymentType_Part-Time' and 'JobLevel_L1' do not appear in the top 10, suggesting that they may not be as critical in this specific analysis as other factors like 'OfficeCountry', 'TotalEquity', 'Department_Production', and 'Division', which do appear and thus warrant consideration. The variability depicted by the boxplot whiskers for each feature's importance speaks to the robustness of each feature's predictive power. For instance, 'NumberOfPositionsChanges' and 'MonthlyRemuneration' exhibit some variability, implying that their impact on the model's accuracy might vary across different data subsets. In contrast, 'TotalEquity' and 'Department_Production' show tighter distributions, indicating more consistent predictive power.

Overall, this analysis validates the importance of these features in the predictive model and offers valuable insights. HR departments can use this information to prioritize retention strategies that focus on career development, effective management engagement, competitive remuneration, and recognition of employee achievements to effectively reduce turnover risks.

4.1.1.4. Partial Dependence Plots

The partial dependence plots serve as a powerful interpretative tool, revealing the relationship between key features and the likelihood of employee turnover. These graphical

representations, as will be illustrated in Figure 12, elucidate how variations in the top 5 features impact the likelihood of an employee's departure from the organization.

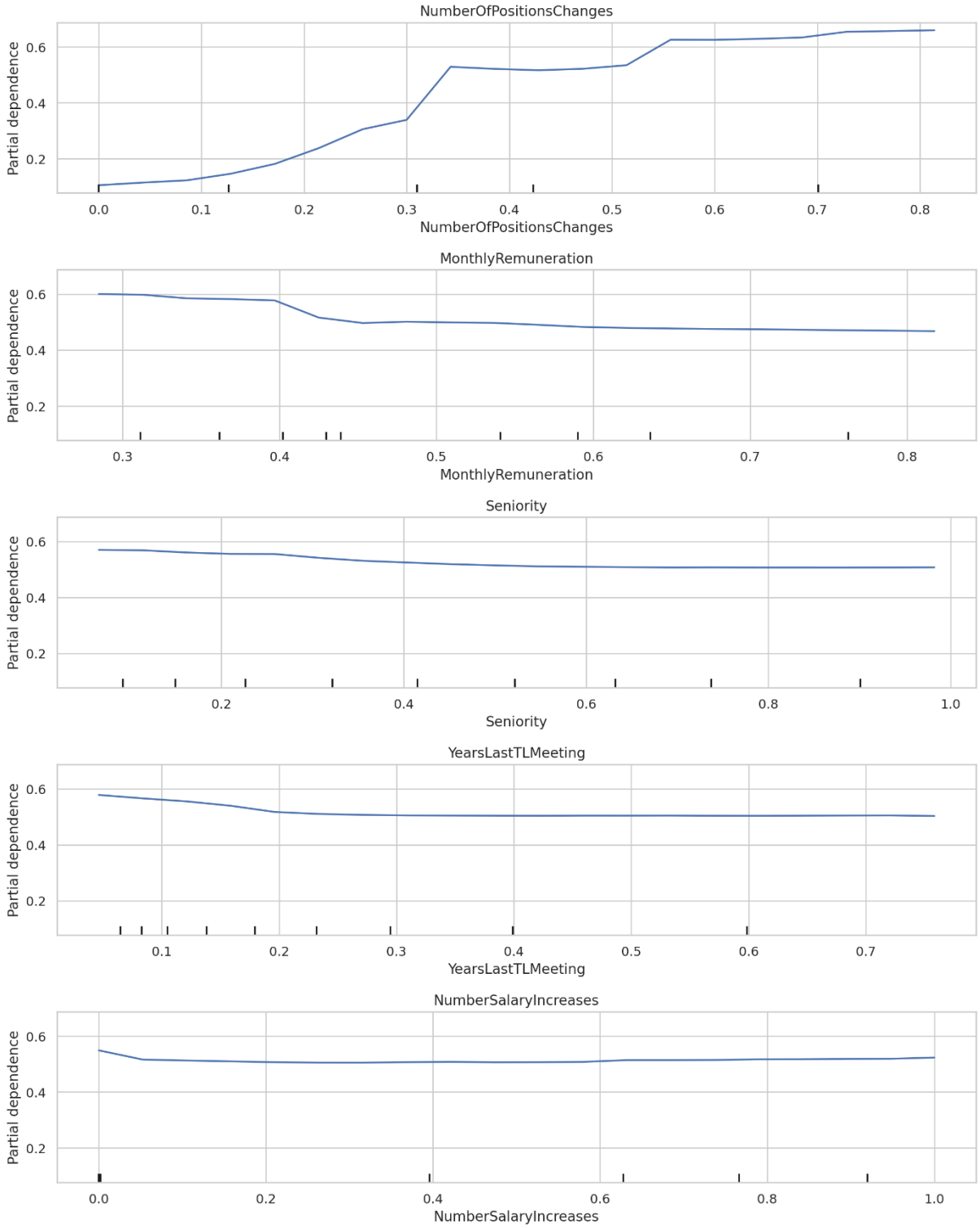


Figure 12 - Partial Dependence Analysis on the Top 5 features

Our analysis begins with the feature 'NumberOfPositions,' which exhibits a progressive increase in turnover probability as the number of positions an employee holds rises. This trend suggests a potential mismatch between the employee's career expectations and the

opportunities available within the company, prompting considerations for talent development and career progression pathways. In 'Seniority', inversely correlates with turnover probability, reinforcing the notion that employees with extended tenures are more inclined to stay. These findings underline seniority as a proxy for deeper organizational attachment and the accumulation of valuable company-specific knowledge, which can discourage turnover. The 'MonthlyRemuneration' plot delineates a negative gradient, asserting the critical role of compensation in retention strategies. The inverse relationship here reaffirms the fundamental principle that competitive remuneration packages are essential in mitigating turnover risk. Further scrutiny of 'YearsLastTLMeeting' reveals a downward trend, indicative of the importance of regular leadership engagement in nurturing employee commitment. This aspect of the analysis points to the potential of managerial support as a catalyst for enhancing employee satisfaction and loyalty. Lastly, the declining trajectory observed in 'NumberSalaryIncreases' suggests that recognition through salary increments plays a pivotal role in retention. This trend likely reflects the dual impact of financial reward and acknowledgment of an employee's contributions, bolstering their decision to remain with the company.

Collectively, these plots not only confirm well-established theories within the field of human resource management but also shed light on nuanced predictors of employee turnover. This synthesis of quantitative insights provides a valuable framework for human resource professionals to devise comprehensive retention strategies responsive to employee turnover's multifaceted nature.

4.1.1.5. Sensitivity Analysis

Sensitivity analysis, represented on Figures 13, 14, 15, 16 and 17 examines how changes in input variables affect the output of a model. This analysis helps understand the robustness and stability of model predictions under varying input conditions. In the context of predictive modeling is a technique used to assess how the predictors of a model are affected by changes in input features. It gauges the robustness of the model and provides insight into which features are most influential in the model's predictions. We can observe the impact on the predicted outcome by systematically varying a single input while holding others constant. The objective is to discern which features have the most substantial sway on the model's decision-making process, hence guiding more informed and strategic data-driven decisions.

