A Real-Time Employee Attrition Prediction and Risk Scoring System

By

Martin Mwangi Kariuki

138536

Master of Science in Data Science and Analytics

2025

A Real-Time Employee Attrition Prediction and Risk Scoring System

By

Martin Mwangi Kariuki

138536

Submitted in Partial fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

July, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Martin Mwangi Kariuki

Sign:  Date: 25th – March – 2025

Approval

The dissertation of Martin Mwangi Kariuki was reviewed and approved for examination by:

Dr. Henry Muchiri

School of Computing and Engineering Sciences (SCES)

Strathmore University

28-03-2025

Abstract

Human Resource (HR) analytics is increasingly being explored around the globe for its potential in addressing employee attrition. Globally, the rate of attrition has been estimated to be about 25% higher in comparison to the pre-pandemic era. The effects of employee attrition including the loss of valuable talent and incurring costs for recruitment and on-boarding of new talent has been felt by companies in different sectors globally. Previous studies have made considerable efforts in not only understanding the concept of employee attrition but also in its early detection. This study aims to advance previous research by moving beyond merely identifying an effective machine learning technique to implementing the model that enables the human resource team to understand and assess employee attrition risk in real-time. This study provides a focus on three specific objectives that utilize human resource analytics approaches to understand the concept of attrition. Firstly, the study aims to use statistical approaches to analyze and identify the factors influencing employee attrition. Secondly, it aims to evaluate the effectiveness of machine learning algorithms in predicting employee attrition. Ultimately, the development of a system that predicts employee attrition and generates risk scores in real-time using relevant HR data marks a pivotal milestone for this study. Generalized Linear Model with interaction terms is the statistical approach which was utilized to assess the contributors of employee attrition. Job satisfaction, job involvement, years at company and monthly income were statistically significant thus are attributed to an employee's decision to quit or stay. In this study, a performance evaluation and comparison of XGBoost, Random Forest (ensemble techniques) and Support Vector Machine, K- Nearest Neighbors as well as LogisticRegression machine learning models was conducted. Leveraging the employee records from the IBM dataset, Random Forest outperformed all the other models with an Accuracy of 80%, Precision of 91%, Recall of 85% and F1 Score at 88%. Insights from the first two research objectives were used to develop a real-time employee attrition and risk scoring tool. The solution provided under this study can be utilized in companies to provide data driven insights on attrition of their employee base. This study provides invaluable insights that can be used by various stakeholders including but not limited to, companies, data solution providers and the government to provide proactive measures to address attrition such as salary adjustment and management of employee work involvement. In conclusion, this study has contributed to the various on-going human resource analytic research which can be incorporated within organizational systems to address employee attrition and reduce costs incurred in recruitment and training of new talent.

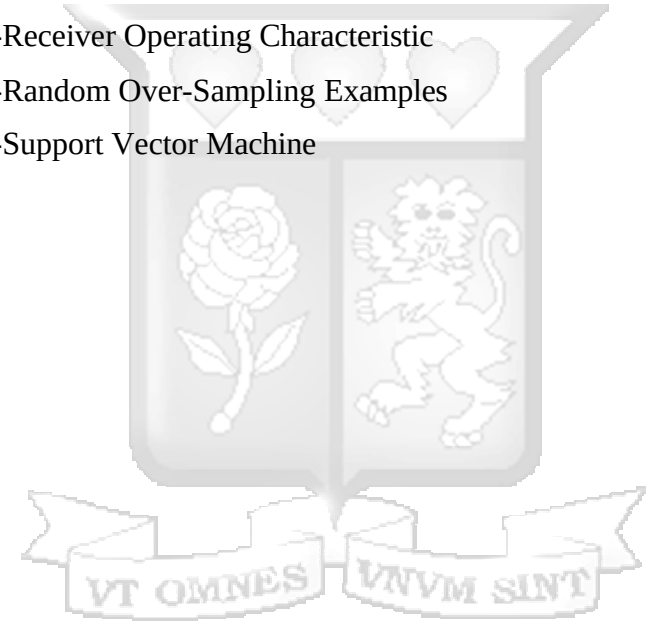
Keywords: Attrition, Human resource, Machine learning, Risk Scoring.

Table of Contents

List of Figures

List of Abbreviations

AUC	- Area Under ROC Curve
CLARA	- Clustering for Analysis and Remedial of Attrition
EDA	- Exploratory Data Analysis
EWS	- Early Warning System
HR	- Human Resource
KNN	- K-Nearest Neighbors
ML	-Machine Learning
RFE	- Recursive Feature Elimination
ROC	-Receiver Operating Characteristic
ROSE	-Random Over-Sampling Examples
SVM	-Support Vector Machine



Acknowledgement

I begin by acknowledging the Grace of God, which provided the strength and opportunity to complete this study. I am sincerely grateful to everyone who supported this endeavor. I would particularly like to express my appreciation to Dr. Henry Muchiri, whose expertise and sustained guidance were crucial to the project's development.

Finally, I acknowledge the invaluable contributions of fellow researchers, and authors whose work provided the essential foundation and inspiration for this research.

Dedication

With gratitude to Jesus Christ for His strength and guidance, I dedicate this dissertation.

To my family and friends, whose unwavering support sustained me throughout this journey. A special thank you to Lewis Mwangi, for his exceptional support and encouragement.

To my mentors, classmates, and Strathmore University, for their invaluable guidance.

To my future wife, for her support, and to my future children, may this work inspire them to pursue their dreams.

Chapter 1: Introduction

1.1 Background of the study

Employee attrition refers to a situation where an employee leaves their company due to various reasons including resignation and voluntary retirement (Alsheref et al., 2022). The gradual reduction of employees has various implications to a company including lowered efficiency, productivity and customer service (Mahtre et al., 2020). Organizations might also suffer financial loses and incur more cost in recruiting, training and onboarding of new talent (Nuel et al., 2022). Therefore, it is crucial for organizations to retain key employees who hold and add value in form of tactic expertise and highly knowledgeable in running business (Jiang et al., 2012). To address attrition, researchers as well as human resource professionals within contemporary organizations have given priority in understanding the key drivers that are attributed to employees' departure. Numerous studies have explored the key motivators that influence an employees' decision to quit through statistical and qualitative approaches. These attributes that explain the increase in employee attrition in organizations include autonomy, work-family balance, future opportunities as well as ability to work from home (Whitton, 2023). Researchers have also contributed to advancement in talent management strategies through employee attrition prediction leveraging machine learning techniques. While considerable efforts have been made by existing studies to understand the drivers of attrition and further detect employees who are likely to leave, there has been little efforts made in the development of tangible solution for data driven talent management strategies. This study aims to advance previous studies by moving beyond merely identifying an effective machine learning technique to implementing a model that enables the human resource team to understand and assess employee attrition risk in real-time. Employeeattrition prediction, also referred to as people analytics, is the detection of the intentions of an employee on staying or leaving a company (Yahia et al., 2021). It plays a major role in the contemporary organizations assisting them to reduce the rate at which employees leave throughdata driven insights.

Employee attrition has been quite a challenge in various industries across the world. In the 2024 report by the Bureau of Labor Statistic (BLS) the number of employees who quit their jobs in the U.S between July 2023 and July 2024 is estimated to be around 3.3 million. The surge in employee departure has motivated the numerous researches around understanding the key factors that influence employee attrition. In a global survey conducted by De Smet et al.

(2022), lack of career progression, inadequate pay, work conditions, poor leadership, geographical demands and non-inclusive community were the major contributing factors for employees leaving which is common to various organizations in the U.S, Australia, Singapore, U.K, Canada and India.

In Africa, high attrition rates have been recorded in several industries across various countries in the region. This has led to various research geared towards understanding the factors attributed to employee departure in the various industries. For instance, studies show that in South Africa, poor working environment, low pay, high workload as well as autonomy are among other key drivers of employees quitting (Mpundu et al., 2023). Organizations and the government have been urged to review their policies to ensure that attrition rates are reduced since most industries contribute to the economic growth within the countries within African region. One of the initiatives that the government should be taken include enforcing policies that align with the companies within their jurisdiction. On the other hand, organizations should place a focus on retention strategies that enhance job satisfaction among their employee base (Muda et al., 2022).

In Kenya, various sectors including the energy, telecommunication and healthcare have faced the complexities brought about by the gradual reduction of valuable professionals. Several studies conducted in Kenya have revealed a number of factors that are attributed to employee attrition. According to Mwendwa (2017), lack of career growth opportunities, stressful working conditions, performance rewards and payment plan are among the key motivators that are drive professionals to quit. Locally, addressing attrition is paramount as sectors which contribute to the overall economic growth like the energy industry records up to 50% attrition rates every year (Mwendwa, 2017). It is therefore crucial that data driven talent management strategies are incorporated in organizations in order to reduce the number of professionals leaving their jobs which in the long run affects the economic development in the country.

1.2 Problem Statement

While considerable efforts have been made in past research to determine algorithms that effectively predict attrition, this study takes this a notch higher by developing a data product that utilizes the most effective model determined to predict attrition and providerisk scores of attritions which allows proactive decision-making and targeted interventions to retain valuable talent. The rate of attrition globally is alarming across key sectors that contribute to the growth of aneconomy including financial, education and healthcare sectors. In a global lens, attrition rates have been on the rise and is estimated to be 25% higher than the pre-pandemic era (De

Smet et al., 2022). The loss of valuable professionals has seen organizations incur the costs of recruitingand training new personnel (Holtom et al., 2006). The Society for Human Resource Management (SHRM) highlight that as of April 2022, the cost of recruiting new talent was estimated to be around four thousand seven hundred dollars per hire. Such costs and financial loses can be reduced by implementing employee retention strategies incorporating the most sort for job satisfaction components (Muda et al., 2022).

To address employee attrition, this study will apply data driven approaches to explore and understand the root cause of employee departure and design a tangible solution that can predict employee attrition. As an advancement to the previous studies that explored a limited set of variables, this study seeks to utilize historical employee data from IBM to understand the key drivers of employee attrition and provide a holistic view of these factors and develop a more integrated model. Notably, previous studies have also focused on predicting employee attrition using machine learning models, they have largely overlooked the development of a structured risk scoring mechanism to quantify attrition likelihood. This gap limits the ability of HR professionals to prioritize interventions effectively, highlighting the need for a real-time system that not only predicts attrition but also provides actionable risk scores. We utilize this data as it is reliable for the operationalization of this study's research objectives. The insights from this study will equip various stakeholders with the knowledge that can be used to make data-driven decisions pertaining the issues around addressing attrition. The data product under this study can be seamlessly integrated and utilized by organizations to optimize employee retention measures and proactively address attrition.

1.3 Research Objectives

1.3.1 General Objective

The general objective of this research is to develop a real-time employee attrition prediction and risk scoring system.

1.3.2 Specific Objectives

1. To use statistical approaches to analyze and identify the factors influencing employee attrition.
2. To evaluate the effectiveness of machine learning algorithms in predicting employee attrition.
3. To develop a machine learning model to predict employee attrition and provide risk scores.

1.4 Research Questions

1. What are the factors that influence employee attrition?
2. Which machine learning algorithm performs best in predicting employee attrition?
3. How can we develop machine learning model that predicts employee attrition and generate risk scores?

1.5 Significance of the study

The significance of this study extends to various stakeholders within the organizational context. Firstly, the employers can be able to leverage on the insights from this study to ensure that they reduce costs associated to employee attrition which include recruitment costs, training cost and onboarding costs (Nuel et al., 2022). Employee attrition also poses big challenge and risk to companies as it affects not only the continuity of their plans but also their overall productivity (Yahia et al.,2021). By operationalizing the first research objective under this study, that is, to understand the factors that influence employee attrition, organizations can implement proactive strategies that enhance retention among their employee base thus reducing the associated costs of losing valuable talent. By comprehensively understanding the key drivers of attrition, employers can establish measures to ensure workforce continuity and stability. This study will also contribute to an advancement in the use of machine learning techniques to predict attrition which provides valuable insights for future practical applications within the human resource sectors. The development of an employee attrition detection and risk scoring system is the ultimate goal for this study. The tangible solution under this study is aimed to assist organizations in making data-driven decisions to proactively address employee attrition and optimize employee retention strategies. The human resource professionals within various companies are able to utilize the findings of this study to enhance retention by providing targeted interventions to employees with high chances of quitting. The study equips the human resource professionals with a comprehensive framework that allows them address employee attrition and enhancing talent management through data driven enhanced decision making. From a solid understanding of the key drivers of employee attrition to the practical solution that this study provides, the human resource professionals can take proactive steps towards retaining valuable talent within their organizations. This way the overall performance within the company is not affected, as loss of expertise leads to a decline in customer service, poor communication and coordination of activities thus affecting overall corporate performance (Holtom et al., 2006).

On the other hand, Mwangi (2019), highlights the role of not only the employers but also the government in addressing employee attrition. A detailed understanding of the drivers of employee attrition is crucial for the government as their policies affect the companies in a given country. Insights from the study can inform public policies stipulated that affect the professionals from various organizations thus enhancing stability. The government needs to always review its policies to ensure that it provides positive outcomes in market firms which, in turn, helps reduce attrition among professionals (Mwangi, 2019).

Ultimately, the data solution providers for various companies can utilize the robust framework provided in this study towards employee retention. From the statistical perspective of the factors influencing employee attrition and analytical aspects to data product development, the data experts can benchmark and utilize the insights drawn from this study to come up with solutions that organizations can seamlessly integrate and implement optimized talent management plans.

1.6 Scope of the study

This study's insights on the factors influencing employee will rely on IBM's historical employee data. We provide a focus of these attributes based on the features provided in the dataset.

1.7 Limitation of the study

While there are so many reasons that employees consider before leaving an organization, this study is limited to the variables provided in the data.

Chapter 2: Literature Review

2.1 Introduction

This chapter seeks to explore work done by other researchers providing a key focus on the concept of human resource analytics in as far as talent attrition is concerned. An overview of employee attrition, contextual background, the theoretical review, empirical review as well as research gaps identified will be discussed.

2.2 Employee Attrition: An Overview

Employee attrition refers to a phenomenon where an employee leaves a company retired or by voluntarily resigning leading to a decrease in the number of active employees (Alsheref et al., 2022). Employee attrition poses a big challenge and risk to companies as it affects not only the continuity of their plans but also their overall productivity (Yahia et al.,2021). According to Alsheref et al. (2022), the impact of losing employees is immense as this lowers efficiency of the business and it hinders the smooth progression of long-term strategies. Furthermore, loss of expertise leads to a decline in customer service, poor communication and coordination of activities thus affecting overall corporate performance (Holtom et al., 2006). For organizations, the impact of employee attrition is deemed costly since there will be need to spend more on acquiring new talent and preparation to join their teams (Alsheref et al., 2022). In an effort to evaluate the real cost of recruitment, The Society for Human Resource Management (SHRM) highlight that as of April 2022, the cost of recruiting new talent was estimated to be around four thousand seven hundred dollars per hire. This has seen employee attrition research being prioritized in various industries.

2.3 Contextual Background

2.3.1 Global Context

Several studies have explored the impact of employee attrition from a global lens. Globally, 40% of employees are likely to quit their jobs in the near future (De Smet et al., 2022). Attrition rates have been on the rise and is estimated to be 25% higher than the pre-pandemic era (De Smet et al., 2022). According to The Job Openings and Labor Turnover Survey (JOLTS), every single month the estimated number of employees leave their organizations in the U.S ranges from 3 to 4.5 million employees. In the McKinsey report provided by De Smet et al. (2022), they highlight that in the US the number of job openings had gone significantly up from 9.3 million in April 2021 to 11.3 million due to voluntary attrition. In an effort to understand the

factors influencing employee attrition globally, De Smet et al. (2022) conducted survey in various industries in Australia, Singapore, United States, Canada, United Kingdom and India. Through the global survey, several factors revealed as the most impactful in explaining attrition are illustrated in figure 2.1.

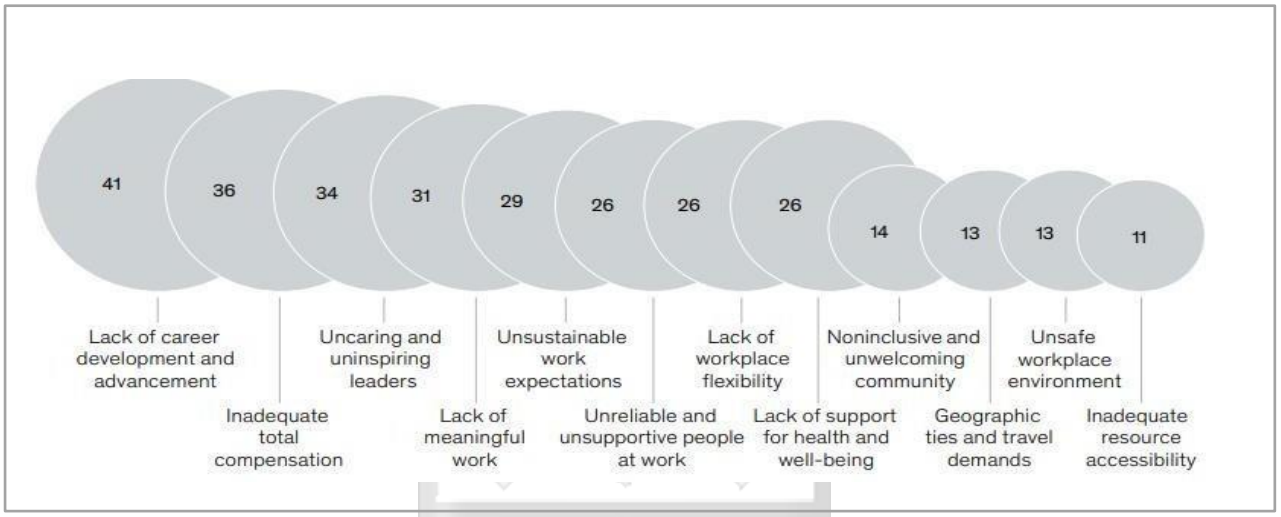


Figure 2.1 Top attributes associated to employee attrition, % (De Smet et al., 2022)

In a study to explore the push and pull factors among Malaysian young professionals, Ho et al. (2010), identified interference with work-family lifestyle as the top demotivator (push factor) that leads to attrition. On the other hand, compensation and benefits were identified as the major pull factors that encourage the young professionals to join another organization (Ho et al., 2010). In India, particularly in the hospitality industry, the attrition rate has been spiking at alarming rate of 10% per year (Gangai, 2013). Records show that in India, between 2010 and 2011, there was a significant increase in attrition rates to 50% up from 25% (Gangai, 2013). In an effort to understand the attrition rates spikes, Gangai (2013), found out that most employees leave due to various push and pull factors including, better opportunities, compensation issues, departmental challenges as well as salary matters.

2.3.2 Regional Context

There are various existing studies that have shed light on the drivers of employee attrition in the African region. Firstly, in an aim to understand the factors influencing attrition among professionals in Gambian public and private sectors between the year 2007 and 2017, Kanteh and Gibba (2019), conducted a survey and performed quantitative analysis to assess the various drivers of attrition. Their study revealed that inadequate pay, lack of career advancement as well as unmet job expectations were the major drivers of attrition. The impact associated with employee attrition which employers reported and quantified in the study conducted Kanteh and

Gibba (2019), are distributed as follows: a 50% reduction in productivity, a 16.7% impact on ongoing projects, and 30% allocated to training costs. This shows that employers should take necessary precautions to help reduce employee attrition by ensuring that they provide avenues for career growth as well as better pay for their employee base (Kanteh & Gibba, 2019).

Secondly, over the years there has been a spike in employee attrition within the education sector in South Africa (Mpundu et al., 2023). This has been associated with poor working environment, low pay, high workload as well as autonomy among other drivers. An intriguing discovery about the study conducted by Mpundu et al. (2023) is that the major cause for the spike in educators' attrition rates shifts from employer related challenge to a government issue, that is, the enforced pension funding system. Employee attrition is therefore an issue that all stakeholders should come together to ensure that the welfare of employees is guaranteed. Price Water Coopers reported that in South African banking sector the employee attrition rates wereat 23% as of 2018. Iwu et al. (2012), conducted a study to investigate and understand the key factors that enhances retention among South African healthcare professionals. The researchers reported that working conditions, fairness in performance evaluation, self-efficacy, clarity of job responsibilities and trust in leadership are most significant factors that healthcare professionals in South Africa consider when deciding to leave or stay. These reports from existing studies show that various sectors in South Africa are facing the problem of employees quitting and that the demands of these professionals in those sectors are quite similar.

In Ghana, a case study conducted by Muda et al. (2022), revealed that there is direct relationship between the components of job satisfaction and an employee's intent to stay or leave an organization. Some of the components of job satisfaction components highlighted include, better pay, promotion, avenues for better career prospects, better conditions of service as well as good employee-supervisors relationship (Muda et al., 2022). Since the financial sector contributes to about 14.5% of the Ghanian economy, there is need to nurture employee retention strategies incorporating the most sort for job satisfaction components (Muda et al ., 2022).

In Nigeria, employee attrition has been quite a challenge in various sectors including health, education and manufacturing sector (Nuel et al., 2022). In the study to investigate the impact of attrition on performance within the manufacturing sector, Nuel et al. (2022), revealed that employees leaving affects the overall productivity. The major elements considered in any type of attrition whether voluntary or non-voluntary are the level of education as well as working

conditions in the organization (Nuel et al., 2022). Some of the adverse effects of losing employees highlighted by Nuel et al. (2022) include, reduced productivity, training costs, stalling of on-going projects and high recruitment costs. In light of these consequences of employee attrition, employers need to ensure that their employee base work in good conditions to enhance performance and help reduce attrition (Nuel et al., 2022).

2.3.3 Local Context

In Kenya, various sectors are affected by the adverse effects of employee attrition. This fact has seen researchers explore the underlying causes, implications and potential solutions to thisissue. In the major telecommunication industries in Kenya, that is Safaricom, Airtel and Telkom, employee performance is directly related to the intent of professionals to leave (Chepkirui & Atambo, 2024). To enhance performance and retain talent, the organizations should ensure that they compensate professionals based on market rates, invest in better and functional infrastructure and ensure that employees are conversant with their job role (Chepkirui & Atambo, 2024). In a different study conducted by Mwangi (2019), the major causes of attrition within the telecommunication industry in Kenya include, discontent with jobrole, lack of career growth opportunities, poor working environments and unreasonable set performance goals. Mwangi (2019), highlights the role of not only the employers but also the government in addressing employee attrition. The government needs to always review its policies to ensure that it provides positive outcomes in market firms which, in turn, helps reduce attrition among professionals (Mwangi, 2019). On the other hand, employers should take pre-emptive measures to ensure that their employee base get better working conditions and career advancement opportunities (Mwangi, 2019).

In the energy industry, the issue of employee attrition has been a challenge as well. With attrition rates of about 50% every year, there is need to understand the factors influencing attrition among call centers professionals within the energy industry (Mwendwa, 2017). Mwendwa (2017), conducted a study to explore the factors influencing employee attrition in the energy sector particularly Kenya Power call centers. The study revealed that supervisor support, performance rewards, career growth, job role characteristics and training were the major drivers of attrition among call center agents within the energy sector. While call centres are often associated with stressful work environments, it is crucial for employers to provide improved working conditions to help reduce employee attrition (Mwendwa, 2017). This can include offering better support systems by supervisors, opportunities for career growth, clear

and realistic goals and training (Mwendwa, 2017). By doing so, organizations can nurture a more positive and sustainable workplace, ultimately lowering employee attrition rates.

In the healthcare sector, the major causes of attrition include retirement, voluntary resignation and death of employees (Chanvoka et al., 2009). In the study conducted by Chanvoka et al. (2009), attrition rates in provisional, district, and other healthcare centres were attributed to retirement, accounting for 48% to 58%, voluntary resignation, contributing 25% to 40%, and death making up to 9% to 17% of total attrition. The researchers advocate for employers to implement policies that reduce voluntary resignation such as improving the working conditions of healthcare professionals within public sectors as well as reviewing their salaries.

2.4 Theoretical Review

2.4.1 Push-Pull-Mooring (PPM) Framework

The Push-Pull-Mooring (PPM) Framework has served as theoretical lens to aid in understanding the factors that influence employees' decisions to leave an organization (Fu, 2011). The PPM model is one of the most widely recognized and utilized in studying human migration within various contexts including within organizations (Haldorai et al., 2019). In an aim to understand the career commitment of professionals in the IT industry, Fu (2011), highlighted the career satisfaction as the most impactful determinant. From a perspective of the PPM framework, the employees in different career stages had different attitude towards their commitment to their jobs. Senior professionals were majorly influenced by push effects (professional obsolescence and career satisfaction) whereas the junior professionals were primarily driven by mooring (professional self-efficacy and career investment) and push factors (Fu, 2011). In a study conducted by Whitton (2023), the push and pull factors that explain the surge in employee attrition by life stage include autonomy, work-family balance, future opportunities as well as ability to work from home. Within the hotel industry, some of the push factors that strongly leads to employee attrition include lack of career advancement, high workload, issues with work-family balance, emotional labor as well as interpersonal conflict (Haldorai et al., 2019). On the other hand, social status, community fit as well as the travel opportunities are the pull factors attributed to lower attrition rates. Personal involvement is a mooring effect that moderate attrition especially for medium term employees (Haldorai et al., 2019). Globally, several push pull factors were identified in a survey conducted by De Smet et al. (2022) which are deemed important for employers to understand especially if they are invested in retaining their employee based. Figure 2.3 indicates the push and pull factors that

are associated to the intent of an employee leaving or staying as established from the global survey conducted by De Smet et al. (2022).

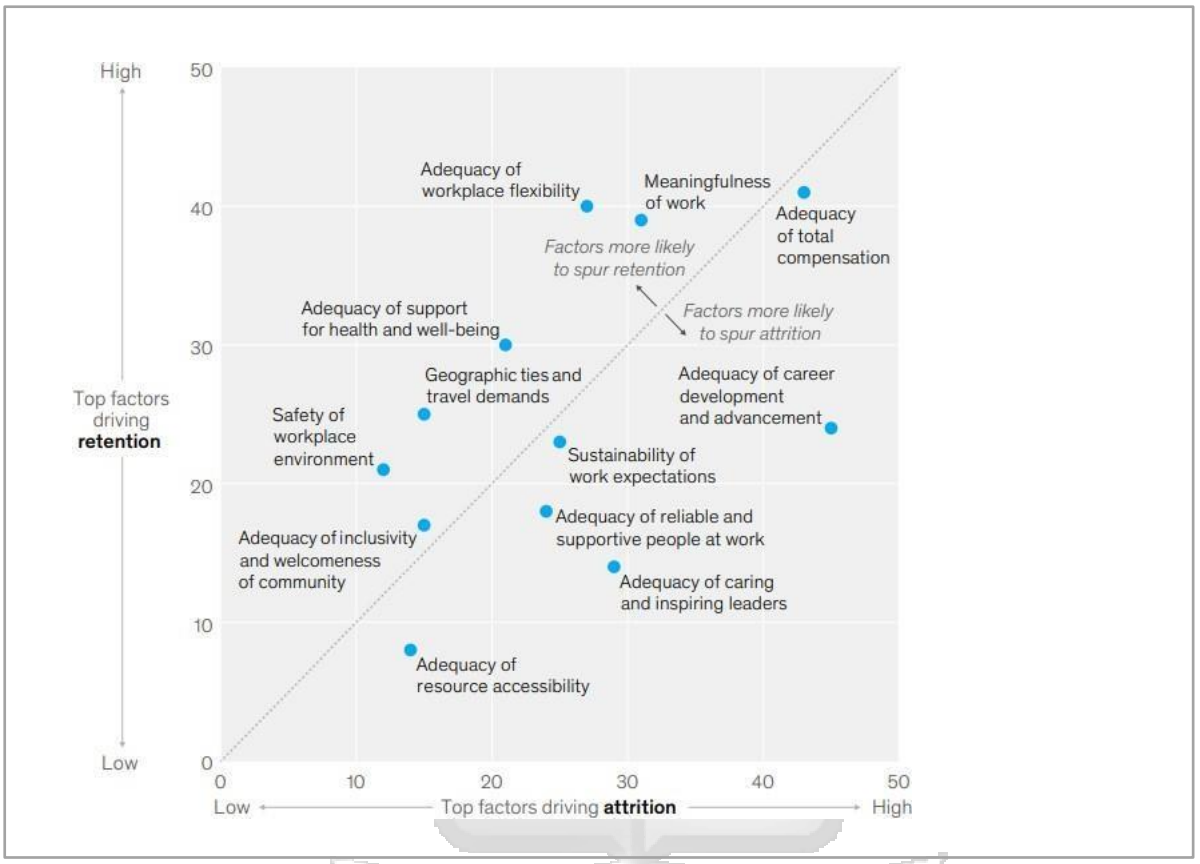


Figure 2.2 Push-Pull Factors Affecting Employees Globally, % (De Smet et al., 2022).

2.4.2 Job Embeddedness Theory

Job embeddedness theory encompasses critical aspects that can be dissected into three dimensions, that is, fit (compatibility with an organization/community), sacrifice (perceived costs and benefits) and links (interactions within the organization/community) (Holtom et al., 2006). This theory provides a focus of the retention elements and strategies. It encompasses both on-job and off-job factors that influence an employee to stay or leave their jobs (Jiang et al., 2012). The value of the concept of job embeddedness has been demonstrated in various organizations with efforts to retain their workforce. It is substantial for organizations to retain key employees who hold and add value in form of tactic expertise and highly knowledgeable in running business (Jiang et al., 2012). Every organization might be required to assess the unique needs of their employees as this would enable the implementation of strategies that lower chances of attrition (Holtom et al., 2006). Low embeddedness (organizational fit,

sacrifice and links) increases the chances of employee attrition within an organization (Jiang et al., 2012). Companies need to embrace the impact of investing in employee retention rather than incur costs of turnover such as that of training new employees, reduced production and customer service (Holtom et al., 2006). The cost of acquiring new talent in the hotel industry in United States has been estimated to range from six thousand to twelve thousand dollars (Jiang et al., 2012). This underpins the importance of leveraging the on-job and off-job factors that affect the unique employee base of an organization to mitigate such costs. According to Narayanan (2016), Human Resource department should evaluate the impact of job embeddedness on retention of their high value employees as well as it's input towards talent management strategies for retention. Figure 2.1 depicts the link between talent management strategies and an employee's intention to stay or leave with the various dimensions of the theory of job embeddedness.