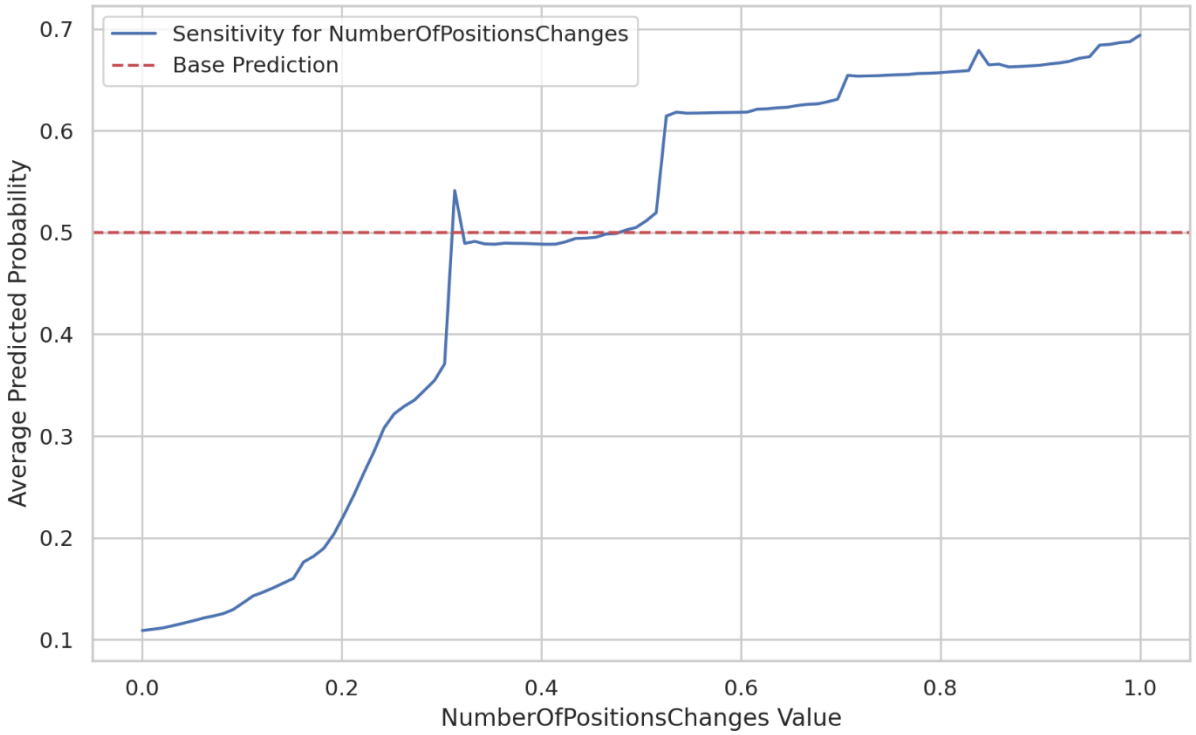


Figure 13 - Sensitivity Analysis on feature “NumberOfPositionsChanges”

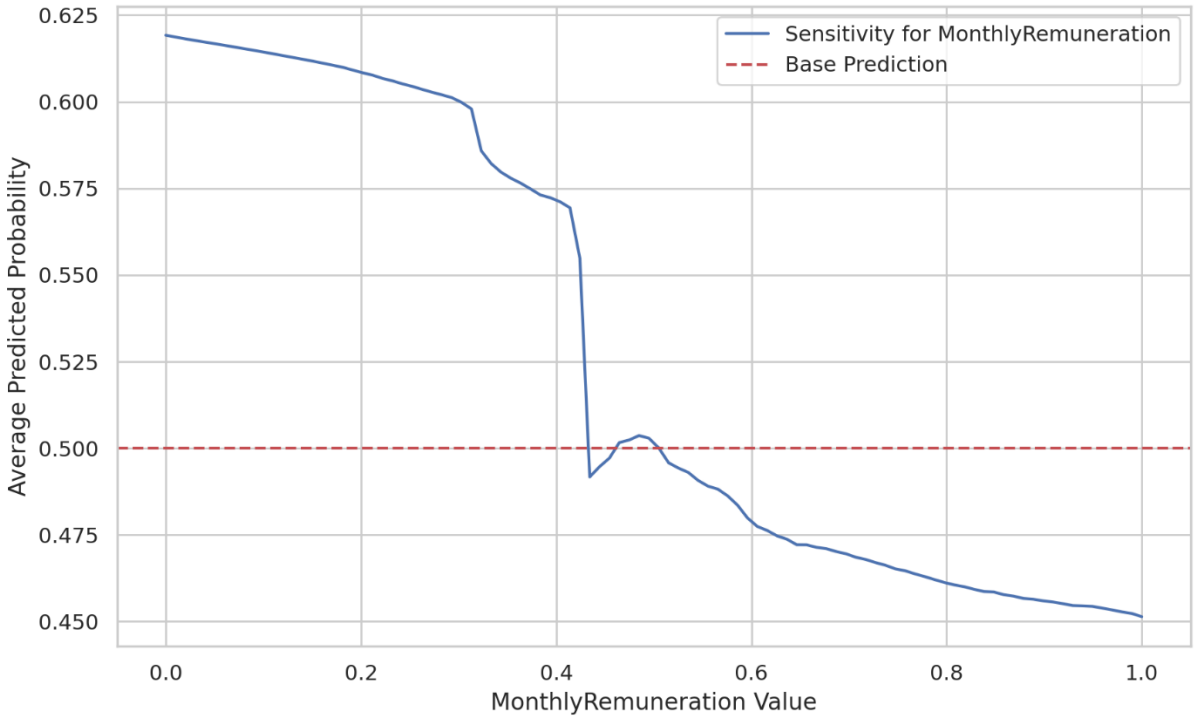


Figure 14 - Sensitivity Analysis on feature “Monthly Remuneration”