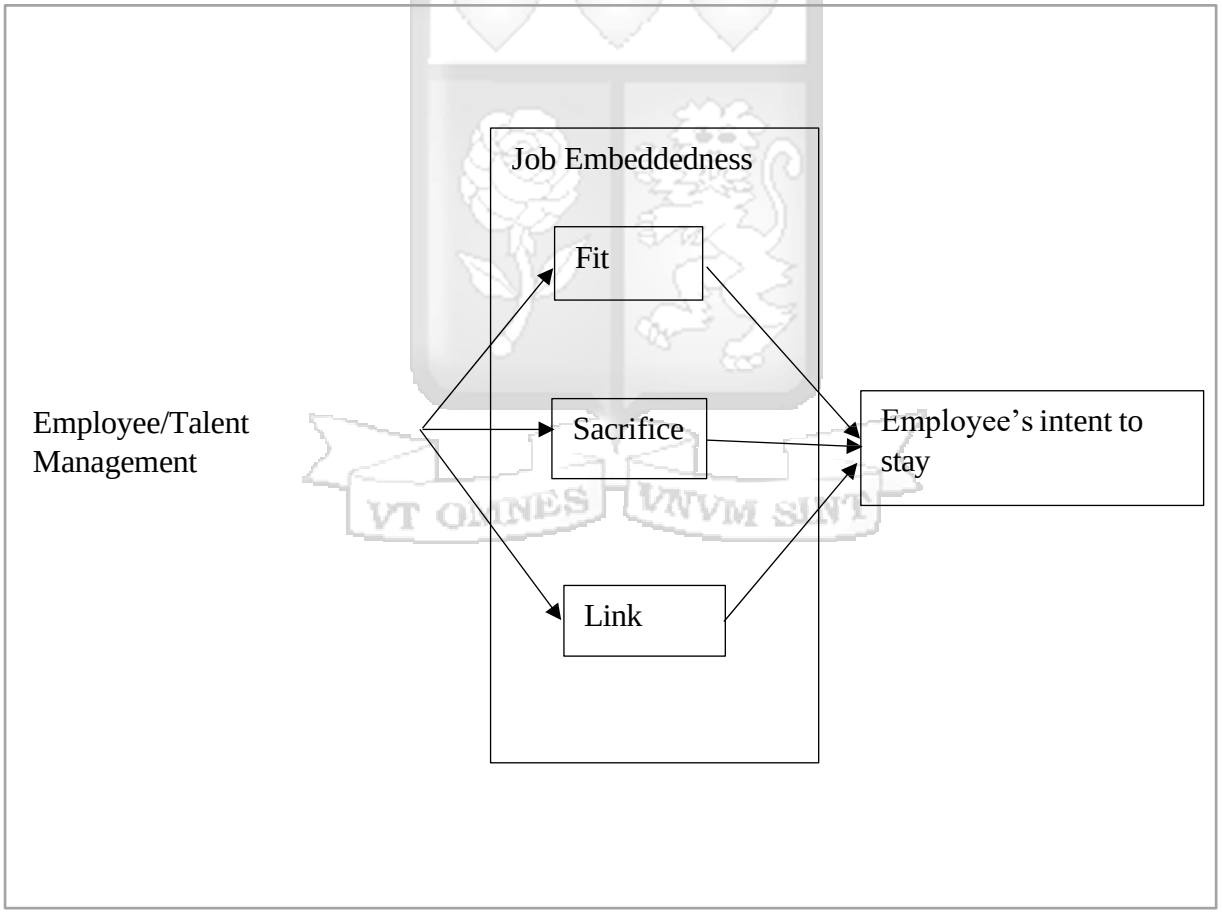


Figure 2.3: Job Embeddedness Theory (Narayanan, 2016)

2.5 Empirical Review

2.5.1 Factors Influencing Employee Attrition

Past studies have utilized both utilized statistical and qualitative approaches to understand the factors that influence employee attrition. In an effort to understand the reasons as to why employees quit in both public and private sectors within organizations, Frye et al. (2018) employed logistic regression coefficients to determine the attributes that are statistically significant. According to the study carried out by Frye et al. (2018), the length of service provided by an employee is the most impactful factor associated to their likelihood to quit. The odds of an employee quitting significantly reduces as their tenure within an organization increase. The odds of an employee leaving also increases or decreases depending on their ageas well (Frye et al., 2018). Employees on a standard payment plan are also less likely to quit (Frye et al., 2018). The attributes age, tenure, rewards and payment have also been highlightedin a study conducted by Yahia et al. (2021) as factors influencing employee attrition within organizations. Following rigorous feature selection using selectKbest and Recursive Feature Elimination (RFE), Yahia et al. (2021) also indicated that other factors associated to the likelihood of an employee leaving a company include, marital status, job satisfaction, grade, environment satisfaction, business travel as well as job involvement. Their study places a focus on business travel and rewards as features with the highest importance and a motivating factor for retention strategies implementation since these attributes have been less common in previous related works. Organizations should focus on retention of their employees by nurturing a conducive work environment as well as ensuring they utilize data driven insights to mitigate the risk they might face in the event of employee attrition (Yahia et al., 2021).

In research conducted by Raza et al. (2022), Employee Exploratory Data Analysis came into play in determining the factors that influence an employees' decision to leave an organization. In their exploration, age, hourly rates, job level and monthly income were the main factors that influence employee attrition. In a study conducted by Srivastava and Eachempati (2021), Multiple Linear Regression Model was utilized to understand employee attrition through evaluating the contribution of various factors. This technique does not only allow one to evaluate the contribution of a feature on their own but also as a set of features, that is, impact of interaction of features on the outcome (attrition) as this is what is expected in the real-world scenario that various factors might lead to one quitting their job (Srivastava & Eachempati, 2021). The appraisal rating, employee satisfaction and level of annual renumeration and

number of tasks assigned to an individual are factors that are considered to simultaneously interact (interplay) thus playing a major role in an employees' decision to quit (Srivastava & Eachempati, 2021).

2.5.2 Predictive Analytics in Employee Attrition

Predictive analytics has been applied in various studies in an effort to determine the likelihood of an employee quitting their work. Yahia et al. (2021) refer to this as people analytics which plays a major role in the contemporary organizations assisting them to reduce the rate at which employees leave through data driven insights. Predictive analytics is meant to detect the intentions of an employee on staying or leaving a company (Yahia et al., 2021).

Yahia et al. (2021) applied deep, ensemble and machine learning models to predict employee attrition on three different datasets (two simulated datasets and one real dataset). Precisely, the deep learning predictive models utilized were, Deep Neural Networks (DNN), Long-Short Term Memory Networks (LSTM) and Convolution Neural Networks (CNN). For the ensemble techniques, XGBoost, Random Forest, Voting Classifier and Stacked ANN based models were used in their study and as for machine learning models, Logistic Regression, Decision Trees as well as Support Vector Machines (SVM) were utilized in predicting attrition. Upon evaluating the performance of the models on the three datasets Voting Classifier outperformed all the other models with a record of 96%, 98% and 99% accuracy on the three data sets utilized respectively (Yahia et al., 2021). In another study conducted by Frye et al. (2018), Logistic Regression model outperformed K-Nearest Neighbors and Random Forest when it came to predicting employee attrition effectively. According to their study, Logistic Regression recorded the highest accuracy with greater than 74% success rate. The findings from study done by Srivastava & Eachempati (2021), revealed that DNN outperformed gradient boosting and random forest algorithms in predicting attrition with an accuracy rate of 91.6%.

In a study conducted by George et al. (2022), the predictive algorithms employed to detect attrition included, Extra Tree classifier, AdaBoost, Random Forest, Gradient Boosting Classifier and XGBoost. The performance evaluation techniques used include precision, accuracy as well as recall. The Extra Tree classifier which employs the concept of integrating outcomes from several un-correlated trees connected in a forest to provide optimal result, outperformed the other models with 97% accuracy, precision and recall (George et al., 2022). According to Raza et al. (2022), the most effective model in predicting employee attrition was

also the Extra Tree Classifier with an accuracy rate of 93%. This model outperformed the Support Vector Machine, Decision Tree Classifiers as well as Logistic Regression.

With the aim of development of a tool that detects employee attrition as well as provides retention measures, Brockett et al. (2019), utilized the best performing model that is, K-Means clustering algorithm and frequent pattern mining scores. The study evaluated the effectiveness of other algorithms such as Random Forest, XGBoost, Support Vector Machine and spectral clustering. The tool developed under the study utilized the function with a linear combination of the clustering scores from K-Means clusters as well as frequent pattern mining scores which recorded an accuracy rate of 65% (Brockett et al., 2019).

In a study conducted by Fallucchi et al. (2020), the models evaluated for their effectiveness in predicting employee departure included, the Bernoulli Naïve Bayes, Gaussian Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Tree, Support Vector Classifier and Linear Support Vector Classifier. Their evaluation identified the Gaussian Naïve Bayes algorithm as the most effective model in predicting employee attrition with sensitivity (recall rate) of 54%. Poornappriya & Gopinath (2021), indicated that Neural Network performs best in predicting employee attrition with an accuracy of 91%, sensitivity (recall rate) of 89.7% and specificity of 93.3%. In their study Support Vector Regression, Decision Tree Regression, Random Forest Regression, as well as the Neural Network Regression were all assessed. In a different study by, Adeusi et al (2024), a key focus on predicting attrition was placed on employees who work in high stress environments. The study utilized Logistic regression, Decision Trees, Random Forest and Neural Networks where the Random Forest and Neural Networks outperformed the other two models in predicting attrition.

2.5.3 Talent Attrition Risk Scoring System

The development of tools to aid in detection of attrition has enhanced talent retention within various contemporary companies and this consequently lowers their operational costs. Brockettet al. (2019), were able to come up with a solution that allows HR catch the employees at high risks of attrition. The tool, Clustering for Analysis and Remedial of Attrition (CLARA), usesa scoring function based on clustering and frequent pattern mining (Brockett et al., 2019). K-Means clustering technique was implemented to group employees based on the feature similarities they exhibit. Frequent pattern mining aids in the analysis of features of the

employees that have already exited the organization such that an active employee with higher similarities to the extracted patterns records high score of attrition (Brockett et al., 2019). The score from CLARA, that is a linear combination of frequent pattern score and clustering score,is used to predict the likelihood of an employee staying or leaving the company. The solution developed also provides measures to be taken to enhance retention (Brockett et al., 2019).

Mhatre et al. (2020) also came up with a solution that allows organizations to predict attrition and segregate the employees into two categories, high risk and low risk employees. The development of this tool, referred to as an Early Warning System (EWS), began with understanding data of about 12000 employees to uncover the factors that influence employee attrition. Mhatre et al. (2020) then proposed and evaluated the effectiveness of five models in predicting employee attrition that is, K- Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree and XGBoost. XGBoost turned out to be the best model and was deployed in the EWS. The data solution has the ability to classify employees using a red, blue and green indicators where red implies those at high risk and green implies those at low risk (Mahtre et al., 2020). With the system, organizations have the ability to identify employees at high risk of attrition and provide timely retention measures to ensure they are not losing their most valuable and knowledgeable talent (Mahtre et al., 2020).

2.6 Research Gap

The review of past studies has brought to the forefront the issues that are associated with the concept of employee attrition which has been a challenge for most organizations. It has also revealed significant gaps that need to be addressed in as far as data-driven approaches for employee retention is concerned. When we look into the concept of understanding the factors attributed to employee attrition, studies took various statistical approaches to get insights from the data they utilized. Despite the significant effort to uncover the factors influencing attrition by Yahia et al. (2021), Raza et al. (2022) and Frye et al. (2018), we note there is a limited set of variables being explored that include age, period of service, work environment as well as payment plans. Some of the implications of having limited data include, reduced model flexibility as well as underfitting therefore integration of a wide range of relevant features to this study is paramount to mitigate the constraints brought about by utilizing a limited set of variables. This study will thereby provide a key focus on various variables in an effort to bring holistic view of the major drivers of employee attrition. Relevant features which include demographic information (age, marital status, gender), performance metrics (overtime,

performance rating), job characteristics (department, job role, job satisfaction) and compensation variables (salary, bonuses) provided in the HR data to be utilized under this study will ensure that we provide a holistic view of the relationship between these features and employee attrition. This will also ensure that we provide a more generalizable and integrated model that incorporates other variables. Looking at the predictive analytics section, we have seen past studies employ various techniques to forecast employee attrition. Adeusi et al. (2024), proposed the use of new models such as Extreme Gradient Boosting in future studies, therefore as an advancement we will incorporate XGBoost among the proposed models under this study. While considerable efforts have also been made by various researchers including George et al. (2022), Poornappriya & Gopinath (2021), Fallucchi et al. (2020), Adeusi et al. (2024), Brockette et al. (2019) and Raza et al. (2022) to determine the most effective models that predict attrition, the models were not utilized to develop practical data solutions that could seamlessly be used by organizations to optimize employee retention measures and proactively address attrition. According to the study conducted by Adeusi et al (2024), future advancement around integration of the best models with HR practices using data tools such as business intelligence tools is paramount for organizations to easily take proactive steps towards retention of their invaluable employee base. Therefore, in this study, we aim to utilize the best model among the proposed to develop an employee attrition and risk scoring system as part of an advancement to the existing studies that have been reviewed. The practical solution under this study is aimed to assist organizations in making data-driven decisions to proactively address employee attrition and optimize employee retention strategies. Looking at the existing systems such as that developed by Brockette et al. (2019), they utilized unsupervised machine learning techniques. We will explore the effectiveness of supervised machine learning techniques and integrate it to the data solution towards employee retention. Notably, previous studies have focused on predicting employee attrition using machine learning models, they have largely overlooked the development of a structured risk scoring mechanism to quantify attrition likelihood. This gap limits the ability of HR professionals to prioritize interventions effectively, highlighting the need for a real-time system that not only predicts attrition but also provides actionable risk scores.

2.7 Conceptual Framework

The conceptual framework for this study is structured around three key components encompassing, the predictor variables, the most effective machine learning model and the response variables, that is employee attrition. The predictor variables also known as independent variables are crucial when it comes to reliably training the machine learning model. This study draws insights from scholarly work such as that of Frye et al. (2018), where statistical approaches were utilized to determine potential predictor variables to be included in the training of the machine learning models. The predictive model is the other crucial component of this study's framework as it performs the predictive task to provide reliable output, that is whether or not an employee is likely to leave. It also serves as a foundation that allows the quantification of the employee attrition likelihood thus generating risk scores for easier interpretability and a more focused intervention strategy. The response variable also the dependent variable, is the output of attrition prediction and risk scoring. Basically, this is the ultimate goal sought after in this study, that is, likelihood attrition and the scores to quantify the output for intervention strategies based on severity.

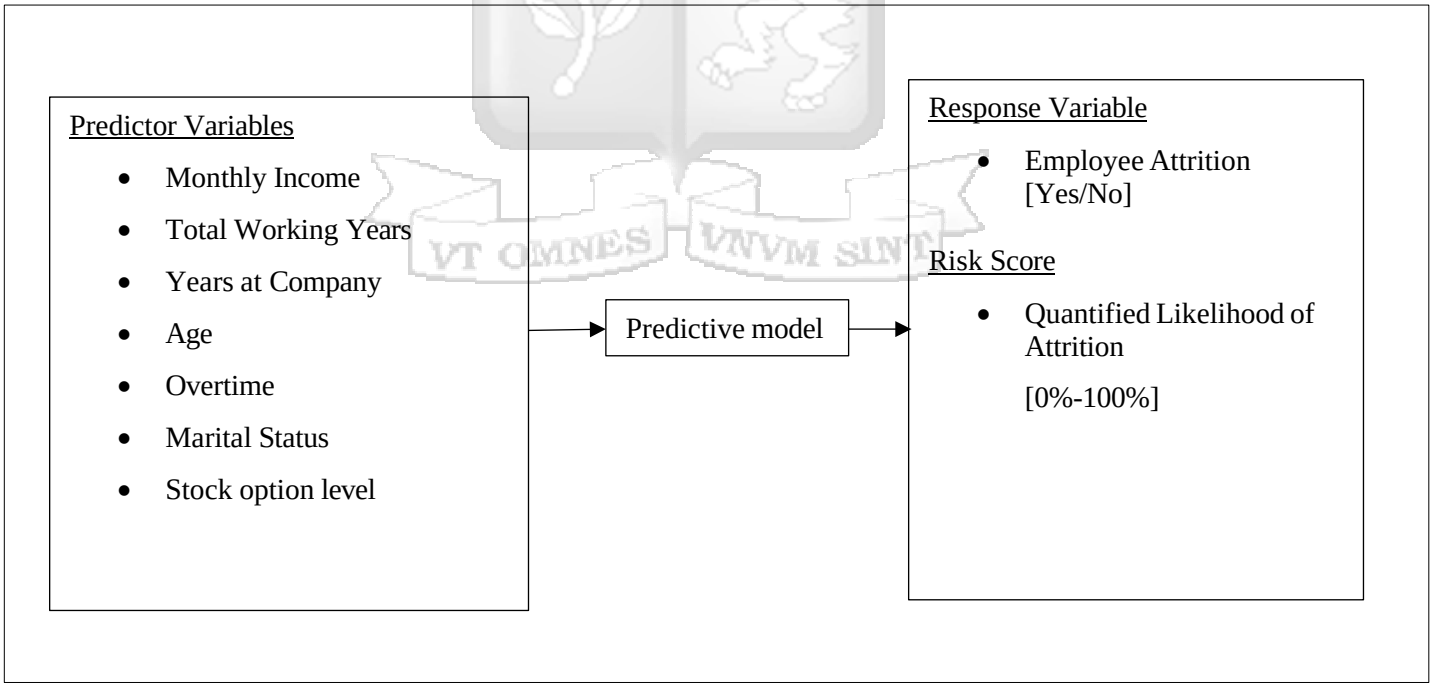


Figure 2.4 Conceptual Framework

Chapter 3: Methodology

3.1 Introduction

In this chapter, we delve into the methodology to be implemented to operationalize the research objectives under this study. The research design, data collection, data preprocessing, exploratory data analysis, feature selection, machine learning algorithms, performance evaluation, overall methodological approach, model deployment as well as the ethical considerations undertaken will be discussed.

3.2 Research Design

The research design that was implemented for this particular study is quantitative research.We relied on quantitative data to operationalize the general objective of this study which is to predict employee attrition and develop a risk scoring data solution that can be used in organizations for targeted talent retention. For the first research objective which is to understand the factors influencing employee's decision to leave an organization, we utilized generalized linear models with interaction terms coefficients to understand the factors that are attributed to an employee's decision to leave an organization. This technique does not only allow one to evaluate the contribution of a feature on their own but also as a set of features, that is, impact of interaction of features on the outcome variable, attrition, as this is what is expected in the real-world scenario that various factors might lead to one quitting their job (Srivastava & Eachempati, 2021). In order to operationalize the second research objective, that is, to evaluate the effectiveness of algorithms in predicting employee attrition, we conducted correlation analysis to avoid features with overlapping information coming into our models. Generally, during the development of a predictive model, it is advisable to train the algorithm on features that are not highly correlated to one another to avoid redundancy (Mehta & Modi, 2021). Therefore, in this research we assessed the correlation of features to ensure that we have no features relaying overlapping information (redundant features) coming into our predictive algorithm. Development of a tangible solution towards employee retention that provides risk scores of employees leaving an organization is the ultimate objective which heavily relied on the findings of the first two research objectives.

3.3 Data Collection

This study utilized secondary data which was made public by IBM data scientists for the purpose of understanding employee attrition through prescriptive, descriptive and predictive analytical approaches. The data set contains 1470 number of records and 35 features (both categorical and numerical) with 'attrition' being the target variable. The wide range of relevant features which include demographic

information (age, marital status, gender), performance metrics (overtime, performance rating), job characteristics (department, job role, job satisfaction) and compensation variables (salary, bonuses) makes this particular dataset reliable in conducting analysis to have a holistic understanding of the major contributors of employee attrition. The features provided in the dataset are also relevant for machine learning task under this study, that is, prediction of employee attrition. The insights from this particular dataset are shared through the tangible solution developed in this particular research.

3.4 Data Preprocessing

Data preprocessing is an essential step towards realizing the most effective model for any kind of task (Mehta & Modi, 2021). When adequately executed, one can be able to improve the performance of the proposed models for a given task and make handling of large data more efficient. Data preparation begins with ensuring that the quality of the data to be used is up to the required standards and that it is accurate and complete (Mehta & Modi, 2021). In this research we checked for missing values within the data to ensure that we have a complete set being utilized for machine learning. The data was complete and as such no imputation or dropping of missing values was done. The next step was feature transformation, this is an important step that ensures that the data can be utilized for statistical and machine learning procedures (Mehta & Modi, 2021). Most machine learning techniques demand that data is converted to numerical input variables as this simplifies statistical computation and helps models learn more efficiently (Mahtre et al., 2020). Ensuring data set constituency by standardizing features to a common scale is also paramount (Mehta & Modi, 2021). In our study we looked at possible features that might need transformation to enhance our statistical analysis as well as integration in machine learning. Categorical features were transformed to numerical leveraging one-hot encoding for the features with low cardinality. Standard scaling was also conducted to ensure that the data points were in the range -1 to 1 which allows effective machine learning.

3.5 Exploratory Data Analysis

Exploratory data analysis allows one to understand the nature of the data and obtain useful insights (Raza et al., 2022). It also helps to uncover trends in data which cannot be directly seen (Mahtre et al., 2020). We conducted univariate, bivariate and multivariate analysis to better understand the contribution of several features to employee attrition. The analysis enhanced the evaluation of features that are important to be included in our algorithms as well to avoid redundancy. According to Mahtre et al. (2020), a correlation matrix comes in handy in exploring the impact of different attributes on each other thus picking the most important features for the algorithms.

3.6 Feature Selection

Feature selection is key for handling the curse of dimensionality thereby enhancing the effectiveness of models (Raza et al., 2022). For this study, as the first step to selecting key indicators for our algorithms we excluded redundant features from multivariate analysis. Thereafter, we also employed recursive feature elimination to retain the most significant indicators to be utilized to the machine learning algorithms. Ultimately, by selecting relevant features we enhanced the performance of our models.

3.7 Machine Learning Techniques

This study explored various machine learning techniques to predict the likelihood of an employee leaving an organization. We provided a key focus on XGBoost, Random Forest (ensemble techniques) and Support Vector Machine, K- Nearest Neighbors as well as Logistic Regression machine learning models. These set of models are distinct and provide balanced interpretability, computational efficiency, and predictive power. While Logistic Regression serve as baseline for evaluation, Support Vector Machine and K-Nearest Neighbors capture complex decision boundaries. The ensemble techniques like Random Forest and XGBoost enhance robustness and generalizability, and are therefore suitable for prediction of the likelihood of employees leaving an organization.

The XGBoost which stands for Extreme Gradient Boosting is an advanced Gradient Boosting technique that integrates a regularized framework that curbs overfitting in its classification tasks (Mahtre et al., 2020). This machine learning algorithm builds on the concept of gradient boosting, where a sequence of weak learners is trained several times and converted to stronger learners through gradual improvement of one tree to the next (Mahtre et al., 2020). With each iteration weights are increased and assigned to data points that were initially misclassified. The final prediction is obtained by aggregating the predictions from all iterations through a weighted majority sum (Mhatre et al., 2020). For risk scoring, the probability output can be utilized as the risk scores and assigned to employees such that low and high-risk individuals are identified.

The mathematical intuition of the XGBoost model can be represented as follows;

Objective: minimizing the loss function L

$$L(t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(\text{ft})$$

Where:

y_i represents the true label for the i^{th} data point.

$\hat{y}_i^{(t-1)}$ denotes the predicted value from previous iteration

$f_t(x_i)$ denotes the function of tree learned in the current iteration

l represents the loss function for classification (e.g., mean squared error for regression or log-loss for classification).

$\Omega(f_t)$ denotes the regularization term that penalizes model complexity

The final prediction output which is an aggregation of the previous predictions is obtain as follows;

$$\sum_{t=1}^{T} f_t(x_i)$$

Random Forest is an ensemble technique that uses bagging to combine several decision trees to enhance predictive accuracy for classification tasks (Mehta & Modi, 2021). By aggregatingthe predictions of individual trees, this technique minimizes overfitting and can efficiently be utilized for large datasets (Mehta & Modi, 2021). For risk scoring, this machine learning technique returns probability estimates which can be assigned to each employee as their risk score. Random Forest model can be mathematically represented as follows;

$\hat{y}$= majority vote $\{h_1(x), h_2(x), \ldots, h_n(x)\}$

Where:

$\hat{y}$ denotes the predicted class

$h_i(x)$ is the prediction made by the i^{th} decision tree for input x

N is the total number of decision trees in the forest

For a classification task, the final prediction output is the class that receives majority votes from all the individual trees (Mehta & Modi, 2021).

Support Vector Machine (SVM) is another useful algorithm with the ability to tackle both linear and non-linear binary classification problems (Raza et al., 2022). The underlying concept behind this algorithm involves creating a hyperplane in higher dimensional space to achieve separation between two classes (Raza et al., 2022). The hyperplane is positioned in such a way

that it maximizes the geometric margin between itself and the closest data points thus the name support vector. This concept makes this particular algorithm a maximum margin classifier (Raza et al., 2022). Although obtaining probabilities might not be direct for SVM, the use of confidence scores for probability score estimation using platt scaling can be liable to enhance scoring the employees who are at risk of departure. The mathematical intuition of Support Vector Machine can be represented as follows;

Objective: Penalizing the instances that are misclassified and those within the margin;

Minimize $\frac{1}{2} \| \boldsymbol{w}^2 \|$

Constraints:

$$y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 \; for \; all \; i$$

Where;

$\boldsymbol{w}$ represents the weight matrix

b represents the bias value

$\boldsymbol{x}_i$ represents feature vector for the i^{th} training

This formula is designed to identify the hyperplane that maximizes the margin.

K- Nearest Neighbors is also one of the most commonly used algorithms in classification problems. According to Mahtre et al. (2020), KNN model is underpinned in the concept of identifying the K nearest data points in the training set to a new instance such that the instanceis classified based on the majority votes of these K neighbors. The distance metric is a crucial element in the model and is meant for determining data point (neighbors) proximity. Techniques such as Manhattan, Minkowski and Euclidean distances can be utilized to measure the distance from the data points (Mahtre et al., 2020). For risk scoring, KNN returns predicted probability of attrition which can be interpreted as risk scores.

The K- Nearest Neighbors' Euclidean distance metric can be mathematically represented as follows;

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2}$$

Where;

x represents the point where the distance from a different point is being evaluated.

x_i is the second point such that $d\ (x, x_i)$ represents the coordinates from first point to the secondpoint being determined.

$x_{i,j}$ represents the value for the j^{th} feature for the second point.

Once an optimal distance is obtained, K neighbors are selected and new instances are classified accordingly (Mahtre et al. 2020).

Logistic Regression is a supervised machine learning model that is designed for binary classification problems (Raza et al., 2022). The model uses a sigmoid function that provides probability values between zero and one which is suitable for classifying instances into two groups. The S-shaped function is therefore deemed fit for prediction (Raza et al.,2022). It maps predicted values to probabilities between zero and one. For risk scoring, logistic regression assigns a probability score, usually between zero and one, to every employee which can be interpreted directly as their risk score for leaving an organization.

The S-Shaped logistic function can be mathematically represented as follows;

$$y = \frac{e^{b_0+b_1x}}{1 + e^{b_0+b_1x}}$$

Where;

y represents the predicted class output

b_0 represents the bias term

b_1 represents the coefficient for input feature x

3.8 Performance Evaluation

This study evaluated the effectiveness of the models in predicting employee attrition using the following proposed metrices; Accuracy, Precision, Recall and F1 Score.

Accuracy is a model evaluation metric that measures the overall predictive power of our models giving us how often they make correct classifications (Raza et al., 2022). In this study we sought forthe model that gives us the highest accuracy which was Random Forest with 80% Accuracy. The formula for obtaining accuracy is provided below;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP denotes the True Positives, that is, the number positive instances that are correctly predicted

TN denotes the True Negatives, that is, the number of negative instances that are correctly predicted.

FP denotes the False Positives, that is, the number of positive instances that are incorrectly predicted.

FN denotes the False Negatives, that is, the number of negative instances that are incorrectly predicted.

Precision on the other hand, is a model evaluation metric that measures the proportion of instances correctly identified as true positives out of all the positive instances (Raza et al., 2022). We sought for the model with the highest precision rate as this would imply that it can effectively identify employees with highest risk of leaving an organization. This was the Random Forest with a Precision score of 92%. High precision also indicates that the false positives are minimal and this gives the company insights to ensure they incorporate pre-emptive measures for their human capital retention without implementing unnecessary intervention (Brockett et al., 2019). The formula for obtaining precision isprovided below;

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Recall is a model evaluation metric that measures the proportion of correctly identified true positive instances (Raza et al., 2022). For this particular study, we chose the model that recorded a high recall (Random Forest with 75% Recall) as it had a better ability to correctly identify actual cases of attrition. This is deemed important for organizations as they will take up pre-emptive measures to ensure they address employee attrition. The formula for obtaining recall is provided below;

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

F1- Score is a model evaluation metric that summarizes the performance of an algorithm by combining the precision and recall (Raza et al., 2022). We sought for the model that recorded a high F1 score as this implies that it can effectively predict employee attrition while ensuring that organizations provide optimal interventions for the employees who are likely to leave withoutoverwhelming them with false alerts. This was Random Forest with F1 Score of 83%. The formula for obtaining F1-Score is provided below;

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.9 Overall Methodological Approach

This study took a step-by-step approach in the development of a model that effectively predicts employee attrition. As stated in the previous sections we began with a complete dataset which was then preprocessed for machine learning. The training and testing of the machine learning models discussed in section 3.7 was done to uncover the best model for the tangible solution under this study. Figure 3.1 illustrates the methodological approach taken towards an effective model for the solution under this study.

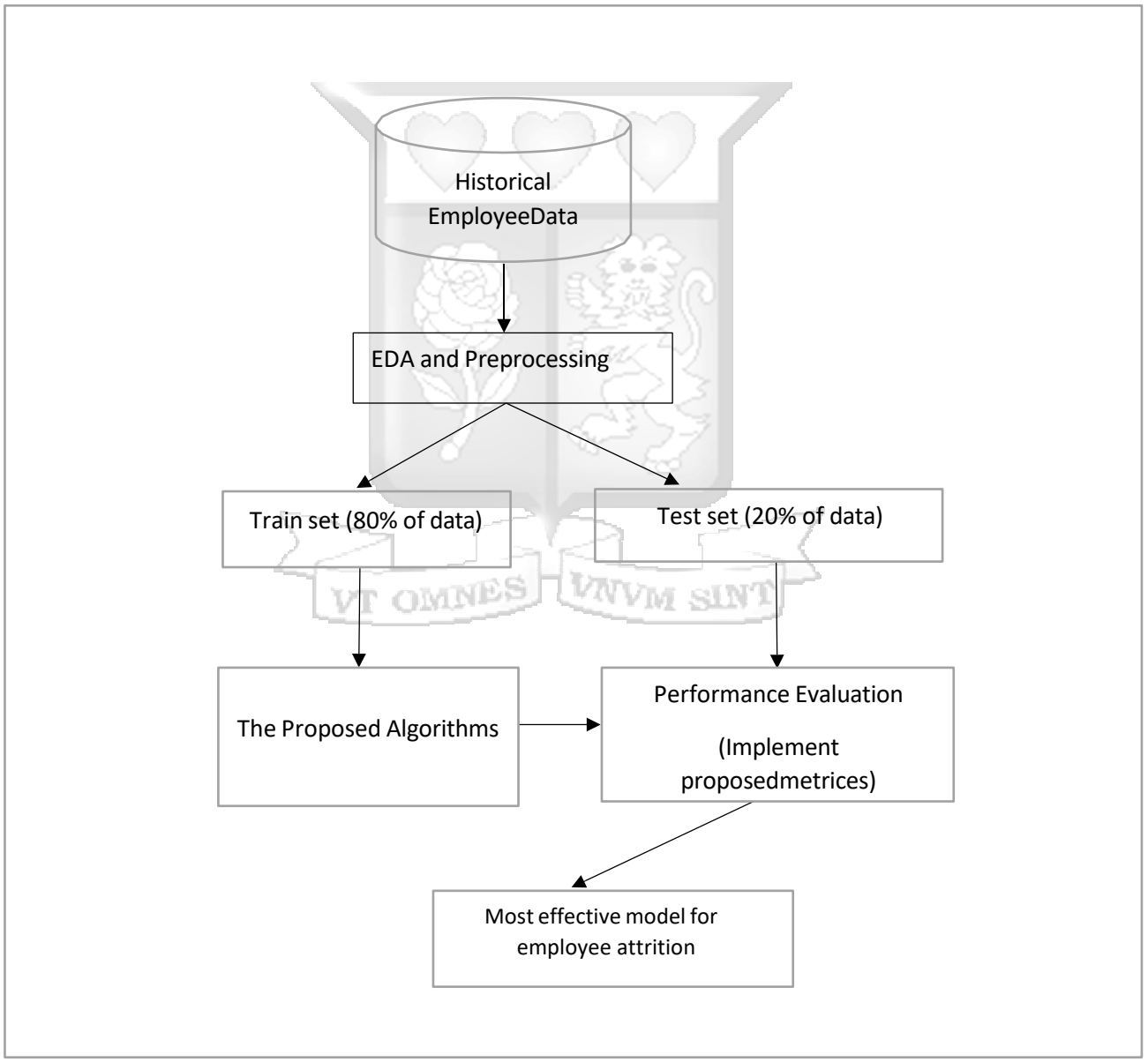


Figure 3.1: Overall methodological approach.