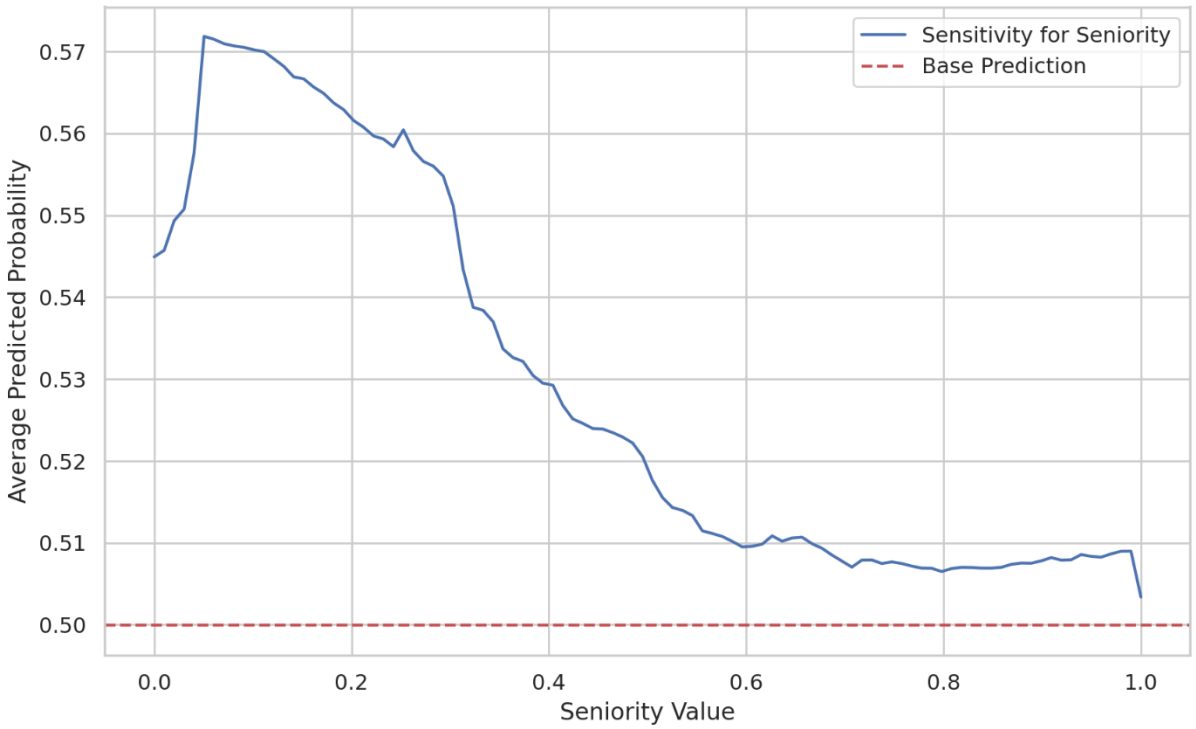


Figure 15 - Sensitivity Analysis on feature “Seniority”

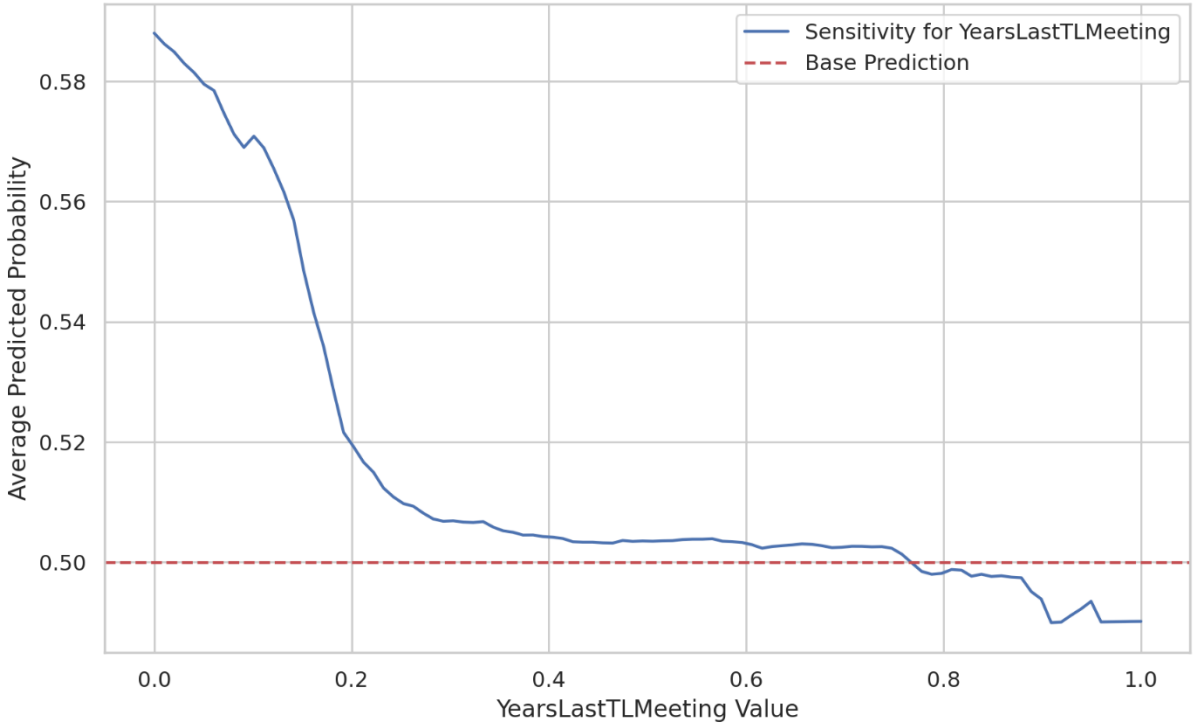


Figure 16 - Sensitivity Analysis on feature “YearsLastTLMeeting”