3.10 Model Deployment

Once an effective model for predicting employee attrition is developed, it was deployed to production environment that allows real-time risk scoring for employees who are most likely to leave an organization. We utilized the R-Shiny interface where the model took up data and made real-time prediction returning the risk scores indicating the likelihood of an employee quitting their job. The back-end, that is, the server function is responsible for the logic behind the tangible solution under this study. The front-end, that is, the user interface isresponsible for the outcome, risk scores displayed for the employees who are likely to leave an organization. This ultimately ensures that organizations have a clear, data-driven view of employee attrition risks, enabling proactive decision-making and targeted interventions to retain valuable talent.

3.11 Ethical Considerations

Ethical considerations are essential to enhance the validity and reliability of this study. A key focus was placed on the acquisition of the secondary data to utilized under this research. We utilized a publicly available dataset for research that is provided by Institute of Electrical and Electronics Engineers. Employee data is highly sensitive and therefore having the data anonymized was paramount. In order to ensure overall integrity and this study, the proposal was submitted to the Strathmore University Institutional Scientific Ethics Review Committee (SU-ISERC) for review (Strathmore University Institutional Scientific Ethics Review Committee, n.d.).

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter provides the system design and architecture for the data solution developed under this study. A clear discussion on the system requirements and system design will be provided under this chapter.

4.2 System Requirements

A set of requirements are needed to ensure that the system under this study achieves its intended objective on attrition prediction, risk scoring and reporting. In this section we provide both the functional and non-functional requirements for the operationalization of the data solution under this study.

4.2.1 Functional Requirements

Aligning with this study's objectives, several functional requirements are necessary for the operationalization of the system. First, the system should have real-time data ingestion capabilities, ensuring that users can load relevant data for attrition prediction results. Second, the system should support attrition prediction by utilizing the random forest algorithm to predict the likelihood of an employee leaving an organization based on the loaded data. The system should also support risk scoring based on the prediction outputs. Ultimately, the system should support reporting through the implementation of a user-friendly dashboard that displays risk scores for the employees to allow easier decision making and targeted retention strategies.

4.2.2 Non-Functional Requirements

The system under this study has several non-functional requirements to be met that ensure it upholds its effectiveness. First, reliability is key, the system should capture data and provide timely insights on employee attrition. Second, usability is an essential requirement that allows seamless system integration and application by users. The system's accessibility is also key given that users require internet connection to be able to access the web-based solution. Ultimately, security is also key, the system should comply with data protection guidelines to ensure that employee data is safeguarded.

4.3 System Components

The key components of the real-time attrition prediction and risk scoring system which will be implemented on R-Shiny include, data, server-side and user-interface. The data component simply refers to the relevant employee records loaded up in the system. The server-side component, is basically where functions are designed to allow attrition prediction based on the loaded data and risk scoring based on the prediction outputs. Ultimately, the user-interface component is also a key component for the system under this study. The user-interface allows real-time reporting of prediction results and risk

scores for decision making and actionable insights. These components ensure that the system meets the ultimate research objective under this study and remains effective and efficient while at it.

4.3.1 Interaction between System Components

The components of the employee attrition prediction and risk scoring system interact to ensure smooth functionality of solution. The data is loaded and manipulated into the server-side where customed functions perform predictive analytics based on it. The prediction outputs are then utilized in the generation of risk scores which are ultimately displayed on the R-Shiny user interface. The figure 4.1 below is a representation of the interaction that takes place between the primary architectural components which facilitate functionality of the system under this study.

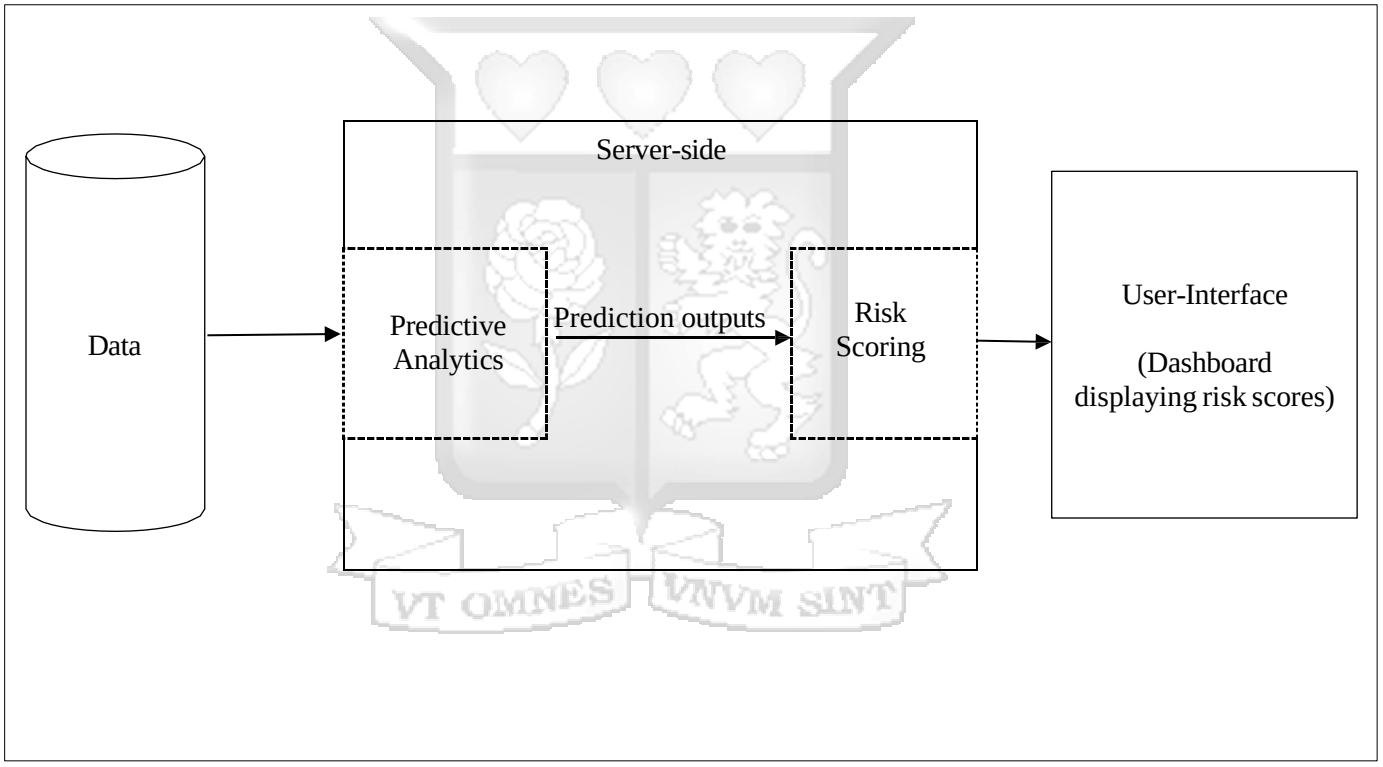


Figure 4.1: System Architecture

4.3.2 Data Flow within the system

The flow of data within the system ensures seamless predictive modeling and reporting of insights to the end-users. Employee data flows to the server-side where it is manipulated and utilized in predictive analytics and risk scoring. The insights are thereafter delivered to users' dashboard enabling decision making and formulation of data-driven retention strategies.

The flow of data within the system is illustrated in the figure below.

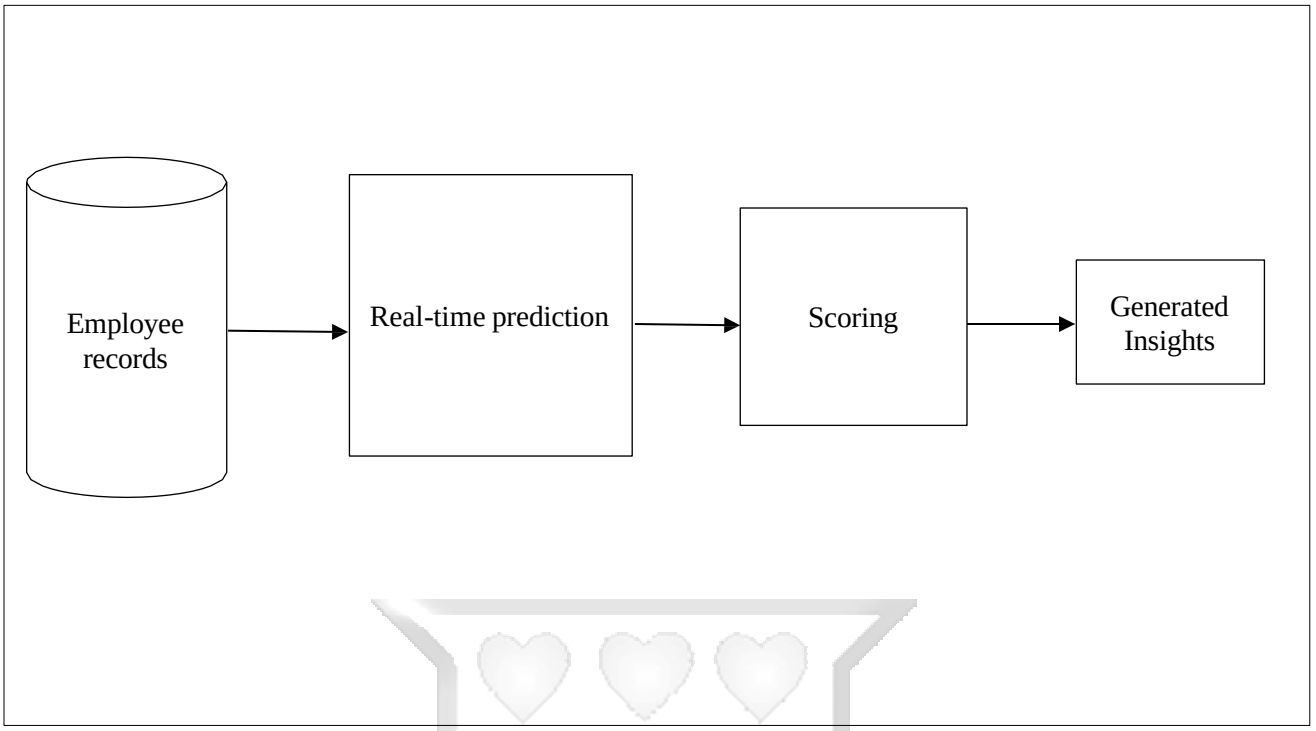


Figure 4.2: Data flow within the system.

4.4 System Design

This section focuses on the design of the real-time attrition prediction and risk scoring system developed under this study. This section discusses the database design, user-interface design and a use case discussion describing how the solution under this study can be seamlessly used.

4. 4.1 Database Design

The real-time employee attrition prediction and risk scoring system developed under this study leverages the relational database MySQL for storing employee data, real-time prediction results, and feature importance scores. MySQL is a key component that allows efficient data management of structured tables thus enabling fast data retrieval and storage. The system integrates with R Shiny through API-based interactions, where employee details entered via the employee input panel are sent to the server-side model, which processes the data and returns **a** prediction with a risk score. The results are then stored in MySQL and dynamically retrieved for visualization in the UI. The seamless API server interactions allows HR professionals to interact with real-time attrition insights, improving decision-making and talent retention/intervention strategies. Figure 4.3 illustrates the API calls encompassed between the key components reliable for the development a flexible and scalable solution.

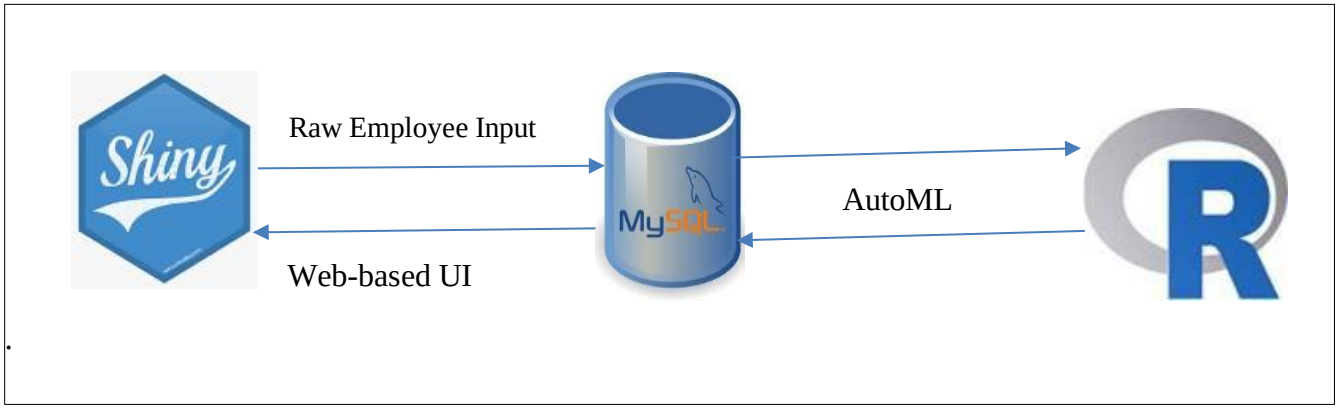


Figure 4.3: API calls between Shiny and MySQL

4.4.2 User-Interface Design

The User Interface (UI) of the Real-Time Employee Attrition Prediction and Risk Scoring System**,** built in R Shiny**,** consists of three key panels**:** Employee Input Panel, Prediction and Risk Scoring Panel, and Feature Importance Panel**.** The selected employee features are manually input and the prediction button is clicked thus triggering server-side action of utilizing the random forest machine learning algorithm to predict employee attrition (Yes/No) and generate risk scorse, ideally between 0% and 100%. The system dynamically visualizes key contributing factors using feature importance bar plots, helping HR understand contributors of attrition such as monthly income and overtime. The responsive UI seamlessly integrates with MySQL, ensuring real-time data processing and decision-making for effective employee retention strategies. The figure below, figure 4.4 provides a use-case diagram showcasing the real-time prediction workflow and the definition of how HR professionals interact with the system under this study.