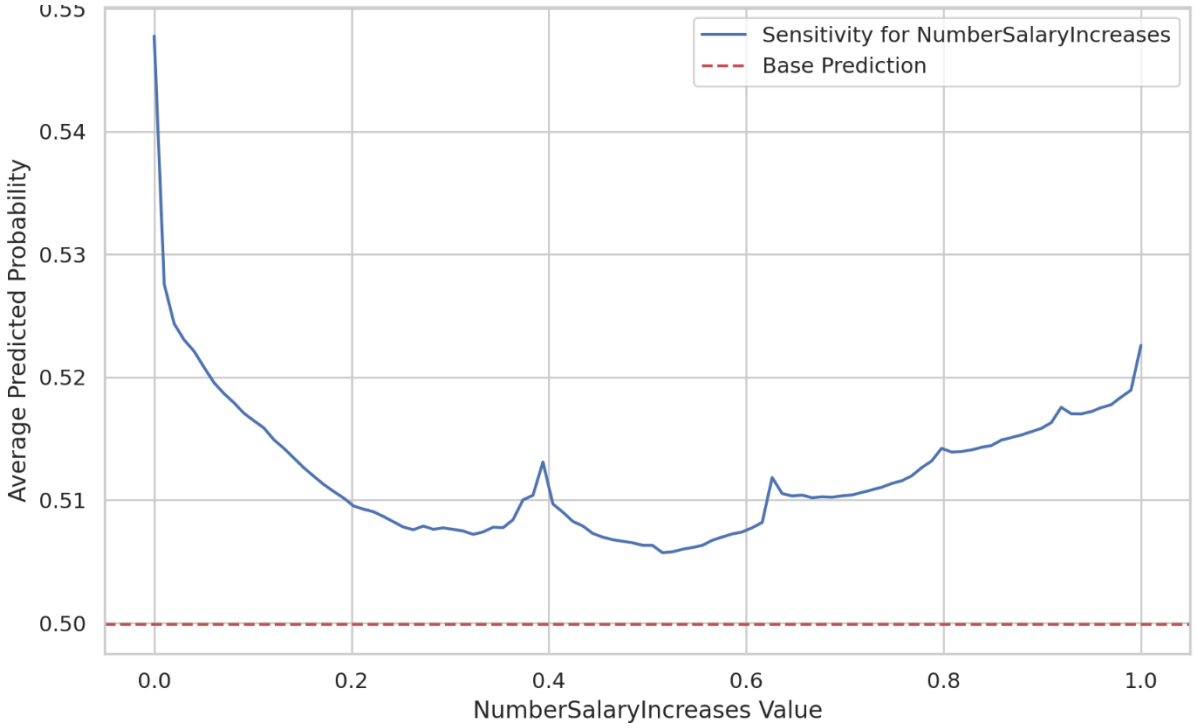


Figure 17 - Sensitivity Analysis on feature “NumberSalaryIncreases”

In Figure 13, the graph illustrates a sharp increase in the predicted probability of turnover as the number of position changes increases, particularly after a certain threshold. This suggests a strong positive correlation between the frequency of position changes and the likelihood of turnover. Organizations may interpret this as an indicator that employees seeking frequent role changes might be at a higher risk of leaving, potentially pointing to underlying issues with job satisfaction or career progression. By analyzing Figure 14, we found that the curve indicates a drop and then a gradual decline in turnover probability as monthly remuneration increases, with a notable dip and subsequent fluctuations. The initial decrease may reflect the retention power of competitive salaries, yet the fluctuations suggest that beyond a certain salary point, other factors may come into play that influence turnover intentions. It indicates that while salary is an important factor in retention, its impact can be non-linear and potentially moderated by other variables. Figure 15 shows a general downward trend in turnover probability as seniority increases, implying that employees with longer tenure are less likely to leave. This trend is consistent with the belief that more tenured employees have stronger ties to the company and potentially more to lose from leaving, such as accumulated benefits and social bonds. In Figure 16, the plot of the sensitivity analysis on "YearsLastTLMeeting" shows a downward trend, indicating that the longer it has been since the last team leader meeting, the higher the predicted turnover probability. This may suggest that regular and recent interactions with leadership can stabilize employees, possibly through improved communication, recognition, and engagement. Finally, in Figure 17 analyzing the "NumberSalaryIncreases", the chart presents a U-shaped curve, with the turnover probability

decreasing and then slightly increasing with the number of salaries increase. It may imply that while salary increases can initially improve employee retention, the effect might plateau or reverse if increases are too frequent, possibly due to heightened expectations or other factors becoming more prominent in the employee's decision to stay or leave.

In conclusion, the sensitivity analysis provides valuable insights into employee turnover factors. It suggests a multifaceted approach to retention, one that recognizes the complexity of the workforce dynamics where stability, financial incentives, and recognition play vital roles. Yet, their impact is not always linear or straightforward. These insights can be instrumental for HR professionals in crafting nuanced strategies to foster a stable and engaged workforce. The sensitivity analysis across these features illustrates a complex interplay of factors influencing employee turnover. Financial compensation, stability, recognition, and managerial engagement are central retention themes. However, the non-linear trends observed for 'Monthly Remuneration' and 'NumberSalaryIncreases' highlight that employee retention is multifaceted and cannot be distilled into purely transactional elements. These insights are imperative for developing informed HR strategies, where a balanced approach addressing financial and non-financial drivers is necessary to enhance employee retention effectively. The professional takeaway for HR practitioners is the need for a strategic, data-informed understanding of turnover drivers that account for individual and organizational factors in concert.

4.1.1.6. Interaction Effect Analysis on Top 5 Important Features

In predictive modelling, understanding the individual contribution of each feature is often insufficient to capture the complex relationships inherent within the data. This is especially true in the context of employee termination prediction, where the interplay between different factors can provide deeper insights into the underlying patterns that drive outcomes. To address this, we employ a method known as SHAP (Shapley Additive exPlanations) interaction value analysis, which extends the SHAP value concept by quantifying the interaction effect between pairs of features. Through the different analysis done before, we did this analysis to the priori identified top 5 important features, as they are the five most influential features within our predictive model. Due to computational limitations and the nature of ensemble models, the interaction effect analysis is performed on a single model rather than the entire ensemble used in the termination prediction. We chose the Gradient Boosting Classifier as our reference model for this analysis, considering its robust performance and interpretability within the ensemble. This approach enables us to explore the intricate relationships between the top features and their collective impact on the predictive performance. Below we have the graphical result of our analysis, the heatmap where is possible to visualize the SHAP interaction values among the top five features, reflecting the degree to which pairs of features synergistically influence the model's output.

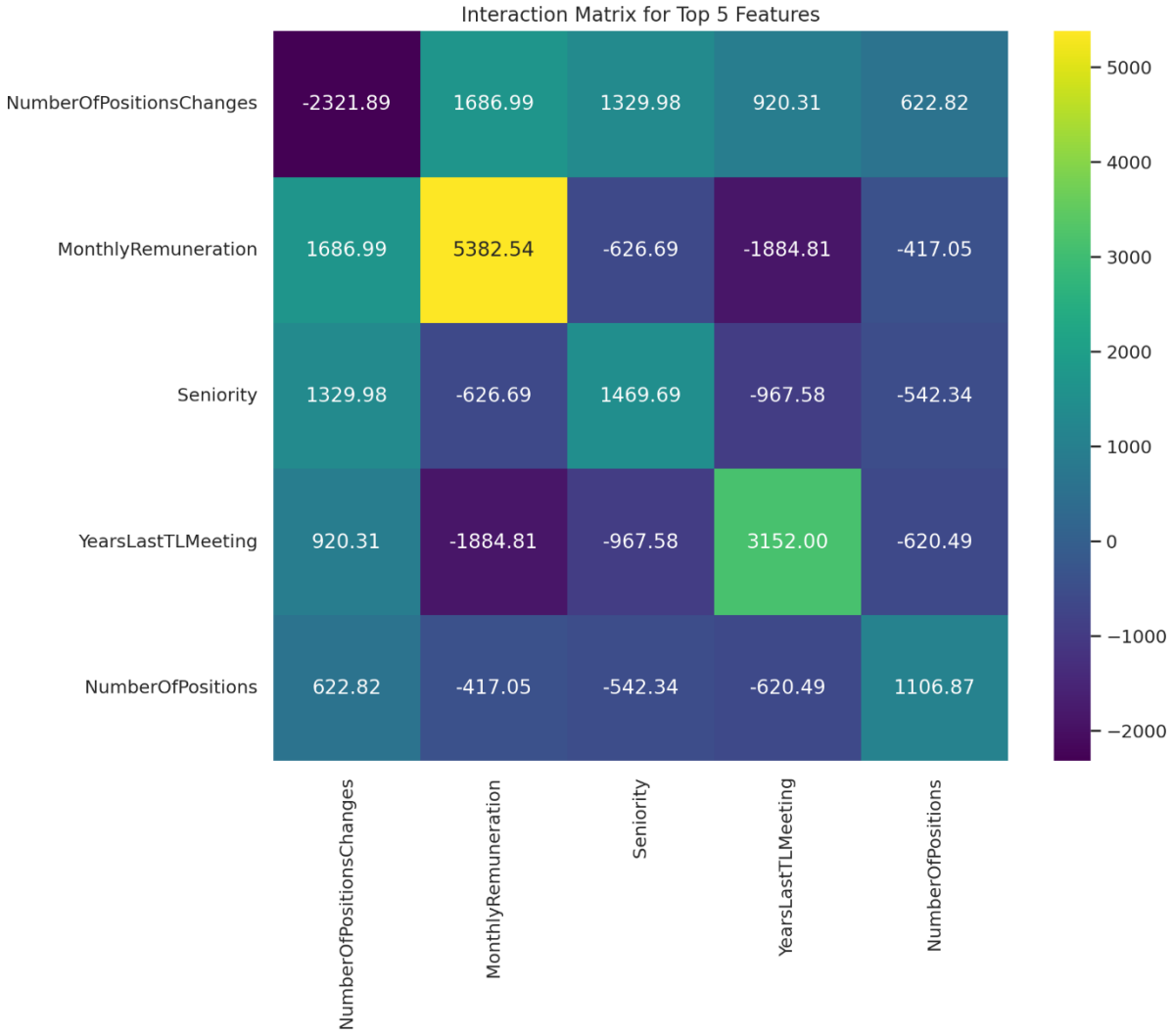


Figure 18 - Interaction Effect Analysis on Top 5 Important Features

In analysing the top predictors of employee turnover from our predictive model, we delve into the relationships between the most important factors. 'NumberOfPositionsChanges' and 'MonthlyRemuneration' interact in a way that really stands out, with a strong positive effect suggesting that employees who've changed roles frequently and earn more are more likely to leave. This could indicate that well-paid employees who have had several roles might be on the lookout for new challenges or better offers elsewhere. On the other hand, when we look at 'MonthlyRemuneration' and 'Seniority', there's a strong negative interaction. It seems that employees who have been with the company longer and earn higher salaries are less likely to quit. This makes sense when you think about it – if you're well-compensated and have put in the years, you might be quite settled and less inclined to leave. 'Time since the last meeting with a team leader', represented by 'YearsLastTLMeeting', also shows an interesting pattern. By itself, it's a strong predictor of whether someone might leave, with the longer the time since the last meeting, the more likely they are to quit. But curiously, when you combine this with high pay ('MonthlyRemuneration'), the risk of them leaving seems to go down. Lastly, the

number of different positions an employee has held, labelled as 'NumberOfPositions', also interacts with pay in an intriguing way. Higher pay makes it less likely for someone to leave, even if they've had several positions within the company.

This interaction analysis gives us a clearer picture of what might drive an employee to stay or go. It's not just about how much they earn or how long they've been around; it's about how these and other factors come together to influence their decisions. Understanding these complex relationships helps us to anticipate turnover better and perhaps even to address it proactively. The SHAP interaction analysis of the top five features reveals complex, non-linear relationships critical to understanding employee termination dynamics. The positive and negative interaction values provide insights into how combinations of factors can either increase or decrease the likelihood of termination beyond their individual effects. This nuanced understanding can guide strategic human resource interventions, enhance predictive accuracy, and inform more effective retention policies.

By analysing the interaction effects, we gain a clearer picture of how features contribute individually and in concert with others, which can have a transformative impact on the interpretability of our predictive model. This, in turn, highlights the necessity of incorporating such interaction effects into the modelling process for a more holistic and accurate representation of the phenomenon under study.

4.1.1.7. Probability of the Active Employees

To produce a more robust approach to the topic in analysis and to provide the company with a concrete result and an output, we present an analysis of termination probabilities among currently active employees using the predictive model. The model, which was initially trained to distinguish between terminated and active employees, has been employed to estimate the likelihood that each active employee might leave the company. The rationale for this analysis is to provide a data-driven approach for identifying employees who may be at risk of leaving, allowing for targeted retention efforts, or to use as a combination point for further analysis. The predictive model's output has been translated into a probability percentage, with higher percentages indicating a greater likelihood of employee departure. The threshold for high risk has been set at 80%, a value chosen based on model performance and the desired balance between sensitivity and specificity. This high threshold ensures we focus on employees most likely to leave, allowing HR interventions to be as effective and efficient as possible. The result of this application was the one presented in Figure 19.