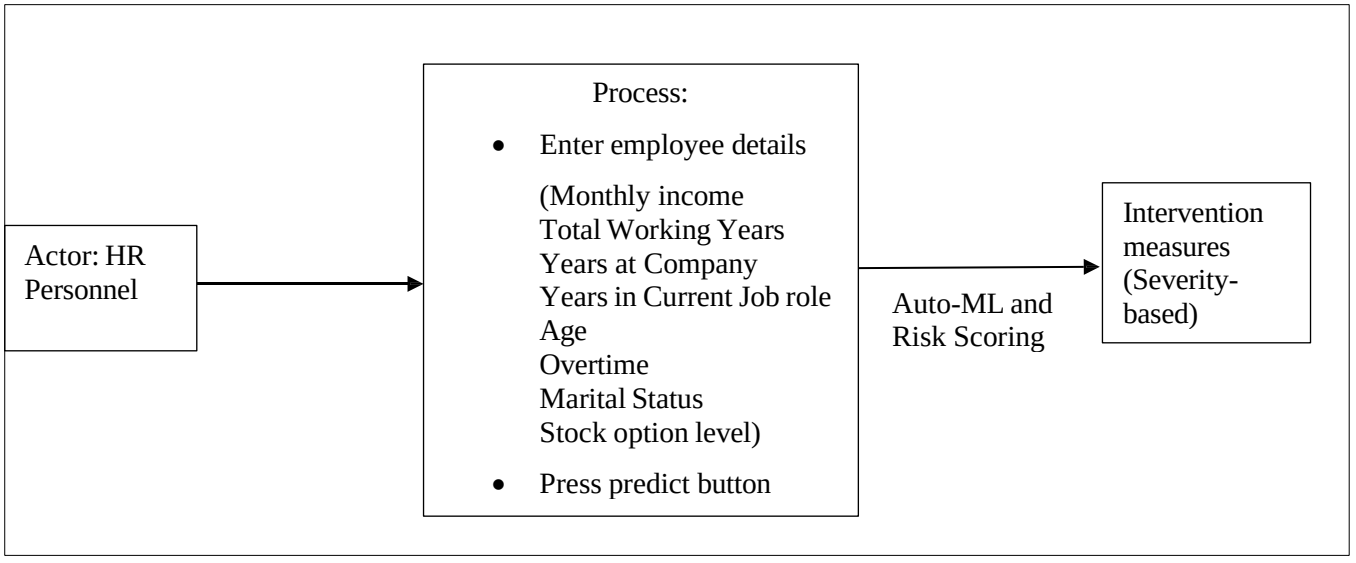


Figure 4.4: System's Use-Case

Chapter 5: Results

5.1 Introduction

This chapter outlines the results derived from the analysis of employee records utilized under this study. The focus is placed on the outcomes for the first two research objectives, that is, to understand the factors that influence an employee's decision to leave an organization and to evaluate the effectiveness of algorithms in predicting employee attrition. Exploratory data analysis, feature importance and model performance are discussed.

5.2 Exploratory Data Analysis

This section focuses on presenting the findings derived in an effort to understand and familiarize with the employee records utilized under this study. Univariate, bivariate and multivariate analysis were conducted to uncover the distribution, interactions and relationships within the dataset.

5.2.1 Univariate Analysis

There are 1470 rows within the dataset utilized under this study. The target variable, that is, 'Attrition' is distributed with 1233 data points labeled as "No" and 237 data points labeled as "Yes". This indicates an imbalanced data. The study utilized ROSE (Random Over-Sampling Examples) technique to address this class imbalance by generating new synthetic samples in the under-represented class. The figure below is a graphical representation of the distribution of the dataset utilized in this study.

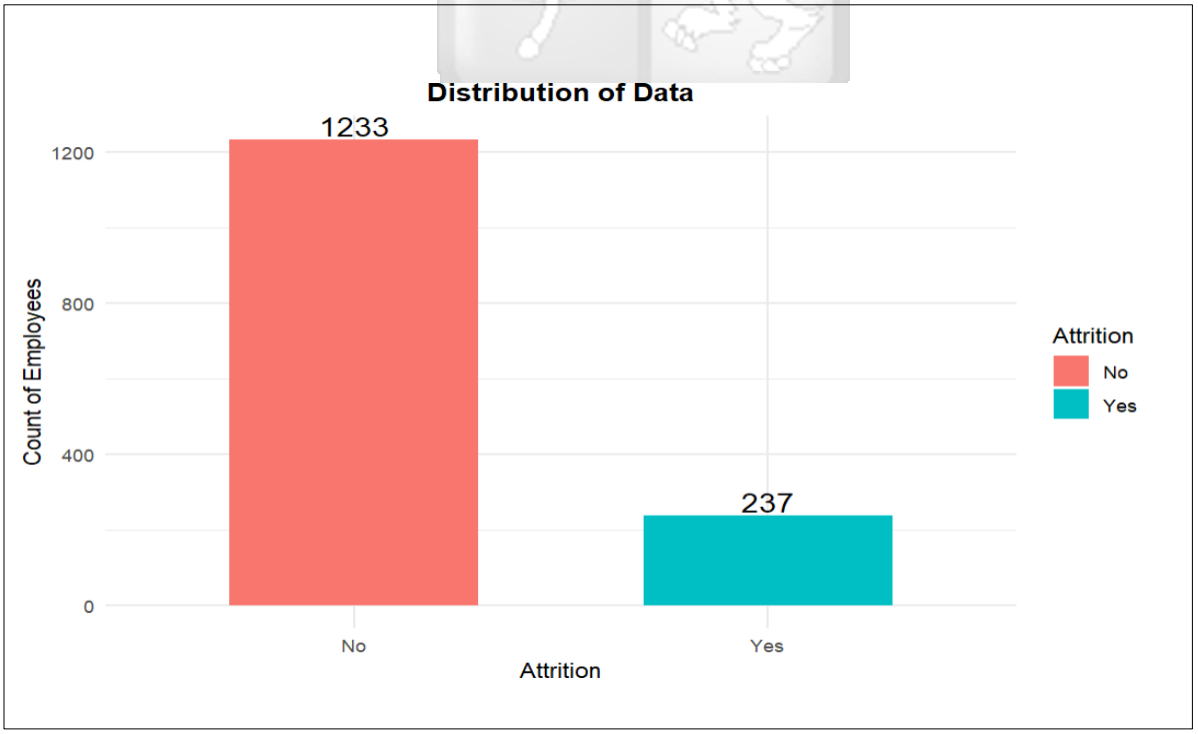


Figure 5.1: Distribution of Data

The data was balanced using ROSE oversampling technique and the figure 5.2 below is an illustration of the balanced data used in machine learning.

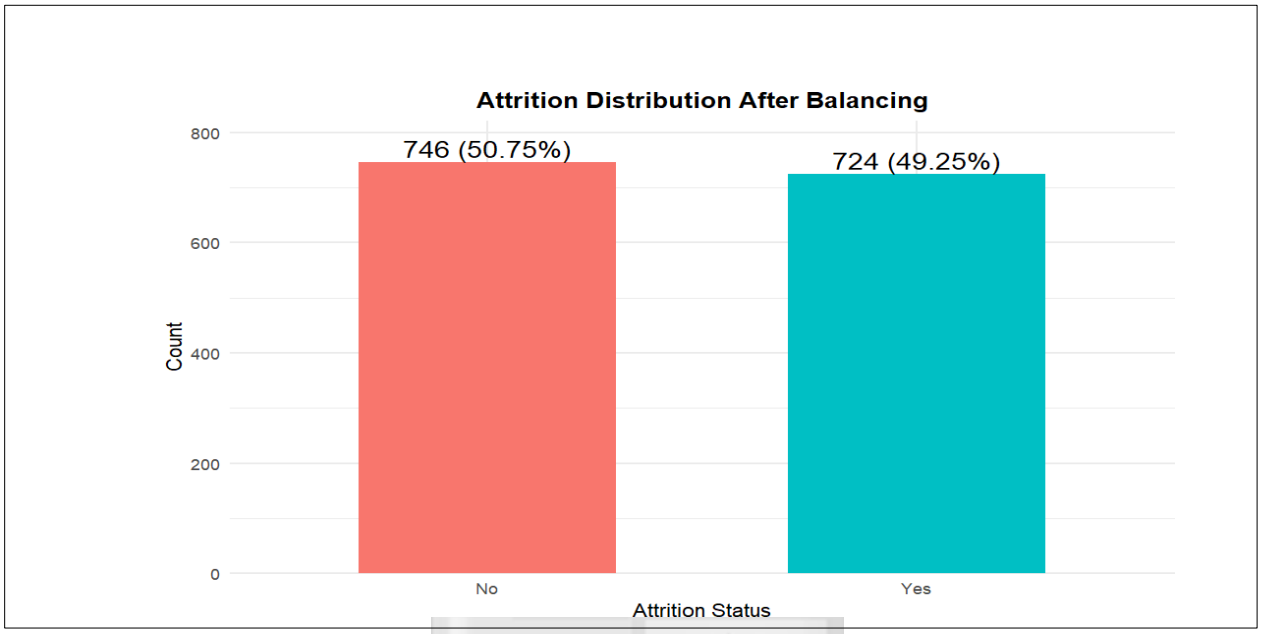


Figure 5.2: Attrition distribution after ROSE oversampling

5.2.2 Bivariate Analysis

Bivariate analysis was conducted to understand the relationship between two variables. The exploration focused on the commonly known factors of attrition including income, job satisfaction, tenure at company versus attrition. From the density plot of the monthly income and attrition, the peak of the "Attrition=Yes" curve is at lower income levels, that is between $1009-5000. This clearly suggests that employees with lower income tend to leave. From the violin plot of tenure (Years at Company) and attrition, the majority of employees who leave have very few years at the company. Looking at the bar plots depicting the relationship between the job satisfaction and attrition, employees under category 1, that is, with lower job satisfaction tend to leave the company. Employees with high satisfaction also tend to leave which could attributed to other factors such as career advancement opportunities.

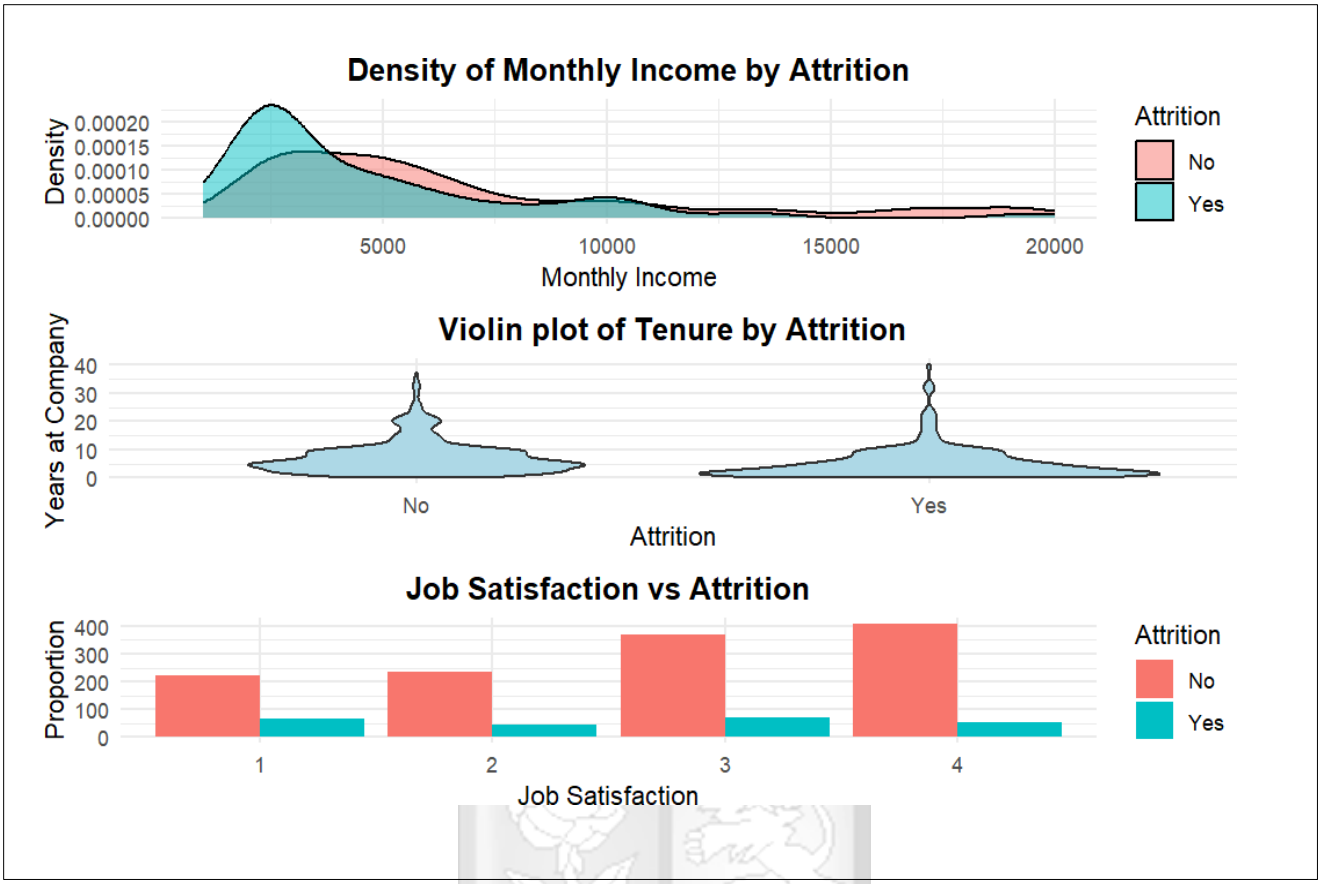


Figure 5.3: Bivariate Analysis

5.2.3 Multivariate Analysis

A correlational analysis was conducted to uncover the relationship between the variables within the dataset. It also served as an initial screening for potential candidate variables to be included in the predictive algorithms suggested under section 3.7. The figure below is a representation of a correlation heatmap depicting the relationship between the variables within the dataset.

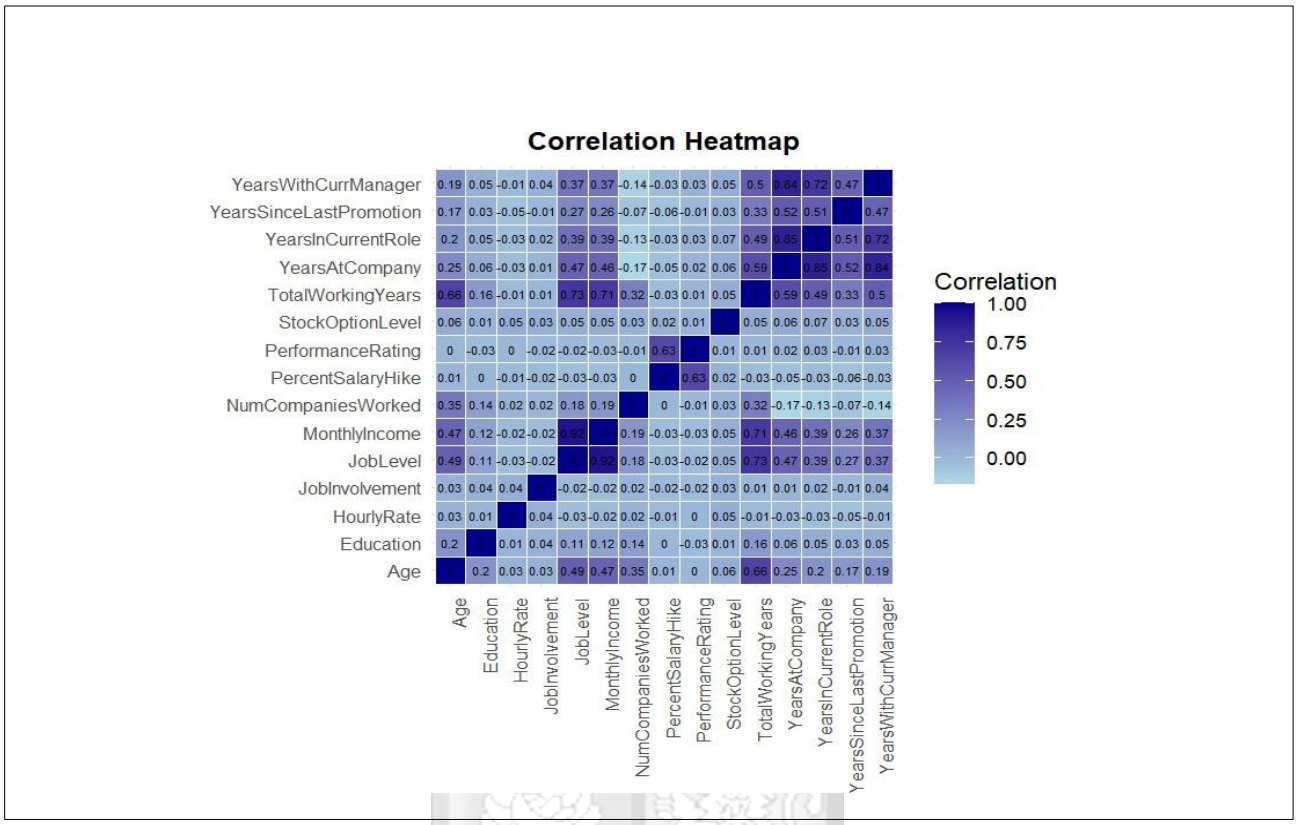


Figure 5.4: Correlation Heatmap

A strong positive correlation exists between the Years at company and Years with current manager (0.84), indicating that employees with longer tenure tend to have worked with their current managers. Similarly, a strong positive correlation of 0.85 is recorded between the Years at company and Years at current role suggesting that employees with longer tenure at the company do not frequently change their roles. Another noteworthy correlation is the one between Total working years and Job level which is at 0.73, clearly indicating that employees with more professional experience hold higher job levels within an organization. The correlation between the variables presented in figure 5.4 are beneficial for initial screening of potential variables for inclusion in modeling. High correlation between variables indicate that they provide similar information if both are introduced in the model as a result of multicollinearity. To ensure the stability and interpretability of the model multicollinearity is handled.

5.3 Factors Influencing Employee Attrition

To operationalize the first research objective, this study relied on generalized linear model coefficients as a statistical approach of understanding the factors that influence an employee's decision to leave an organization. The choice of the statistical technique was primarily informed by the nature of the response variable (binary outcome) in the dataset utilized in this study. Figure 5.5 displays the summary statistics.

Term	Estimate	StdError	zValue	Pr
(Intercept)	-2.100	0.104	-20.145	< 2.22e-16
Age	-0.147	0.098	-1.499	0.13380391
JobSatisfaction	-0.285	0.079	-3.591	0.00032962
MonthlyIncome	-0.509	0.144	-3.533	0.00041070
YearsAtCompany	-0.405	0.123	-3.279	0.00104052
JobInvolvement	-0.338	0.076	-4.426	9.6015e-06
Age_JobSatisfaction_Interaction	0.056	0.088	0.636	0.52483747
MonthlyIncome_YearsAtCompany_Interaction	0.283	0.072	3.920	8.8459e-05
MonthlyIncome_JobSatisfaction_Interaction	0.008	0.105	0.078	0.93785911
Age_MonthlyIncome_Interaction	0.171	0.103	1.667	0.09542137
JobInvolvement_MonthlyIncome_Interaction	0.039	0.088	0.436	0.66248508

Figure 5.5: GLM Summary Statistics

Based on the above generalized linear model output, the P-Values of job satisfaction, monthly income, years at company and job involvement are less than 0.05 indicating that they are statistically significant. These features have an impact on attrition. Intuitively, a higher job satisfaction, income and longer tenure reduces the rate of attrition. This study not only assessed the impact of each stand-alone factor, but also the contribution of more than one factor towards attrition as this scenario is most likely expected in real-life. From this assessment, the interaction between monthly income and years at company was statistically significant indicating that longer serving employees with low income are more likely to leave an organization as they may feel undervalued.

5.4 Machine Learning and Model Performance

To operationalize the second research objective, this study utilized a set of identified features on five machine learning models, that is, XGBoost, Random Forest(ensemble techniques) and Support Vector Machine, K- Nearest Neighbors as well as Logistic Regression model. Recursive Feature Elimination (RFE) was used to determine the most significant employee features for predicting attrition by iteratively removing less important variables. The key features including, monthly income, total working years and overtime were identified and integrated into the machine learning models. The identified features are as shown in figure 5.5

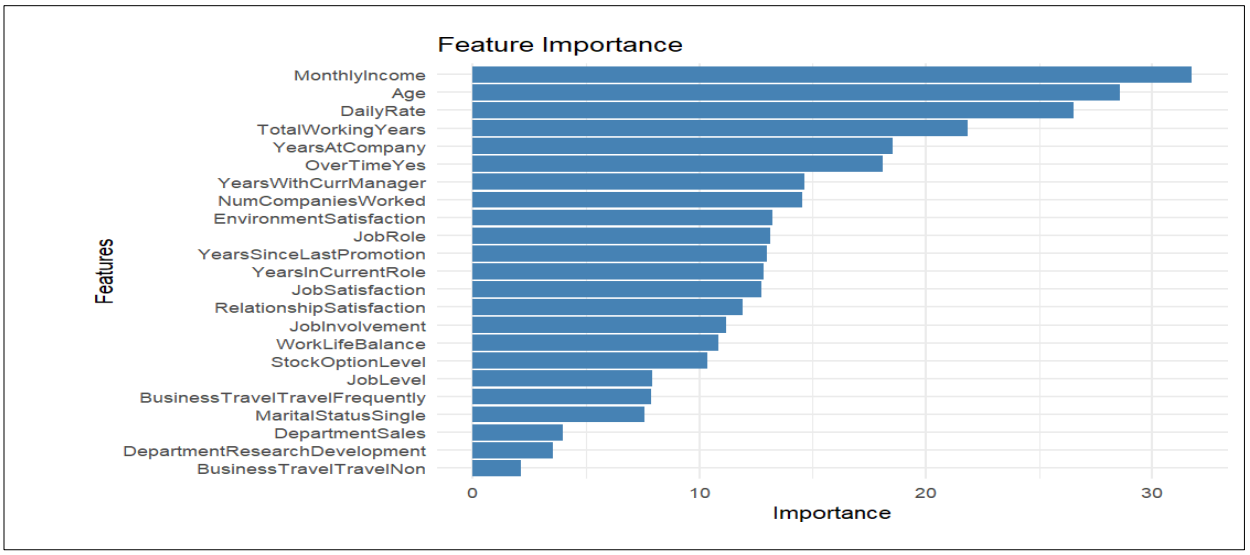


Figure 5.6: Representation of Feature Importance

The important features from the representation above and intuition from multivariate analysis conducted under subsection 5.2.3 were utilized in modeling. To determine the best model the performance of the models was evaluated using Accuracy, Precision, Recall and F1 Score metrices. R programming language was used to train the machine learning models and further evaluating their performance. Key libraries for implementing the machine learning models including xgboost, randomForest, class and e1071 were employed. Figure 5.6 provides the performance of each of the models utilized in this study.

Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1_Score
XGBoost	0.7337884	0.8925234	0.7764228	0.8304348
Random Forest	0.8088737	0.9130435	0.8536585	0.8823529
KNN	0.7098976	0.8926829	0.7439024	0.8115299
SVM	0.6996587	0.9247312	0.6991870	0.7962963
Logistic Regression	0.7167235	0.9267016	0.7195122	0.8100686

Figure 5.7: Performance Comparison of ML Models

Based on the output on performance of each model, Random Forest outperformed all the other models in predicting attrition. It has the highest accuracy at 80%, indicating that the model correctly classified 73% of employees as either leaving or staying. It also strikes a good balance in terms of precision and recall as clearly indicated by the F1 metric (83%) indicating that the model maintains the balance in providing minimal false positives (high precision) and the ability to correctly identify actual cases of

attrition (high recall). Logistic regression follows with an accuracy of 72% and F1 Score of 92%. The Random Forest model demonstrates the best overall performance in this study. The Logistic Regression and XGBoost also provide strong performance output making them viable alternatives.

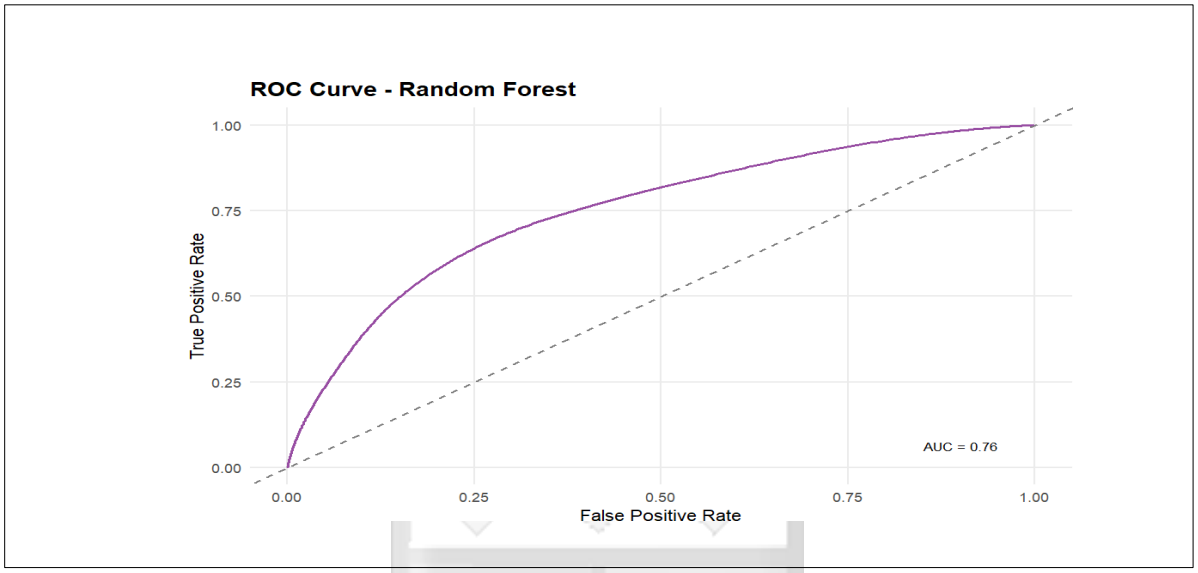


Figure 5.8: Area Under ROC Curve for Random Forest Machine Learning Model

The Random Forest model has AUC of 0.76, the score suggests the model is well-suited in predicting employee attrition. It reliably flags at-risk employees, thus ensuring targeted retention strategies within the organizations. The imperfect discrimination of this model underpins the importance of evaluating the output of this model before decision -making and formulation of intervention strategies.

Chapter 6: System Implementation

6.1 Introduction

This chapter focuses on the implementation of the tangible solution developed under this study. The system provided utilized the insights obtained from the findings of the first two research objectives discussed in the previous chapter. Essentially, this chapter focuses on the third research objective under this study, that is, to use R-Shiny to design and implement tangible solution purposely to predict attrition and provide risk scores. User-Interface, server and system functionality testing will be discussed.

6.2 Server Implementation

The server-side serves as the most critical part for the realization of the solution under this study. It contains functions behind the logic of the system that is, the ability to predict employee attrition based on data input provided in the UI and the pretrained Random Forest model. Furthermore, it evaluates the risk score of individual employees utilizing the probabilities generated by the model. The implementation of the server-side allows real-time prediction and risk scoring thus facilitating easier and faster data-driven actions in an organization.

6.3 User-Interface

As indicated in sub section 4.3.1, the User-Interface (UI), is one of the primary components of the system developed under this study. Shiny and shinydashboard are the libraries utilized in crafting the layout of the UI ensuring that it allows seamless usage. The UI displays three panels, one where employee's input, the prediction and risk scores results and feature importance panel which enables drawing of insights of the root cause of either leaving or staying in an organization. With all these features, the UI provides easy interpretability and decision making on retention strategies. Figure 6.1 below displays the systems front-end.

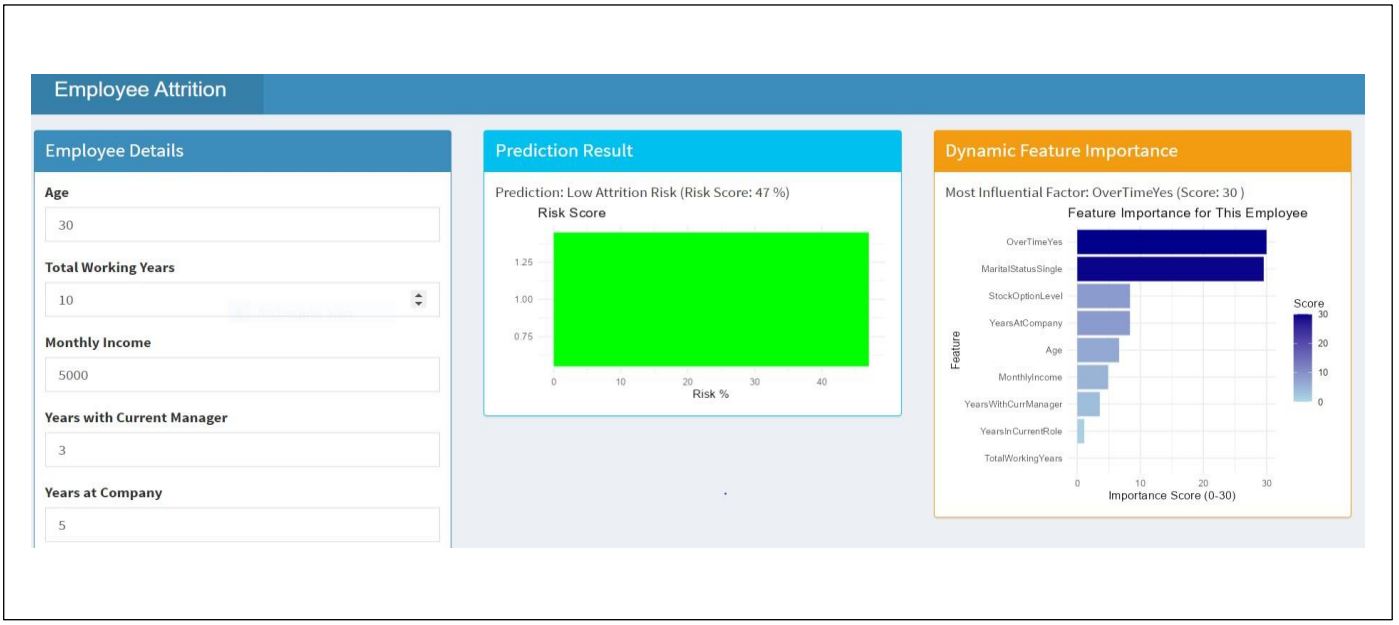


Figure 6.1: User-Interface

6.4 System Functionality Test

The system developed under this study is meant to provide real-time predictions and risk score for new employee input. The functional requirements outlined in section 4.2.1 were assessed. The system has the ability to process new employee input provided and predicting whether one will leave an organization in real-time. It also provides risk scores based on the prediction probabilities with no computational delays. The average server execution time is less than one second and figure C.1 provides a code snippet showcasing the systems responsiveness, providing real-time insights on the front-end. Meeting these functional requirements, the system developed under this study provides instant data driven insights that is resourceful to an organization in terms of decision making and formulating strategies towards retention of its employee base.

Chapter 7: Conclusion, Recommendation and Future Works

7.1 Conclusion

The primary objective of this study was to develop a real-time employee attrition prediction and risk scoring tool. The realization of the tangible solution under this study was guided by three specific objectives outlined in section 1.3.2. The IBM dataset was used in the operationalization of the objectives under this study. This section provides the insights drawn from the exploration of the data, predictive analytics conducted and the development of the tangible solution under this study.

7.1.1 Factors Influencing Employee Attrition

This study's first objective was to use statistical approaches to analyze and identify the factors influencing employee attrition. Previous studies including that of Yahia et al. (2021), Frye et al. (2018) and Raza et al. (2022), provided their insights on factors influencing employee attrition based on statistical approaches. In this study, given the nature of the response variable, that is, binary outcome, generalized linear model with link "logit" was utilized to assess the factors attributing to an employee's decision to leave or stay. The model utilized also incorporated interaction terms to provide an understanding of the contribution of more than one predictor variable to the response variable (attrition). From the summary statistics of the GLM with interaction terms model, job satisfaction, monthly income, years at company, job involvement and the interaction between monthly income and years at company were statistically significant indicating that they contribute to an employee's decision to either leave or stay in a company. This aligns with the findings from the logistic regression summary statistics obtained by Yahia et al. (2021), which indicated that tenure, job satisfaction, income and job involvement contributed to employee attrition.

7.1.2 Model Selection

This study's second objective was to evaluate the effectiveness of machine learning algorithms in predicting employee attrition. The performance of five machine learning models including XGBoost, Random Forest (ensemble techniques) and Support Vector Machine, K- Nearest Neighbors as well as Logistic Regression model was assessed using Accuracy, Precision, Recall and F1 Score performance metrices. The features incorporated in the model were selected using Recursive Feature Elimination as well as through multicollinearity analysis to avoid overfitting of the mentioned machine learning models. Random forest outperformed all the other models with an Accuracy of 73%, Precision of 92%, Recall of 75% and F1 Score at 83%. This showcased the power of Random Forest as an ensemble technique to effectively predict employee attrition. This is in contrast to what researchers such as Falluchi et al. (2020), Brockett at al. (2019) and George et al. (2022), who assessed Random Forest and

other machine learning techniques and concluded that Random Forest did not outperform the others.

7.1.3 Prediction and Risk Scoring System

This study's third and ultimate objective was to develop a system that predicts employee attrition and generates risk scores in real-time using relevant HR data. The solution under this study was developed using R- Shiny with its core aspects embedded on the server function. The server function utilized the findings of the first two research objectives under this study, that is, employee input of relevant factors uncovered and Random Forest for real-time prediction and risk scoring. This web-based solution provides insights on likelihood of an employee leaving and further quantifies this likelihood thus ensuring proactive interventions and retention strategies based on the severity of the output. The key contributing factors are also displayed on the user-interface which ensures tailored interventions are provided.

7.2 Implications of Findings

This section provides the implications of the key findings drawn from the practical exploration of this study towards understanding the concept of employee attrition. The focus will be on the following key stakeholders, human resource teams, data solution providers and the government.

7.2.1 Human Resource

Organizations may suffer financial loses and incur more cost in recruiting, training and onboarding of new talent (Nuel et al., 2022). The Human Resource (HR) professionals can utilize this study's valuable insights to make data driven decisions to reduce employee attrition and costs associated with employee attrition. By understanding the major contributors of employee attrition, HR teams can formulate retention strategies. Managing employees job involvement, ensuring job satisfaction and salary adjustments are some of strategies that this study find useful to reduce employee attrition. By utilizing the tool developed under this study, organizations can make swift decision and implement proactive measures upon identifying high-risk employees. Ultimately, this would reduce the cost of recruiting and training new employees as Nuel et al. (2022) highlighted.

7.2.2 Data Solution Providers

Data solution providers are part of key stakeholders in contemporary business landscape. This study provides a clear approach of utilizing machine learning models to come up with a solution that predicts employee attrition and provides risk scores to assist organizations in formulating proactive measures towards retention. From the statistical perspective of the factors influencing employee attrition and analytical aspects to data product development, the data solution providers can benchmark and utilize the insights drawn from this study to come up with solutions that organizations can seamlessly integrate and implement optimized talent management plans.

7.2.3 Government

The business-related policies enacted by a government is more likely to affect companies within its jurisdiction (Mwangi, 2019). In light of this, government can leverage in this study's findings on the factors influencing employee attrition to review its policies to align with organizations requirements to ensure job stability, job satisfaction and fair compensation among employees. The need to assess the factors attributed to high attrition within various sectors as well as addressing them is paramount as most of them contribute to a countries economic development (Mwendwa, 2017).

7.3 Recommendation

This section provides a set of recommendation for organizational adoption for the solution developed under this study. A discussion of what can be explored in future research is also provided.

7.3.1 Organizational Adoption

To utilize the real-time employee attrition and risk-scoring system developed under this study, organizations must ensure that the product can be seamlessly integrated within their existing systems. The HR professionals must also be with clear understanding of the product usage, as they will make decisions and provide proactive measures towards retention based on the systems output. Their interpretation on risk scores will influence the effectiveness of the tool in talent management. Organizations must also ensure that the data from their employee base is of high quality to ensure accurate predictions.

7.3.2 Further Studies

This study focused on utilized machine learning techniques to develop a real-time employee attrition prediction and risk scoring tool. The variables that this study focused on were those provided in the IBM dataset and as such future research could explore the contribution of other features to employee attrition. Another research avenue is evaluating the effectiveness of more models in predicting employee attrition. These avenues will ensure the generalizability of the insights drawn from future studies.

References

Adeusi, K. B., Amajuoyi, P., & Benjami, L. B. (2024). Utilizing machine learning to predict employee turnover in high-stress sectors. *International Journal of Management & Entrepreneurship Research*, *6*(5), 1702-1732.

Ajmal M S, TANMAY DESHPANDE, IBM Data Scientists, February 17, 2023, "IBM HR Analytics Employee Attrition & Performance", IEEE Dataport, doi: https://dx.doi.org/10.21227/2m1g-6v47. Retrieved from: https://ieee-dataport.org/documents/ibm-hr-analytics-employee-attrition-performance

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), 7728668.

Brockett, N., Clarke, C., Berlingerio, M., & Dutta, S. (2019, December). A system for analysis and remediation of attrition. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2016-2019). IEEE.

Chankova, S., Muchiri, S., & Kombe, G. (2009). Health workforce attrition in the public sector in Kenya: a look at the reasons. *Human resources for health*, *7*, 1-8.

Chepkirui, A., & Atambo, W. (2024). HUMAN RESOURCES ANALYTICS AND EMPLOYEE PERFORMANCE IN TELECOMMUNICATIONS COMPANIES IN KENYA. *International Journal of Social Sciences Management and Entrepreneurship (IJSSME)*, *8*(2).

De Smet, A., Dowling, B., Hancock, B., & Schaninger, B. (2022). The Great Attrition is making hiring harder. Are you searching the right talent pools. *McKinsey Quarterly*, *58*, 1-13.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86.

Frye, A., Boomhower, C., Smith, M., Vitovsky, L., & Fabricant, S. (2018). Employee attrition: what makes an employee quit?. *SMU Data Science Review*, *1*(1), 9.

Fu, J. R. (2011). Understanding career commitment of IT professionals: Perspectives of push–pull–mooring framework and investment model. *International Journal of Information Management*, *31*(3), 279-293.

Gangai, K. N. (2013). Attrition at work place: how and why in hotel industry. *IOSR Journal of Humanities and social science*, *11*(2), 38-49.

George, S., Lakshmi, K. A., & Thomas, K. T. (2022, December). Predicting Employee Attrition Using Machine Learning Algorithms. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 700-705). IEEE.

Haldorai, K., Kim, W. G., Pillai, S. G., Park, T. E., & Balasubramanian, K. (2019). Factors affecting hotel employees' attrition and turnover: Application of pull-push-mooring framework. *International journal of hospitality management*, *83*, 46-55.

Ho, J. S. Y., Downe, A. G., & Loke, S. P. (2010). Employee attrition in the Malaysian service industry: Push and pull factors. *IUP Journal of Organizational Behavior*, *9*.

Holtom, B. C., Mitchell, T. R., & Lee, T. W. (2006). Increasing human and social capital by applying job embeddedness theory. *Organizational dynamics*, *35*(4), 316-331.

https://www.bls.gov/news.release/pdf/jolts.pdf

Iwu, C. G., Allen-Ile, C. O., & Ukpere, W. I. (2012). Key factors of employee satisfaction for the retention of health-related professionals in South Africa. *African Journal of Business Management*, *6*(39), 10486.

Jiang, K., Liu, D., McKay, P. F., Lee, T. W., & Mitchell, T. R. (2012). When and how is job embeddedness predictive of turnover? A meta-analytic investigation. *Journal of Applied psychology*, *97*(5), 1077.

Kanteh, L., & Gibba, A. (2019). A study on employees' attrition in public and private institutions in the Gambia, 2007-2017. *Arabian J Bus Manager Review*, *9*, 377.

Mehta, V., & Modi, S. (2021, December). Employee attrition system using tree-based ensemble method. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-4). IEEE.

Mhatre, A., Mahalingam, A., Narayanan, M., Nair, A., & Jaju, S. (2020, December).

Predicting employee attrition along with identifying high risk employees using big data and machine learning. In *2020 2nd international conference on advances in computing, communication control and networking (icacccn)* (pp. 269-276). IEEE.

Mpundu, M., Assan, T. E., & Mokoena, M. (2023). An analysis of high teacher turnover and attrition in the North-West Province of South Africa. *E-Journal of Humanities, Arts and Social Sciences*, *4*(4), 404-414.

Muda, V., Opoku, O. A., Anim, J., & Opoku-Dadzie, I. (2022). The effect of job satisfaction on staff retention and attrition at GCB bank plc in upper east region of Ghana. *Journal of Corporate Finance Management and Banking System (JCFMBS) ISSN*, 2799-1059.

Mwangi, J. W. (2021). *Causes of Attrition on Organization Performance in The Telecommunication Industry in Kenya A Case of Airtel Kenya Limited* (Doctoral dissertation, Daystar University, School of Business and Economics).

Mwendwa, L. (2017). *Factors influencing call center agent attrition: A case of Kenya Power call center* (Doctoral dissertation, University of Nairobi).

Narayanan, A. (2016). Talent management and employee retention: Implications of job embeddedness-a research agenda. *Journal of Strategic Human Resource Management*, *5*(2).

Nuel, O. I. E., Ifechi, A. N., & Chike, N. K. (2022). Staff Attrition and Performance of Selected Manufacturing Firms in Southeast Nigeria. *Journal of International Business and Management*, *5*(4), 01-12.

Poornappriya, T. S., & Gopinath, R. (2021). Employee attrition in human resource using machine learning techniques. *Webology*, *18*(6).

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting employee attrition using machine learning approaches. *Applied Sciences*, *12*(13), 6424.

Srivastava, P. R., & Eachempati, P. (2021). Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach. *Journal of Global Information Management (JGIM)*, *29*(6), 1-29.

Strathmore University Institutional Scientific Ethics Review Committee. (n.d.). *Ethics review*. Strathmore University. https://research.strathmore.edu/ethics-review/

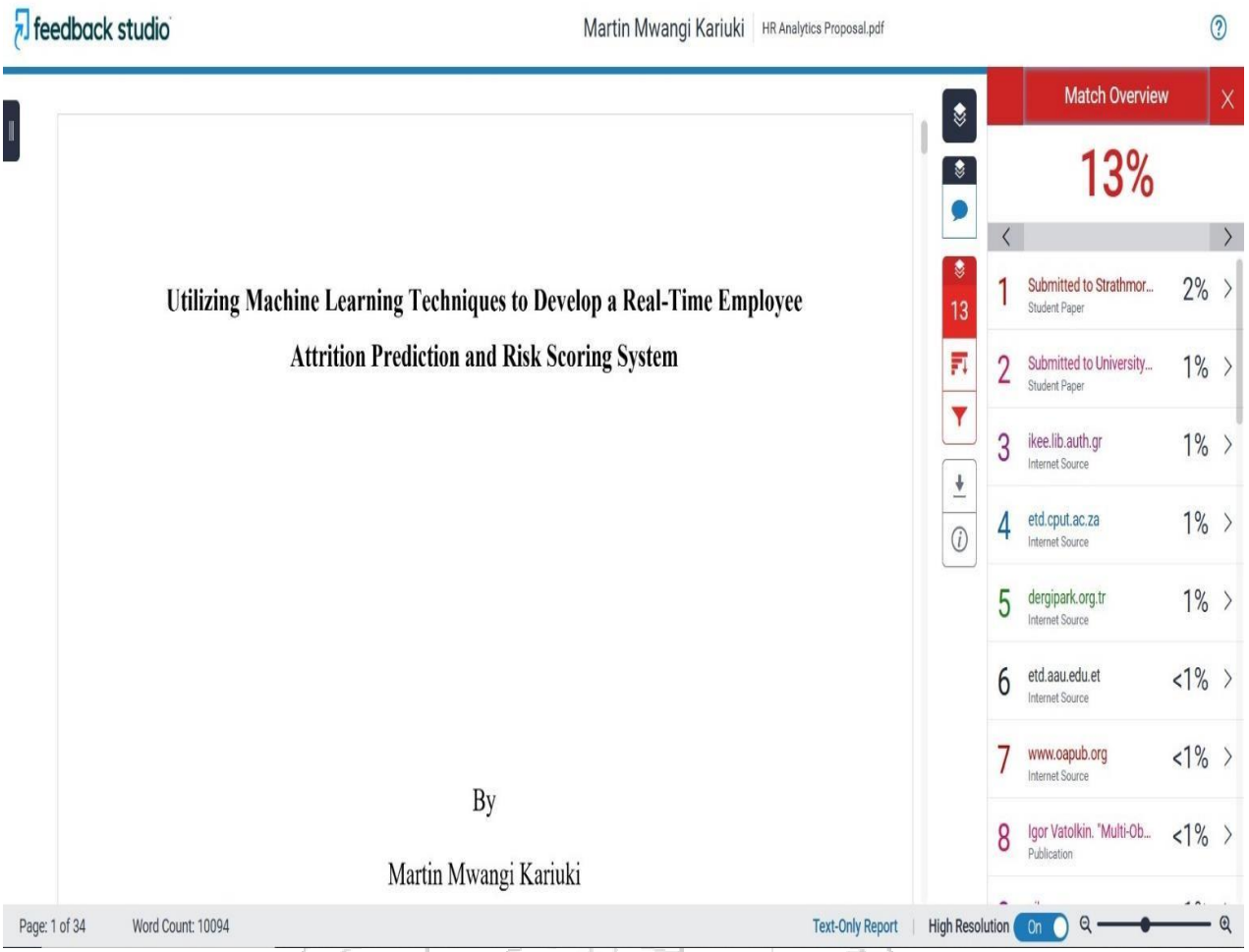
Whitton, R. J. (2023). Exploring Factors for Employee Attrition and Retention by Life Stage.

Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From big data to deep data to support

people analytics for, employee attrition prediction. *Ieee Access*, *9*, 60447-60458.

Appendices

Appendix A: Similarity Report



Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

22nd January 2025

Mr Kariuki Martin,
martin.kariuki2021@strathmore.edu

Dear Mr Kariuki,

RE: Utilizing Machine Learning Techniques to Develop a Real-Time Employee Attrition Prediction and Risk Scoring System

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2478/25.** The approval period is from **22nd January 2025 to 21st January 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C: System Implementation

C.1: System Responsiveness

```
+     output$feature_importance <- renderPlot({
+       ggplot(imp, aes(x = reorder(Feature, Importance), y = Importance)) +
+         geom_bar(stat = "identity", fill = "steelblue") +
+         coord_flip() +
+         labs(title = "Dynamic Feature Importance", x = "Feature", y = "Importance")
+     })
+
+     end_time <- Sys.time()
+     execution_time <- end_time - start_time
+     print(paste("Execution Time:", execution_time))
+   })
+ }
>
> # Run the application
> shinyApp(ui = ui, server = server)

Listening on http://127.0.0.1:4399
[1] "Execution Time: 0.118211984634399"
[1] "Execution Time: 0.0891909599304199"
[1] "Execution Time: 0.0931580066680908"
[1] "Execution Time: 0.111138105392456"
[1] "Execution Time: 0.0893170833587646"
[1] "Execution Time: 0.0900301933288574"
[1] "Execution Time: 0.0904459953308105"
[1] "Execution Time: 0.119696855545044"
```

Figure C.1: Code snippet showcasing system responsiveness