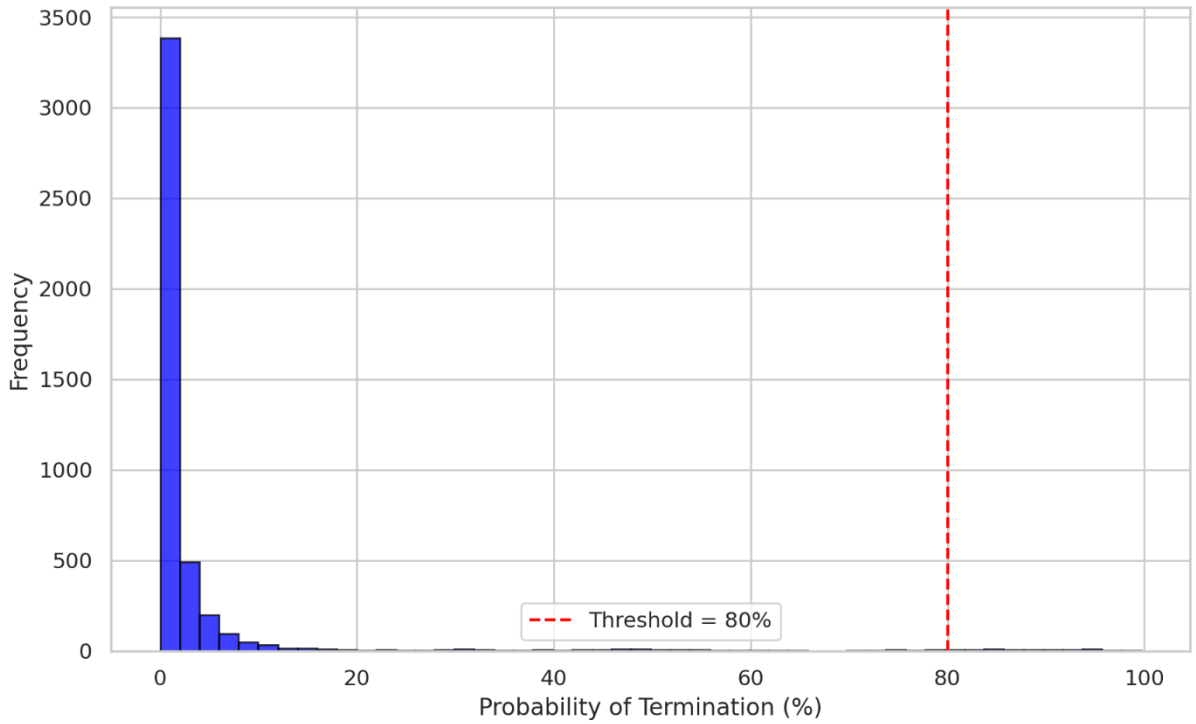


Figure 19 - Distribution of Termination Probabilities for Active Employees

We present the distribution of termination probabilities for active employees in the accompanying chart. According to the model's current estimation, most employees fall below the 20% probability mark, indicating a low likelihood of termination. However, a tail in the distribution extends towards higher probabilities. These are the employees that, based on the model's prediction, could be considered at a higher risk of turnover. We identified 58 employees regarding the threshold applied. Such information could be instrumental for human resources to pre-emptively address potential turnovers, perhaps through targeted engagement strategies or career development opportunities. Additionally, the output from this analysis can create synergies with other HR initiatives, such as performance management, employee satisfaction surveys, and strategic planning, where the model results cannot be solely overlooked when deciding on the action to take. By integrating predictive termination probabilities with these other dimensions, the organization could develop a more nuanced understanding of employee dynamics and design personalized retention programs.

Overall, this probability-based approach seeks to leverage predictive analytics to inform and enhance human resource management, demonstrating the potential of machine learning in operational decision-making within organizations.

4.2. DISCUSSION AND FINDINGS

The study's results are compelling, indicating that the ensemble models employed, particularly the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC), provide predictive solid accuracy in the context of employee turnover. This is evidenced by performance metrics that consistently approach a 0.90 threshold for accuracy and precision, particularly notable given the use of a real-world dataset. The high performance of the predictive models suggests that they are robust tools for understanding and forecasting employee turnover. This capability allows for more targeted and effective HR interventions aimed at reducing turnover rates, thus offering substantial value to organizations. Moreover, the detailed examination of feature importance highlights specific factors such as 'Number of Positions', 'Seniority', and 'Monthly Remuneration' that significantly influence turnover decisions, reinforcing the need for strategic HR management. The findings directly affect HR practices, particularly in crafting strategies that address the identified key predictors of turnover. For example, understanding that higher 'Monthly Remuneration' and greater 'Seniority' reduce turnover propensity can lead to better compensation strategies and career development programs to retain talent. This study contributes to the existing literature by applying advanced machine-learning techniques to real-world data, thus bridging the gap between theoretical turnover models and practical, actionable insights. It extends our understanding of turnover predictors in the modern workplace and demonstrates the applicability of ensemble models in HR analytics. The robustness of the RFC and GBC in predicting turnover is consistent with existing studies that highlight the effectiveness of ensemble methods in handling complex datasets and providing reliable predictions. This study enhances the literature by integrating these methods with real-world data, thereby offering a nuanced view of employee turnover that is grounded in empirical evidence. The findings align with prior research on the significance of compensation and job tenure in influencing employee turnover. However, this study also introduces nuanced insights into how different features interplay in the context of turnover, such as the role of frequent position changes, which have been less emphasized in previous studies. These insights are valuable for developing more comprehensive and tailored retention strategies. In summary, this thesis not only reaffirms established theories on employee turnover but also expands them by integrating advanced predictive modeling techniques and a rich, real-world dataset. The use of ensemble models has proven particularly effective, offering a high degree of predictive accuracy and valuable insights for strategic HR management.

4.3. CONCLUSIONS

In this chapter, we dive into two major points, the performance results obtained by the different models applied in the study and following the mechanism snowed in the methodology chapter, and the deep dive in an analysis to comprehend the models and the problem itself. By exploring the different points of view regarding the importance of features

classified by the models, the permutation importance analysis, and the remaining ones it permits us to identify key point of focus to explore and mark as future guides for work. This represents a good opportunity for new exploration and combination with other datasets where the features could be arranged in the same order.

In conclusion of the results reflected on the development and efficacy of predictive models designed to predict employee termination, and addressing the model's performances, the findings suggest that the Random Forest and Gradient Boosting Classifiers, following hyperparameter optimization, demonstrated superior performance across several metrics. These models not only achieved high accuracy and precision but also maintained robust recall and F1 scores, which are crucial for the balanced classification of both 'Active' and 'Terminated' employees. Particularly, these models excelled in identifying the negative class (non-leavers), which is vital given the initial class imbalance. This ability ensures that the models can predict turnover without overfitting to the more frequently occurring class. Moreover, the ensemble model, integrating the strengths of the Random Forest, Gradient Boosting, and the considered Tensor model, showed even further improvement in performance metrics. This ensemble approach leveraged diverse methodologies to enhance the predictability of the turnover model, ensuring high reliability and applicability in practical HR scenarios. The ensemble model's high precision in predicting 'Active' status of employees indicates its potential utility for HR departments to implement effective retention strategies based on reliable data. These results are not only significant in terms of achieving high performance metrics but also in understanding the underlying dynamics of employee turnover. Features such as the number of position changes, monthly remuneration, and seniority were identified as significant predictors of turnover. Such insights can help organizations tailor their HR strategies to address the root causes of turnover effectively. This study's comprehensive approach—from initial data analysis to model refinement and validation—demonstrates the potential of machine learning techniques in addressing complex HR issues like turnover. The successful application of these models provides a foundation for further research and development in this field, suggesting pathways for future studies to explore additional data sources, alternative modelling techniques, and cross-industry validations.

With the objective on going beyond the model's performance and contribute for the HR study on employee termination, and to give specific guides on the behaviour of terminated employees and to give them a tool to anticipate employee turnover, we worked on detailed analysis regarding the final ensemble model that represented the best model. The results and findings from the detailed analysis provide significant insights into the predictors of employee turnover and highlight the nuanced interactions between various employee characteristics and their impact on retention. The comprehensive feature importance analysis identified several key predictors such as the number of position changes, monthly remuneration, and seniority. This analysis points to the significant role that stability and financial incentives play in influencing employee retention. For instance, employees experiencing frequent position

changes are more likely to leave, suggesting that job instability can lead to dissatisfaction and turnover. Conversely, higher remuneration and longer tenure correlate with increased retention, indicating that financial rewards and job security are critical to keeping employees engaged and committed to the company. Additionally, the permutation importance results further emphasized the critical nature of these factors, offering a robust statistical backing to the feature importance findings. This analysis confirmed that not only are these features statistically significant, but their impact on model accuracy is profound, reinforcing the need for HR strategies that focus on enhancing job stability and financial satisfaction to mitigate turnover risks. Moreover, the partial dependence plots provided visual confirmation of these relationships, illustrating how changes in these key features affect the probability of turnover. For instance, increases in the number of position changes were clearly associated with higher turnover rates, while increases in seniority and monthly remuneration demonstrated a protective effect against turnover. These plots serve as a powerful tool for understanding and communicating the complex dynamics at play, offering clear, actionable insights into how specific changes in workplace conditions and policies might impact employee retention.

These findings underscore the importance of strategic HR management in addressing the multifaceted challenges of employee turnover. By leveraging the insights from this analysis, organizations can better tailor their HR practices to address the specific needs and preferences of their workforce, ultimately leading to more effective retention strategies and a more stable, satisfied employee base. The synthesis of quantitative analysis with practical HR implications provides a valuable framework for proactive employee management and strategic decision-making in HR policies.

In conclusion, the results establish a robust predictive model that not only meets the rigorous standards of accuracy and precision but also provides actionable insights for managing employee turnover. These insights are invaluable for strategic human resource management, offering a proactive tool for organizations to enhance employee retention and operational stability.

4.3.1. Limitations

The dataset we worked with presented several challenges that impacted our analysis and model development. From the beginning, understanding, and manipulating the data to create a usable data frame that could yield valuable insights and form the basis for an accurate model predicting employee termination proved complex. The dataset underwent numerous transformations to address various inaccuracies and discrepancies. We believe that the initial methodological steps effectively resolved these issues. However, the iterative process of analyzing, understanding, and correcting the data, including using support tables to standardize the data transformations, highlighted a potential limitation for future applications. This approach, necessary for our current analysis, would likely need to be

replicated for future datasets the company collects, which could complicate the model's deployment in a production environment. Another significant challenge was missing data, which posed a major limitation to our study. This issue requires further investigation to enhance the dataset's completeness and reliability. Data reliability itself emerged as a crucial concern. We encountered errors where the data did not accurately reflect reality, such as discrepancies between information from different features that should complement each other. An example of this was finding a 'PayType' classified as "monthly" while the associated 'PayRate' suggested an hourly rate, indicating misclassification. While we corrected these points, they underscore the possibility of other inaccuracies within the dataset that we might not have identified. These issues, if unresolved, could affect the model's performance and reliability, representing a limitation in our study.

4.3.2. Future Research

In the domain of predictive modeling for employee turnover, our investigations have opened avenues for significant insights and raised awareness about the complexities involved. Recognizing the importance of data integrity, we underscore the necessity of initial data cleanup, the establishment of rules for data imputation, and the implementation of regular data validation rounds. These foundational steps are imperative to ensure the data's accuracy and relevance for future analyses.

Advancing from these preliminary measures, future research is poised to delve deeper into enhanced data management strategies. The development of algorithms or the application of machine learning techniques for automating data cleaning and validation processes could dramatically streamline these tasks, significantly reducing manual efforts. Moreover, investigating machine learning-based imputation methods presents an opportunity to address missing data with greater sophistication. Techniques such as predictive mean matching, k-nearest neighbors' imputation, or leveraging deep learning approaches promise a more nuanced and precise strategy for data imputation, offering a leap beyond conventional methodologies.

The exploration of predictive models beckons further scholarly pursuit. There is a rich landscape for examining additional predictive models beyond those assessed in this research. The advent of emerging machine learning and deep learning architectures, including attention mechanisms and transformer models, could shed new light on the intricate dynamics of employee turnover. Furthermore, applying and testing these developed models across various industries could provide insights into their adaptability and generalizability, potentially unveiling industry-specific turnover influences. The temporal dynamics of employee turnover also warrant attention, inviting studies on how economic cycles, seasonal trends, or organizational events might impact turnover patterns. Time-series analysis and dynamic modeling could reveal valuable patterns not captured by static models.

Incorporating external data sources offers another promising direction for expanding the predictive capabilities of turnover models. The integration of macro-economic indicators, for example, could enrich the models with a layer of contextual understanding, considering how external economic factors like unemployment rates and industry growth trends influence employee turnover. Similarly, the exploration of social media data and sentiment analysis could tap into employee engagement and satisfaction levels, providing insights beyond traditional datasets through natural language processing techniques.

On the organizational and behavioral front, conducting in-depth analysis of organizational culture, management practices, and employee engagement could offer a comprehensive view of turnover dynamics. This qualitative research would complement quantitative models, providing rich, contextual insights into the factors that influence employee retention or departure. Additionally, studying the stages of an employee's lifecycle within the organization might identify critical junctures that significantly impact retention decisions, guiding targeted interventions to enhance retention at these pivotal moments.

Lastly, as we refine and expand upon these models, the ethical considerations surrounding predictive analytics in the workplace become increasingly pertinent. Ensuring fairness, transparency, and the responsible use of predictive insights is paramount, underscoring the need for ethical frameworks that guide the application of these models in real-world settings. As we venture into these uncharted territories, the horizon for future research on employee turnover appears both broad and deeply compelling, promising novel insights and tools to tackle this pervasive challenge.

5. CONCLUSIONS AND FUTURE WORK

This thesis has tackled the complex issue of predicting employee turnover within a specific company, providing a nuanced understanding of the factors that contribute to employee departures. The study employed a robust methodology involving an extensive data cleaning and transformation process, which was crucial for the development of a reliable predictive model. The model's performance metrics, which indicated high accuracy without overfitting, underscore its effectiveness in capturing the true dynamics influencing employee turnover. Our research identified several critical factors that significantly impact an employee's decision to leave the organization. Key among these were management interaction styles, the transparency and opportunities for career progression, and the overall compensation packages relative to market standards. The model's ability to isolate these factors as significant predictors of turnover provides actionable insights that can inform targeted interventions to improve employee retention. By applying advanced analytical techniques, including feature selection and machine learning algorithms, the study was able to highlight the importance of both quantitative and qualitative attributes in predicting turnover. For example, the inclusion of variables such as job satisfaction scores and past employee engagement levels could further refine the predictive accuracy of the model. These factors often manifest through more subtle interactions and perceptions, which can significantly influence an employee's loyalty and satisfaction. While the study provides significant insights, it also acknowledges its limitation of being confined to a single company's data. This focus, while detailed, restricts the generalizability of the findings across industries or diverse corporate cultures. Future research should aim to replicate this study across various organizations to validate and refine the model's applicability and to incorporate normalization processes that adjust for company-specific variances. Moreover, the predictive model's application to this single company revealed patterns that are potentially applicable to other organizations within similar industries or with comparable organizational structures. For instance, the impact of managerial style on turnover intentions suggests a universal theme across corporate environments where leadership effectiveness directly correlates with employee retention rates.

Focusing on answering the research questions presented in the study, regarding the first question, “1. How can we enhance attrition prediction accuracy by employing artificial neural networks compared to conventional classifiers?” we conclude that despite the good performance on the conventional classifiers used in the study, employing artificial neural networks helped improve the performance even further when applied the ensemble model with ANN model developed, resulting in improved accuracy and robustness of the model overall. However, it is important to notice that the ANN model isolated did not performed better than all the considered conventional models, where gradient boosting classifier and random forest classifier were the models with the better performances.

For the second question, “2. What are the primary indicators and patterns that precede employee departures in business workplaces?” we have identified the specific features and patterns behind that most contribute for the employee turnover prediction and that helps us in make a clear assumption on the probability of leaving the company for the active employees, as these features are the ones that make a clear distinction between active and inactive employees. The primary indicators are the Number of Position Changes, the Monthly Remuneration, the Seniority, the number of years from the last team leader meeting (YearsLastTLMeeting) and the number of salaries increases. Being from the field of People Analytics, these outcome gives a deeper understanding of the problem at consideration, with facts and concrete observations, and the understanding on how powerful the turnover predictive models can be for a company as how impactful they can be. Employees with frequent position changes are more likely to leave. This suggests that job instability and lack of career progression contribute to higher turnover rates. Frequent shifts in job roles may indicate dissatisfaction or a search for better opportunities, leading to increased attrition. Longer tenure within the company is associated with lower turnover rates. Employees with higher seniority tend to stay longer due to established relationships, deeper organizational knowledge, and increased loyalty. This stability factor underscores the importance of career development and long-term engagement strategies, and where companies can act on the specific threshold for the different employees. Higher monthly remuneration is linked to lower turnover propensity. Competitive salaries and financial incentives play a crucial role in retaining employees. Adequate compensation reflects the value the organization places on its employees, thereby enhancing job satisfaction. The permutation importance analysis further corroborated these findings, showing that these features not only significantly impact turnover predictions but also have a stable influence across different model validations. The detailed examination of feature importance provides a robust statistical backing to these conclusions, emphasizing the need for strategic HR interventions focusing on these key areas. Additionally, visual tools such as partial dependence plots illustrated how variations in these features affect turnover probability. For instance, an increase in the number of position changes was clearly associated with higher turnover rates, while higher seniority and monthly remuneration demonstrated a protective effect against turnover.

For the last but not least question, the third “3. Can we accurately predict employee termination and termination probability using real company data?”, the answer is yes, accurately predicting employee termination and termination probability using real company data is feasible. However, it is important to address the model in each reality and company dynamics. The study demonstrates this through the application of advanced machine learning techniques, including ensemble methods and deep learning. By leveraging authentic and representative data from actual employees, the predictive models developed in the study showed high accuracy and reliability, and more important, high precision. The use of comprehensive data preparation, feature selection, and model validation processes ensured that the predictions were robust and applicable to real-world scenarios. This approach

underscores the importance of using real company data to capture the unique factors influencing turnover within a specific organizational context.

These insights are invaluable for developing targeted retention strategies. Organizations can enhance employee retention by focusing on job stability, competitive compensation, and fostering long-term career development. By addressing these critical factors, HR departments can mitigate turnover risks and maintain a stable and committed workforce.

For future research, it is recommended to extend the scope of this model to multiple organizations to validate its effectiveness and adaptability across different corporate cultures and industry settings. This broader application could help in fine-tuning the model to accommodate diverse employee demographics and organizational policies. Additionally, integrating more dynamic datasets, such as real-time employee feedback and ongoing performance evaluations, could enhance the model's predictive capabilities. These data sources would allow the model to capture the immediate effects of policy changes and management interventions on employee turnover intentions. This enhancement would allow for a more dynamic assessment of turnover risks, potentially introducing a predictive mechanism that not only forecasts who might leave but also when they might do so, so bringing an important component on termination study, timing. A proposed advancement is the development of a hybrid model combining regression analysis with classification techniques, offering a comprehensive tool that addresses both the probability and timing of employee turnover. Such techniques could uncover interactions that simpler models might overlook, offering a more comprehensive understanding of the factors driving employee departures. Lastly, a longitudinal study design could be implemented to track changes over time, providing insights into how employee turnover trends evolve with shifts in the economic climate, internal policy changes, or industry disruptions. This approach would be invaluable in understanding the long-term efficacy of retention strategies and HR policies. In conclusion, this thesis not only enriches our understanding of employee turnover but also sets the stage for more sophisticated analyses and practical applications in human resources management. The development of predictive models based on empirical data represents a significant advancement in the strategic management of human capital.

BIBLIOGRAPHICAL REFERENCES

Akasheh, M. A., Malik, E. F., Hujran, O., & Zaki, N. (2023). A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review. *Expert Systems with Applications*, *238*.

Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A. (2021). Employee Attrition Prediction Using Deep Neural Networks. *Computers*, *10*(11), 141.

Alsheref, F. K., Fattoh, I. E., & M. Ead, W. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. *Computational Intelligence and Neuroscience*, *2022*, 1–9.

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.

Chung, D., Yun, J., Lee, J., & Jeon, Y. (n.d.). *Predictive model of employee attrition based on stacking ensemble learning*.

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *International Journal of Organizational Analysis*, *28*(6), 1273–1291.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, *9*(4), 86.

Han, J. W. (2022). A review of antecedents of employee turnover in the hospitality industry on individual, team and organizational levels. *International Hospitality Review*, *36*(1),

Kang, I. G., Croft, B., & Bichelmeyer, B. A. (2021). Predictors of Turnover Intention in U.S. Federal Government Workforce: Machine Learning Evidence That Perceived Comprehensive HR Practices Predict Turnover Intention. *Public Personnel Management*, *50*(4), 538–558.

Mohammed, A. Q. (2019). HR ANALYTICS: A MODERN TOOL IN HR FOR PREDICTIVE DECISION MAKING. *JOURNAL OF MANAGEMENT*, *10*(3).

Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B. (2023). Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. *Benchmarking: An International Journal*, *30*(10), 4140–4173.

Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. *Mathematics*, *9*(11), 1226.

Naz, K., Siddiqui, I. F., Koo, J., Khan, M. A., & Qureshi, N. M. F. (2022). Predictive Modelling of Employee Churn Analysis for IoT-Enabled Software Industry. *Applied Sciences*, *12*(20), 10495.

Tian, X., Pavur, R., Han, H., & Zhang, L. (2023). A machine learning-based human resources recruitment system for business process management: Using LSA, BERT and SVM. *Business Process Management Journal*, *29*(1), 202–222.

Wang, X., & Zhi, J. (2021). A machine learning-based analytical framework for employee turnover prediction. *Journal of Management Analytics*, *8*(3), 351–370.

Wild Ali, A. B. (2021). Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm. *Wireless Personal Communications*, *119*(4), 3365–3382.

Mazzetti, G., & Schaufeli, W. B. (2022). The impact of engaging leadership on employee engagement and team effectiveness: A longitudinal, multi-level study on the mediating role of personal- and team resources. *PLOS ONE*, *17*(6), e0269433

Schneider, B., & Pulakos, E. D. (2022). Expanding the I-O psychology mindset to organizational success. *Industrial and Organizational Psychology*, *15*(3), 385–402.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa