การทำนายการลาออกของพนักงานใหม่โดยใช้การวิเคราะห์เชิงคาดการณ์ : กรณีศึกษาบริษัทการเงินไทยขนาดใหญ่ในกรุงเทพมหานครฯ ประเทศไทย

น.ส.มีนา กิตติคุณศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2566

PREDICTING NEWCOMER'S TURNOVER USING PREDICTIVE ANALYTICS : A CASE STUDY OF THAI FINANCIAL FIRM IN BANGKOK, THAILAND

Ms. Meena Kittikunsiri

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2023

Thesis Title	PREDICTING NEWCOMER'S TURNOVER USING PREDICTIVE ANALYTICS : A CASE STUDY OF THAI FINANCIAL FIRM IN BANGKOK, THAILAND
By	Ms. Meena Kittikunsiri
Field of Study	Computer Science
Thesis Advisor	Assistant Professor Dr.Natawut Nupairoj

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

Dean of the Faculty of Engineering

.....

(Professor Dr. Supot Teachavorasinskun)

THESIS COMMITTEE

..... Chairman

(Associate Professor Dr. Peerapon Vateekul)

..... Thesis Advisor

(Assistant Professor Dr.Natawut Nupairoj)

..... External Examiner

(Dr. Titipat Achakulvisut)

มีนา กิตติคุณศิริ: การทำนายการลาออกของพนักงานใหม่โดยใช้การวิเคราะห์เชิงคาดการณ์ : กรณีศึกษาบริษัทการเงินไทยขนาดใหญ่ในกรุงเทพมหานครฯ ประเทศไทย. (PREDICTING NEWCOMER'S TURNOVER USING PREDICTIVE ANALYTICS : A CASE STUDY OF THAI FINANCIAL FIRM IN BANGKOK, THAILAND) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ณัฐวุฒิ หนูไพโรจน์, 38 หน้า.

การลาออกของพนักงานเป็นหนึ่งในปัญหาสำคัญที่ส่งผลต่อประสิทธิภาพการดำเนินงานของบริษัท การใช้การวิเคราะห์เชิงคาดการณ์เพื่อทำนายการลาออกของพนักงานจึงเป็นสิ่งที่ผู้บริหารและนักทรัพยากรบุคคลให้ความสนใจ งานวิจัยนี้มุ่งเน้นไปที่การทำนายการลาออกของพนักงานใหม่จากข้อมูลแบบสอบถามพนักงานใหม่ของบริษัทการเงินไทยขนาดใหญ่ในกรุงเทพมหานครฯ ประเทศไทย โดยใช้โมเดลการเรียนรู้ของเครื่อง (Machine Learning Model) ผลการทดลองพบว่า Random Forest ได้ผล F1 Score สูงที่สุด นอกจากนั้นงานวิจัยนี้ยังค้นพบปัจจัยที่ส่งผลต่อการลาออกของพนักงานใหม่มากที่สุด เช่น ความพึงพอใจต่อวัฒนธรรมองค์กร นโยบายการ Work-From-Home โปรแกรมดูแลพนักงานใหม่ และความพึงพอใจต่อการขั้นตอนการสรรหาบุคลากร ผลลัพธ์จากงานวิจัยนี้ช่วยให้ผู้บริหารและเจ้าหน้าที่ทรัพยากรบุคคลรู้แนวทางในการปรับปรุงและสร้างประสบการณ์ที่ดีและนำไปสู่การลดการลาออกของพนักงานใหม่ลงได้

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต	.....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาหลัก	.....
ปีการศึกษา	2566		

6370237421: MAJOR COMPUTER SCIENCE

KEYWORDS: PEOPLE ANALYTICS/ APPLIED DATA SCIENCE/ HUMAN RESOURCES MANAGEMENT/ CHURN MODEL/ EMPLOYEE TURNOVER

MEENA KITTIKUNSIRI : PREDICTING NEWCOMER'S TURNOVER USING PREDICTIVE ANALYTICS : A CASE STUDY OF THAI FINANCIAL FIRM IN BANGKOK, THAILAND. ADVISOR : Asst. Prof. Natawut Nupairoj, 38 pp.

Employee turnover, a critical issue impacting workplace productivity, has prompted organizations to leverage machine learning techniques for predictive analysis. This study specifically targets the prediction of turnover among new employees, utilizing data obtained from a survey conducted at a Thai financial firm in Bangkok, Thailand. Through an evaluation of various machine learning models, the results indicate that the Random Forest model surpasses others. Furthermore, this research highlights crucial factors influencing newcomer turnover, such as comfort with workplace culture, work-from-home policies, onboarding programs, and satisfaction with the recruitment process. These findings offer actionable insights for HR professionals to focus on these specific areas and improve the experience for new employees, thereby enhancing their retention within the organization.

Department:	Computer Engineering	Student's Signature	.....
Field of Study:	Computer Science	Advisor's Signature	.....
Academic Year:	2023		

CONTENTS

Page

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Chapter I

INTRODUCTION

Employee attrition is a prevalent issue that organizations worldwide face, defined as the reduction of manpower due to voluntary resignation or quitting of employees [1]. This challenge is particularly significant in today's intensely competitive global market, and Human Resources (HR) managers consider employee attrition to be a major concern [2]. The financial implications of employee attrition include recruitment and training costs, loss of talented employees, and decreased productivity [3]. These financial consequences are exacerbated when employees quit shortly after joining the organization, which is known as newcomer attrition. According to the US Bureau of Labor Statistics, resignation rate of workers who have stayed for a year or less at a company was 45 percent in 2020 [4], while it takes an average of 8.2 months for employees to achieve role clarity and 23.2 months to master their roles [5]. This means that organizations lose the opportunity to fully utilize and gain a return on their investment in newcomers' productivity, and must bear the costs of recruiting and onboarding new employees. Therefore, addressing newcomer attrition is crucial for organizations to maintain their productivity and competitiveness.

Data science has become a powerful tool for predicting customer churn models and has been applied in various fields including Human Resources Management [6]. There is a growing body of case studies demonstrating the successful application of machine learning (ML) techniques in predicting employee attrition or churn across diverse demographics and industries, including global retail, telecommunications, and technology companies [7]. ML applications for employee attrition have also been observed in the pharmaceutical industry [8] and the university setting [9]. And these studies showcase the broad adoption of ML techniques for employee attrition prediction in various countries worldwide, such as the USA, the UK, India [7], Indonesia [6], Belgium, Taiwan, and Thailand [9]. The application of ML in this domain highlights its potential for enhancing workforce management and

retention strategies. Employee Attrition Prediction can assist HR managers in forecasting how many employees will leave the company within a specific timeframe. This enables managers to prioritize valuable employees and implement measures to retain them [10].

However, despite the growing research on employee churn models, there is a relative lack of specific studies on predicting newcomer turnover. This research gap highlights the need for further exploration of predictive analytics in predicting newcomer attrition. Addressing the challenge of high newcomer turnover and ensuring a positive return on Human Resources investment in organizations requires a deeper understanding of this specific segment.

A study by [11] reviewed employee turnover theories and research over the last 100 years. They suggested that turnover research should expand to capture contextual factors and move away from a "one size fits all" approach. Focusing solely on newcomer attrition is one way to achieve this goal.

Therefore, this work focuses on developing a predictive model for newcomer churn. The study utilizes employee data, including 30 attributes, from 132 employees in one of the largest financial firms in Bangkok, Thailand. The proposed research will contribute to the literature by filling the gap in HR data science research on predicting newcomer turnover using predictive analytics. The findings will provide practical implications for HR managers and leaders to identify and retain new hires in their organizations. Moreover, the results will contribute to the academic community by enhancing the understanding of newcomer turnover prediction.

Chapter II

REVIEW OF LITERATURE

2.1 Employee Attrition

Since the birth of turnover research in 1917 [11], there are several theories that have been developed to explain employee turnover. These theories provide frameworks for understanding the underlying factors and processes that contribute to employees' decisions to leave an organization.

Job satisfaction theory suggests that employees' level of satisfaction with their work plays a significant role in turnover decisions. If employees are dissatisfied with various aspects of their job, such as pay, work environment, or relationships with supervisors, they are more likely to consider leaving the organization [12]

Job embeddedness theory posits that employees' decisions to stay or leave an organization are influenced by their connections and fit within the job, community, and organization [13]. It suggests that employees who have stronger links to their job, colleagues, and community are less likely to turnover, as they perceive high costs associated with leaving.

Identity theory suggests that employees' self-identities and the alignment of their identities with the organization play a role in turnover decisions. When employees experience a discrepancy between their personal identity and the organizational identity, it can lead to disengagement and ultimately turnover. This theory highlights the importance of organizational culture, values, and employee identity integration [14].

Push-Pull-Mooring Framework provides a comprehensive perspective on employee turnover. It considers both the "push" factors (dissatisfaction with the current job) and the "pull" factors (attraction to alternative job opportunities). Additionally, it intro-

duces the concept of "mooring," which refers to factors that keep employees in their current organization, such as social connections, organizational commitment, and financial stability. This framework helps to understand the interplay between different forces influencing turnover decisions [15]

In Table 1, we present a summary of factors that have been identified in previous employee turnover researches.

Category	**Factors related to Employee Turnover**
Job Fit	Job mismatch with employee's expectations [16] Salary and various benefits [16; 17; 18], Job mismatch with employee's personality [16] Job mismatch with employee's values or beliefs [16] Lack of job pride [16; 18] Work-life imbalance and increased need for flexibility [19]
Supervisor/Manager	Poor behavior from superiors [20] Lack of recognition [16; 18]
Working Relationship	Unhealthy working relationships [17]
Organization Fit	Loss of trust in receiving support [16] Feeling incompatible with the organization's culture or desiring a better culture [21; 22]
Development	Lack of career development opportunities [23] Underutilization of skills and abilities [24] Insufficient training and development [18; 25]
Autonomy	Participation in decision making [25]
Job security	Job insecurity [25]
Outside Factors	Job offers from other organizations [16]
Overall Satisfaction	Job-related stress and lack of happiness [16]

Table 1: Factors related to Employee Turnover

2.2 Newcomer Attrition

Newcomer attrition is the phenomenon of newly hired employees leaving an organization shortly after starting their employment. It is influenced by factors such as task mastery, role clarity, and social acceptance [5].

Task mastery involves the acquisition and improvement of job-related skills and knowledge. On average, it takes newcomers approximately 21.6 months to reach a level of proficiency in their tasks.

Role clarity refers to newcomers' understanding of their job expectations, responsibilities, and specific contributions within the organization. It takes an average of 8.29 months for newcomers to gain clarity about their roles.

Social acceptance pertains to the extent to which newcomers are accepted, welcomed, and integrated into the social fabric of the organization. It typically takes around 6.27 months for newcomers to feel a sense of acceptance and belonging in their work environment.

According to the SHRM Customized Human Capital Benchmarking Report, the average tenure before employees voluntarily leave the organization is 8 years, with a median tenure of 4.3 years. Therefore, the early resignation of new employees, particularly within the first 6 months of employment, is not a common occurrence and should be of significant concern to HR and managers.

Table 2 presents factors related to new hire turnover based on HR research.

Category	**Factors related to New Hire Turnover**
Job Fit	Job mismatch with employee's expectations (Work Institute, 2020)
Supervisor/Manager	Poor behavior from superiors [26]
Development	Lack of career development opportunities (Work Institute, 2020)
Onboarding	Bad onboarding program [26]

Table 2: Factors related to New Hire Turnover

2.3 Machine learning and employee Attrition

Machine Learning (ML) has become a widely used technology in various business domains, including Human Resources (HR). Although the adoption of ML in HR departments may have been slower compared to other departments, there is a growing number of case studies that demonstrate the successful application of ML in predicting employee attrition or churn [2]. These studies have utilized ML techniques for employee churn prediction across diverse demographics and industries, including global retail, telecommunications, technology companies [7], pharmaceuticals [8], and universities[9]. Moreover, ML applications for employee attrition have been observed in numerous countries worldwide, including the USA, the UK, India [7], Indonesia [6], Belgium , Taiwan, and Thailand [9].

A review of the literature and research background found that human resource data in predicting employee attrition is based on three main sources.

1) HRIS (Human Resource Information System): Data obtained from the HRIS or HR system, which includes employee information such as age, gender, marital status, years of service in the company, previous employment history, salary, other benefits, workplace location, commuting distance, department, job position, job type (e.g., permanent or contract), education level, grade, frequency of business travel, stock options, years with current manager, latest performance evaluation, latest salary increase percentage, latest promotion, training received in the past year, years of work experience, overtime work data, number of dependents or family members supported, number of children, and number of leave days taken in the previous year [27; 6; 7; 9].

2) Survey Data: Data collected through surveys, including job satisfaction level, environment satisfaction level, work-life balance rating, organizational culture and values, job security, growth opportunities, perceived freedom to work, clarity of roles and responsibilities, recognition from supervisors, relationship with supervisors, relationship with colleagues, and sense of pride in work [28; 29].

3) External Data: Data from external sources, such as unemployment rates and

average household income, obtained from government organizations [27].

Most commonly, predictive models are built using data from a single source, either from HRIS or surveys. However, there are cases where data from both sources are combined for analysis. Additionally, apart from using ML techniques for predicting employee attrition or identifying reasons for attrition, data from exit interviews are also used for content analysis to obtain comprehensive insights into the reasons behind employee departures [30].

The below tables briefly documents the literature review findings.

Author	**Problem Study**	**Data Used**	**Model Used**	**Model Recommendation**
[2]	Compare predictive employee churn models	1575 employees and 25 attributes	SVM, Random Forests and Naïve Bayes	SVM
[27]	Prediction of Employee Turnover in global retailer	33 attributes from HRIS and Bureau of Labor Statistics Data	XGBoost, Logistic, Regression, Naïve, Bayesian, Random Forest, SVM, LDA, KNN	XGBoost
[31]	Prediction of Employee Turnover using dataset from Kaggle	14,999 records and 12 attributes including satisfaction, Salary, etc.	Logistic Regression, SVM, Random Forest, Decision Tree and AdaBoost	Random Forest and Decision Tree

Table 3: Summary of Research Papers (Part 1)

Author	**Problem Study**	**Data Used**	**Model Used**	**Model Recommendation**
[6]	Prediction of Employee Turnover in Indonesian telecommunication company	HRIS data including age, gender, base salary, Postion, etc.	Naïve Bayes, Decision Tree, and Random Forest	Random Forest
[32]	Prediction of Employee Turnover using dataset from IBM Watson Analytics	1470 employees with 32 features such as Job Level, Total Working Years, etc.	SVM, Random Forest, KNN	KNN
[30]	Prediction of Employee Turnover during a job transition	Dataset of anonymously submitted resumes through Glassdoor's online portal	Linear Regression, Logistic Regression, Decision Tree, and Random Forest	Random Forest and Decision Tree
[33]	Comparison of different Machine Learning methods using HR dataset from IBM Watson Analytics	14,999 records and 12 attributes	Logistic Regression, SVM, Random Forest, Decision Tree and AdaBoost	Random Forest and Decision Tree

Table 4: Summary of Research Papers (Part 2)

Author	**Problem Study**	**Data Used**	**Model Used**	**Model Recommendation**
[29]	Prediction of Employee Turnover in an IT sector	Questionnaire including satisfaction level, work life balance, etc.	Naïve Bayes, Decision Tree, Random Forest, J48 and K-Means	Naïve Bayes
[10]	Prediction of Employee Turnover using dataset from IBM Watson Analytics	1470 employees with 32 features such as Job Level, Total Working Years, etc.	Logistic Regression, Naive Bayes Classifier, Random Forest Classifier, and XGBoost	XGBoost
[9]	Employee churn prediction in higher education	University HRIS data including Education Level, Working Year, etc.	Naïve Bayes, Decision Tree and Random Forest, logistic regression, and SVM	(Work in progress)
[34]	Prediction of Employee Turnover in Home Care Industry	Data extracted from Alaya Software including Average hour per visit	Logistic Regression, Random Forests, XGBoost, and Multi-layer Perceptron	XGBoost
[35]	Prediction of Employee Turnover using Kaggle Dataset	14,999 records and 12 attributes.	Naive Bayes., Logistic Regression, KNN, Multilayer Perceptron (MLP)	KNN

Table 5: Summary of Research Papers (Part 3)

A common approach in previous studies for building predictive models is to employ multiple algorithms and evaluate their performance to determine the most effective algorithm. Typically, a combination of 3 to 5, or even up to 7, algorithms is used in such cases. These algorithms include SVM, Random Forest, Decision Tree, J48, Naïve Bayes, XGBoost, Logistic Regression, LDA, KNN, ADABoost, and Multi-layer Perceptron. And the algorithms that have been consistently recommended and ranked highly in terms of performance include Random Forest, and Decision Tree.

However, despite the growing research on employee prediction, there is a relative lack of research on newcomer attrition prediction and the reasons why newcomers leave. Therefore, this paper will focus specifically on using Machine Learning algorithms to predict newcomer attrition and report important features that relate most to the decision.

Chapter III

METHODOLOGY

In this paper, the term "Newcomer" refers to an employee who has recently joined an organization and has been working for a duration not exceeding the average time it takes to achieve task mastery, role clarity, and social acceptance, which is 12.05 months [5]

The study is divided into 5 phases. The first phase is data collection from Newcomer Survey. The second phase is data preprocessing which involves Exploratory Data Analysis (EDA), data upsampling and fixing issue of imbalanced dataset. In the third phase, which is model fitting, we use multiple models to predict newcomer churn. We then last compare those models in terms of Precision, Recall and F1 Score.

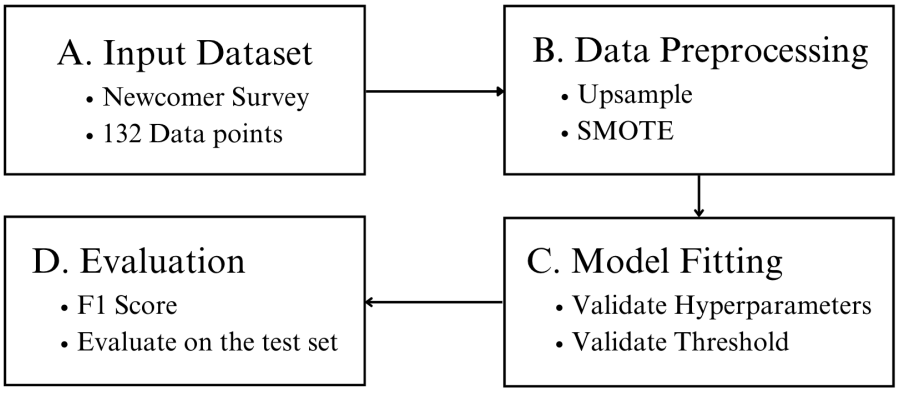

Figure 3.1: Research Stages

3.1 Input Dataset

In this study, data was collected through a Newcomer Survey administered to 132 newly hired employees (with less than 12 months of tenure) at one of Thailand's top three financial companies. The received data was anonymized, with the removal of names and sensitive information, ensuring the protection of privacy and confidentiality. The survey

consisted of 30 attributes as following.

Category	**Features**	**Data Type**
Demographic	Age	Numeric
	Department	Categorical
	Working Months In Organisation	Numeric
Job-Fit	Preferred WFH Percentage	Numeric
	Satisfaction On WFH Policy	Numeric
	Work is As Expected	Numeric
	(Onsite) Satisfaction on Workplace	Numeric
	(Onsite) Satisfaction on Facility at Office	Numeric
	(WFH) Org Support Facility to WFH	Numeric
Supervisor/Manager	Warm Welcome From Supervisor	Numeric
	Supervisor Support To Get Job Done	Numeric
	Supervisor State Clear Expectation And Responsibility	Numeric
	Supervisor Suggest And Follow Up During Probation Period	Numeric
Working Relationship	Teammates give warm welcome and advise	Numeric
	Know who to contact to get the job done	Numeric
	Manager assign buddy	Numeric
	Buddy helps suggest about work and connection	Numeric
	Buddy is friendly and share experience	Numeric
	Feel engaged to be in the team	Numeric
Organisation-Fit	Comfortable with Organisation Culture	Numeric
	Comfortable with Department Culture	Numeric
	Comfortable with Team Culture	Numeric
Recruitment Process and Onboarding	Enough Job Description Before Applying For This Job	Numeric
	Job Application Process is User-Friendly	Numeric
	Get Enough Role and Responsibility Information During Interview Process	Numeric
	Satisfaction on Recruitment Process	Numeric

Table 6: New Hire Dataset Attributes (Part 1)

Category	Features	Data Type
Recruitment Process and Onboarding	Onboarding Program is Effective to Prepare Employee For Work	Numeric
	Overall Satisfaction on Onboarding Program	Numeric
Overall Satisfaction	Working with This Company is The Right Decision	Numeric
	Feel Energized and Enthusiastic to Work	Numeric
Employee Decision	Attrition	Categorical

Table 7: New Hire Dataset Attributes (Part 2)

By focusing on these specific factors related to onboarding and HR processes, this study aims to gain insights into the unique challenges and opportunities in predicting new hire attrition.

3.2 Data Preprocessing

Exploratory Data Analysis (EDA)

Prior to conducting the exploratory data analysis (EDA), it is important to establish hypotheses for further investigation.

Here are Hypothesis on New Hire Churn :

1. Younger newcomers may be more prone to attrition due to their age and relatively lower levels of responsibility and commitments. Being younger, they often have fewer family and financial obligations, making it easier for them to consider job changes or relocate for better opportunities. Their greater mobility and flexibility in terms of career choices and geographical location could contribute to a higher attrition rate compared to older employees.

2. The low level of comfortability with the organizational culture, department culture, and team culture among newcomers may have a strong correlation with attrition.

3. The low level of satisfaction with the work-from-home policy could be strongly associated with attrition, as individuals are increasingly seeking job opportunities that provide the flexibility they desire.

4. The low level of satisfaction with recruitment and onboarding processes could be strongly associated with attrition, as these initial experiences shape newcomers' perceptions and their decision to stay or leave the organization.

Fixing small and imbalanced dataset issues

In our specific dataset, a survey dataset with all requisite questions, no instances of missing data or duplicate records were observed. To address the issues related to small samples and data imbalance, the following data preprocessing steps were implemented.

1. Data Split using Stratified K-Fold :

The dataset, consisting of 132 instances, underwent division into three segments using the Stratified K-Fold method. Each segment served as the testing data for three folds. In each fold, the training set comprised 88 instances, while the testing set contained 44 instances, maintaining a ratio of 67 percent for training and 33 percent for testing. Notably, the testing sets for fold 1, fold 2, and fold 3 included 2, 2, and 3 instances labeled as '1' or 'leavers', respectively.

2. Oversampling of Training Data :

Within each fold, instances labeled as '1' or 'leavers' in the training set underwent ten-fold upsampling. Consequently, the training dataset was duplicated, followed by the application of the Synthetic Minority Over-sampling Technique (SMOTE) to further address the challenge of data imbalance.

Table 8 illustrates the number of data at each stage of the data preprocessing steps.

Fold	Training set (After)				Testing set
	Stratified K-Fold	Ten-fold Upsample	Duplicate	SMOTE	
1	Total 88 '0' : 83 '1' : 5	Total 133 '0' : 83 '1' : 50	Total 266 '0' : 166 '1' : 100	Total 332 '0' : 166 '1' : 166	Total 44 '0' : 42 '1' : 2
2	Total 88 '0' : 83 '1' : 5	Total 133 '0' : 83 '1' : 50	Total 266 '0' : 166 '1' : 100	Total 332 '0' : 166 '1' : 166	Total 44 '0' : 42 '1' : 2
3	Total 88 '0' : 84 '1' : 4	Total 124 '0' : 84 '1' : 40	Total 248 '0' : 168 '1' : 80	Total 336 '0' : 168 '1' : 168	Total 44 '0' : 41 '1' : 3

Table 8: Number Of Data After Preprocessing Steps

3.3 Model Construction

Due to the lack of specific studies on predicting newcomer turnover, we implemented various Machine Learning models recommended by existing employee turnover research. These models include Decision Tree, Random Forest, XGBoost, SVM, Naive Bayes, and KNN. Additionally, we included a basic model, Logistic Regression, as a benchmark.

To optimize model performance, a GridSearchCV was conducted on the training set to fine-tune the model parameters. The predict_proba function was also employed on the training set to determine the optimal threshold. An extensive search spanned thresholds ranging from 0.2 to 0.9, with an increment of 0.01. The models were then trained using their optimal configurations on the training dataset. Finally, the trained models were applied to the testing set, which constituted 33 percent of the data.

Table 9 shows strengths and limitations of the chosen Machine Learning models.

Model	Strengths	Limitations
Decision Tree	Able to capture non-linear relationships in the data, effective with the balanced dataset obtained after preprocessing.	Prone to overfitting in scenarios with a limited number of instances labeled as 'leavers'.
Random Forest	Robust to overfitting, effective in handling imbalanced datasets.	May have reduced performance in very small datasets.
XGBoost	High predictive performance; well-suited for the dataset with oversampling and SMOTE.	The complexity of XGBoost may not be fully utilized in a relatively small dataset.
SVM	Simple and efficient in high-dimensional spaces, suitable for small to medium-sized datasets.	May not perform optimally with very small datasets. The model could face challenges in learning complex patterns and might be prone to overfitting.
Naive Bayes	Simple and efficient, works well with small datasets; potentially effective with newcomer data.	Assumes feature independence, which is not fully aligned with the characteristics of this dataset.
KNN	Effective in capturing local patterns and works well with small datasets.	Performance may degrade with high-dimensional data, and it's sensitive to irrelevant features.
Logistic Regression	Simple and interpretable, efficient for small datasets; suitable for the newcomer dataset.	Assumes linear relationships, which may not capture the complexity of the underlying patterns.

Table 9: Strengths and Limitations of the Chosen Algorithms

3.4 Model Comparison

In the field of Human Resources, one of the primary goals is to accurately pinpoint potential leavers while understanding the reasons behind their departure. Given the

relatively low number of individuals leaving compared to those staying (7 out of 132), using accuracy as an evaluation metric can be misleading, as predicting all individuals as stayers would already yield a high accuracy of 0.947. This study emphasizes the balance between maximizing Recall to capture leavers and maintaining Precision to control false positives. The F1 score will be used as the metric of choice, providing a concise yet comprehensive evaluation in the context of HR and newcomer churn.

Parameter	**Description**
Precision	The proportion of true churners among the predicted churners
Recall	The proportion of true churners that are correctly identified by the model
F1-Score	The harmonic mean of precision and recall, providing a balanced measure

Table 10: Model Evaluation Parameters

Chapter IV

RESULTS

4.1 EDA Results

Hypothesis 1

Younger newcomers are more prone to attrition due to their age and relatively lower levels of responsibility and commitments.

The average age of newcomers who participated in this survey is 27.63, ranging from 22 to 45 years old. Figure 4.1 indicates that a significant portion of newcomers falls within the early 20s age group.

Based on the analysis of Figure 4.1, it is evident that the age group of 40 has the highest percentage of leavers, followed by the age group of 38. Surprisingly, the age group in their 20s does not exhibit the highest attrition rate as initially hypothesized. Therefore, the first hypothesis stating that younger newcomers have a higher attrition rate compared to older individuals is refuted by the findings from the data.

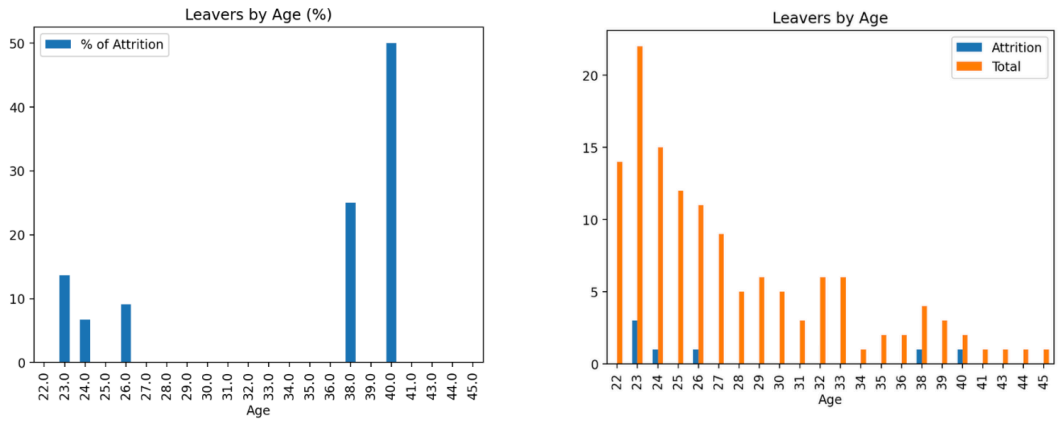

Figure 4.1: Leaver by Age

Hypothesis 2

The low level of comfortableness with the organizational culture, department culture, and team culture among newcomers have a strong correlation with attrition.

The analysis of Figures 4.2, 4.3, and 4.4 reveals a consistent trend: a clear correlation between lower satisfaction levels in organizational, departmental, and team cultures and higher attrition percentages. This suggests that employees who experience lower comfortability in these cultural aspects are more likely to leave the organization. Therefore, hypothesis 2, which posits a connection between culture satisfaction and attrition, is supported.

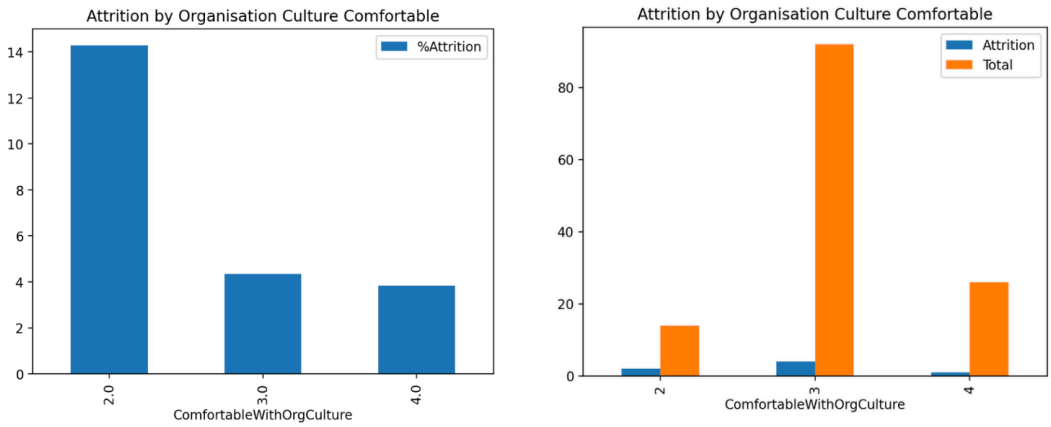

Figure 4.2: Attrition by Organisation Culture Comfortableness

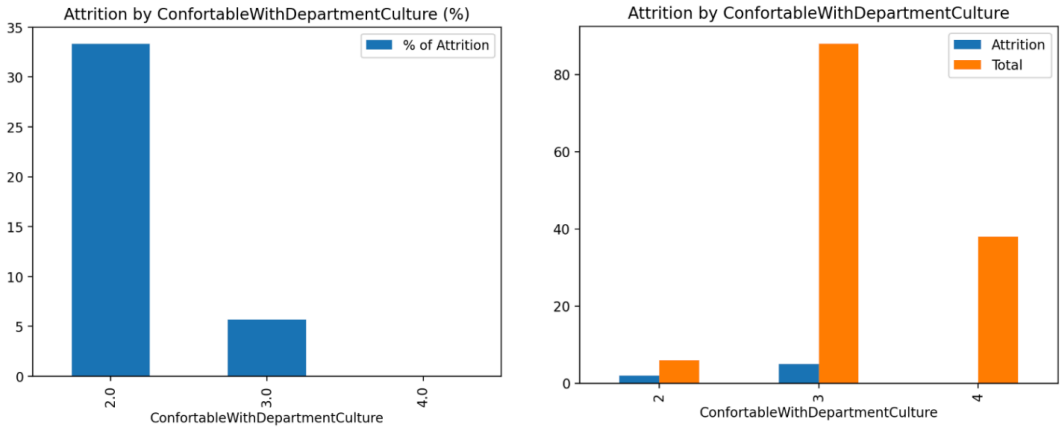

Figure 4.3: Attrition by Department Culture comfortableness

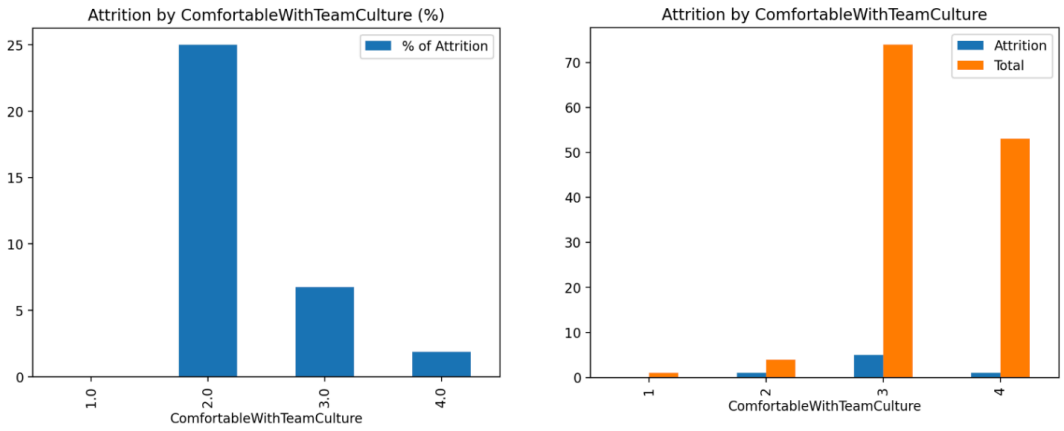

Figure 4.4: Attrition by Team Culture comfortableness

Hypothesis 3

The low level of satisfaction with the work-from-home policy could be strongly associated with attrition.

The insights from Figure 4.5 demonstrate a notable pattern: newcomers who express low satisfaction levels regarding the work-from-home policy exhibit the highest attrition rates. This suggests that individuals who are dissatisfied with the flexibility and remote work options provided by the organization are more likely to leave. Hence, hypothesis 3 is supported.

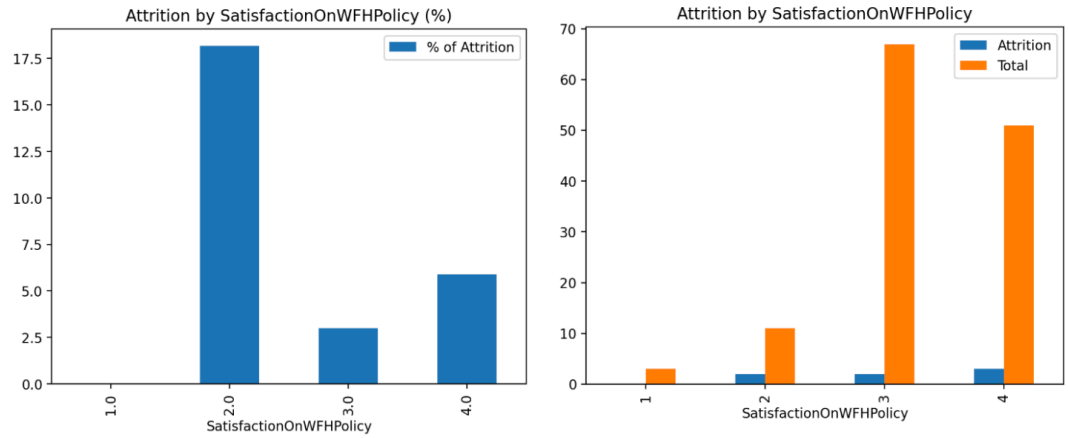

Figure 4.5: Attrition by Satisfaction on WFH Policy

Hypothesis 4

4. The low level of satisfaction with recruitment and onboarding processes could be strongly associated with attrition.

Insights from Figures 4.6 and 4.7 demonstrate a significant relationship between lower satisfaction with the recruitment and onboarding process and higher attrition rates. This finding supports hypothesis 4.

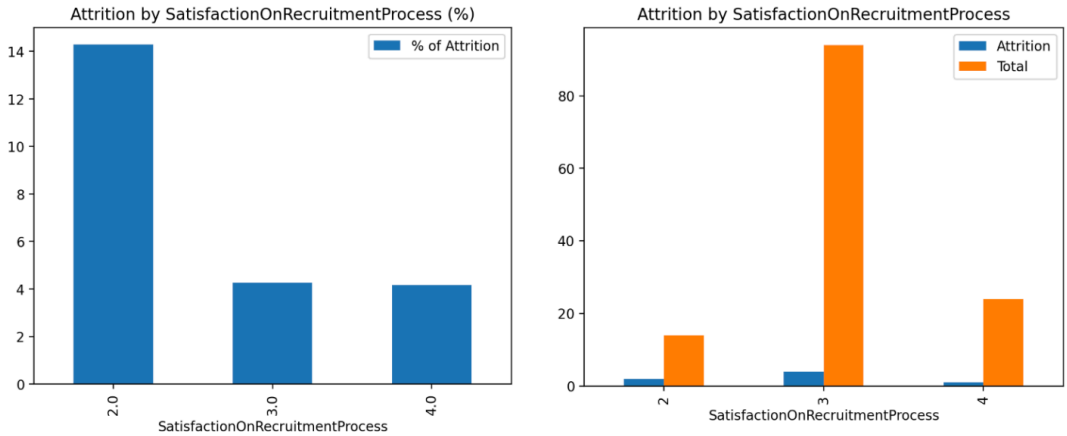

Figure 4.6: Attrition by Satisfaction on Recruitment Process

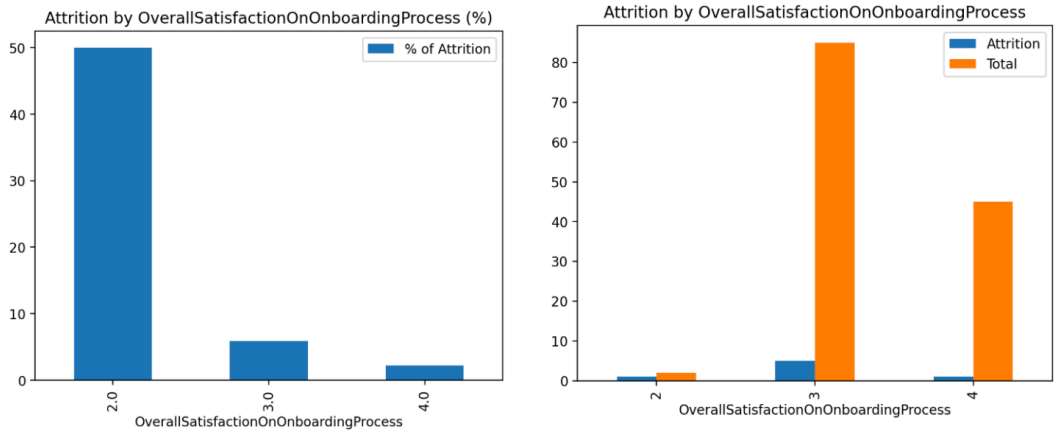

Figure 4.7: Attrition by Satisfaction on Onboarding Process

4.2 Results from Machine Learning Models

Table 11 displays the average ± standard deviation (SD) of results for each model, while Table 12 presents the results from combining three folds.

Model	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**	**Accuracy**
Decision Tree	(0.056, 0.878)	(0.106, 0.672)	(0.047, 0.659)	(0.428, 0.836)	(0.711, 0.985)
Random Forest	**(0.412, 0.632)**	**(0.732, 1.000)**	**(0.590, 0.680)**	**(0.978, 1.000)**	**(0.936, 0.958)**
XGBoost	(0.500, 0.500)	(0.514, 0.930)	(0.511, 0.647)	(0.932, 0.998)	(0.936, 0.958)
SVM	(0.092, 0.908)	(0.070, 0.486)	(0.097, 0.569)	(0.421, 1.000)	(0.881, 0.967)
Naive Bayes	(0.000, 0.268)	(0.000, 0.268)	(0.000, 0.268)	(0.498, 0.626)	(0.729, 0.967)
KNN	(0.000, 0.268)	(0.000, 0.268)	(0.000, 0.268)	(0.498, 0.626)	(0.729, 0.967)
Logistic Regression	(0.196, 1.000)	(0.070, 0.486)	(0.106, 0.672)	(0.594, 0.812)	(0.919, 0.975)

Table 11: Average Performance Metrics with Standard Deviations (SD)

Model	Precision	Recall	F1 Score	ROC-AUC	Accuracy
Decision Tree	0.158	0.429	0.231	0.650	0.848
Random Forest	**0.500**	**0.857**	**0.632**	**0.905**	**0.945**
XGBoost	0.500	0.714	0.588	0.837	0.947
SVM	0.286	0.286	0.286	0.623	0.924
Naive Bayes	0.066	0.143	0.091	0.515	0.848
KNN	0.250	0.143	0.182	0.559	0.939
Logistic Regression	0.500	0.286	0.364	0.635	0.947

Table 12: Performance After Combining Three Folds

Even though Table 10 and 11 present all evaluation metrics, this paper will specifically focus on Precision, Recall, and F1 Score. Both tables show that Random Forest outperforms the other models in terms of Recall and F1 Score.

Table 13 shows best hyperparameters of each algorithms.

4.3 Identify most important features using SHAP Value

SHAP (SHapley Additive exPlanations) values are a technique used in Machine Learning to explain the output of a model by calculating the contribution and impact of each feature on the prediction.

After applying this technique, we identified the top 10 features that have the most impact :

1. Comfort with Department Culture
2. Organisation Support Facility to Work From Home

Model	Best Hyperparameters of Each Algorithm		
	Fold 1	Fold 2	Fold 3
Decision Tree	max depth: 7 max features: sqrt min samples leave: 1 min samples split: 15	max depth: 7 max features: sqrt min samples leave: 1 min samples split: 2	max depth: 9 max features: sqrt min samples leave: 1 min samples split: 2
Random Forest	max depth: 7 min samples leave: 1 min samples split: 2 n estimators: 100	max depth: 7 min samples leave: 1 min samples split: 2 n estimators: 100	max depth: 7 min samples leave: 1 min samples split: 2 n estimators: 100
XGBoost	colsample bytree: 0.8 learning rate: 0.1 max depth: 5 n estimators: 300 subsample: 0.1	colsample bytree: 0.8 learning rate: 0.1 max depth: 3 n estimators: 200 subsample: 0.8	colsample bytree: 0.8 learning rate: 0.1 max depth: 3 n estimators: 100 subsample: 0.8
SVM	C: 10 kernel: linear	C: 10 kernel: linear	C: 1 kernel: linear
KNN	n neighbors: 3 p: 1 weights: distance	n neighbors: 3 p: 1 weights: distance	n neighbors: 3 p: 1 weights: distance
Logistic Regression	C: 10 Penalty: 12	C: 10 Penalty: 12	C: 10 Penalty: 12

Table 13: Best Hyperparameters of Each Algorithm

3. Satisfaction on Work From Home Policy
4. Supervisor Support to Get Job Done
5. Comfort with Organisation Culture
6. Comfort with Team Culture
7. Satisfaction on Facility at Office
8. Teammate Gives Warm Welcome and give Advice
9. Onboarding Program is Effective
10. Satisfaction on Recruitment Process

Figure 4.9 also shows the directionality impact of these important features.

Figure 4.8: Most impactful features

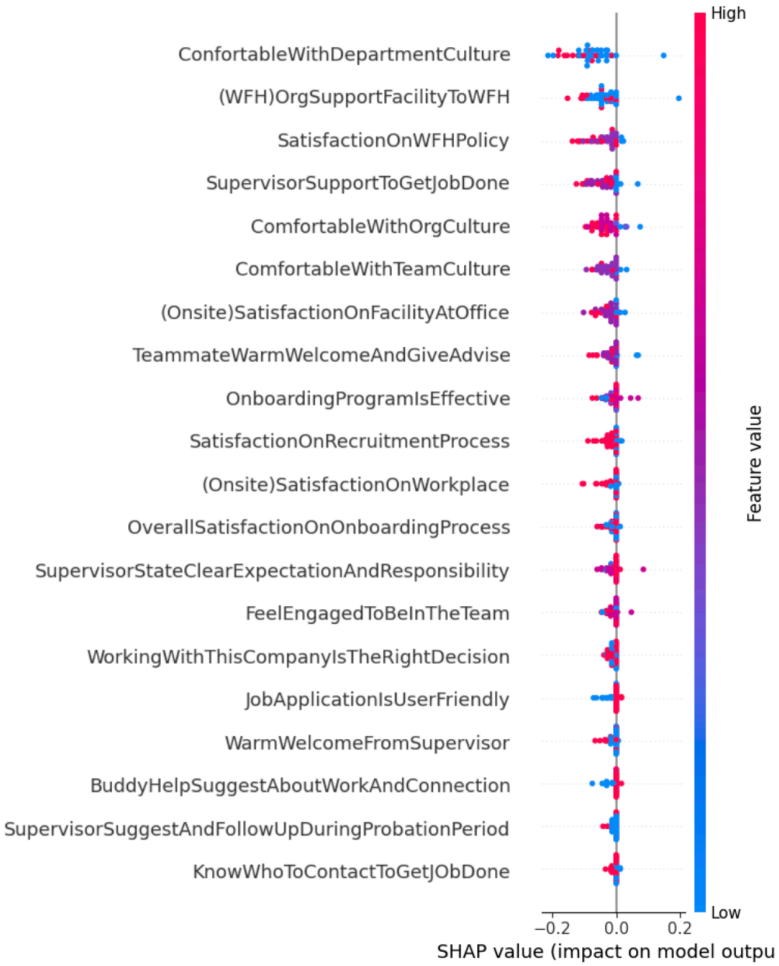

Figure 4.9: Directionality Impact of the Features

Chapter V

CONCLUSION

In this research, we explored the application of Machine Learning algorithms to predict newcomer turnover within organization. Newcomer turnover refers to the phenomenon of newly hired employees leaving the organization within a relatively short period after their hiring. By focusing specifically on the phenomenon of newcomer attrition within organizations, distinct from regular employee turnover, this research gained valuable insights into the factors that contribute to the departure of newly hired employees, shedding light on their unique challenges and needs. The uniqueness of our research lies in its exclusive focus on predicting newcomer turnover, in contrast to typical turnover studies that combine all employee departures.

5.1 Summary of Findings

This study investigated the challenge of predicting newcomer turnover within an organizational setting. Employing preprocessing techniques such as Stratified K-Fold and data upsampling, we effectively addressed issues related to imbalanced and small dataset. Our analysis involved several Machine Learning models, including Decision Tree, Random Forest, XGBoost, SVM, Naive Bayes, KNN and Logistic Regression. F1 score was chosen as an evaluation metric because in Human Resources, one of the primary objectives is to accurately pinpoint potential leavers while also minimize false alarms. The results indicated that the Random Forest model outperformed the others, with Recall of 0.857 and F1 score of 0.632.

Furthermore, we identified the most impactful features in the prediction process using SHAP Value. Comfort with the Department emerged as the top feature, suggesting its critical role in influencing newcomer turnover. Additionally, factors such as Organizational and Team Culture ranked among the top six, emphasizing the significance of

workplace culture for newcomers. The prominence of Work From Home Policy as the second and third most impactful feature indicated the influence of Work From Home policies on newcomer retention. Moreover, the importance of onboarding programs and satisfaction with the recruitment process among the top ten features highlighted the need for specific strategies to engage and retain new employees. These important features have negative directionality impact to the model, implying that the lower they are, the more likely newcomers will leave the company. These findings offer valuable insights for HR professionals to focus on specific areas, ensuring a positive and engaging experience for newcomers, distinct from the strategy for regular employees.

The model's inability to predict all departures could be attributed to the broader turnover context, where external factors such as other job opportunities or family issues significantly influence decisions. In the competitive landscape of the Thai financial industry, the decision to stay or leave is influenced not only by push factors but also by pull factors, namely attractive job opportunities from other financial firms. This complexity adds an extra layer of challenge to predicting departures.

5.2 Limitations

An important limitation in this study is the small dataset, consisting of only 132 examples. This could lead to potentially inflated claims about how well the model works and raises concerns about overfitting. To address these limitations, future research should focus on expanding the dataset to improve the reliability and relevance of the predictive models. Moreover, our data came from a Thai financial firm in Bangkok, introducing a specific cultural and organizational context. This may impact the generalizability of the findings to other locations or industries. Additionally, it's essential to recognize the limitation of focusing solely on one company's data. Therefore, including data from multiple sources or industries could provide a more comprehensive perspective in future research.

REFERENCES

[1] Krishna Kumar Mohbey. Employee's attrition prediction using machine learning approaches. pages 121–128, 2020.

[2] V Vijaya Saradhi and Girish Keshav Palshikar. Employee churn prediction. Expert Systems with Applications, 38(3):1999–2006, 2011.

[3] Alex Frye, Christopher Boomhower, Michael Smith, Lindsay Vitovsky, and Stacey Fabricant. Employee attrition: what makes an employee quit? SMU Data Science Review, 1(1):9, 2018.

[4] UD Labor. Us bureau of labor statistics. Retrieved from, 2020.

[5] Johnna Capitano, Vipanchi Mishra, Priyatharsini Selvarathinam, Amy Collins, and Andrew Crossett. How long are newcomers new in different occupations? Organization Management Journal, 19(3):110–123, 2022.

[6] Andry Alamsyah and Nisrina Salma. A comparative study of employee churn prediction model. In 2018 4th International Conference on Science and Technology (ICST), pages 1–4. IEEE, 2018.

[7] Shikha N Khera and Divya. Predictive modelling of employee turnover in indian it industry using machine learning techniques. Vision, 23(1):12–21, 2018.

[8] Fatemeh Mozaffari, Marzieh Rahimi, Hamidreza Yazdani, and Babak Sohrabi. Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking: An International Journal, (ahead-of-print), 2022.

[9] Jariya Limjeerajarat and Damrongsak Naparat. Preserving talent: Employee churn prediction in higher education. 2022.

[10] Lok Sundar Ganthi, Yaswanthi Nallapaneni, Deepalakshmi Perumalsamy, and Krishnakumar Mahalingam. Employee attrition prediction using machine learning algorithms. In Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 1, pages 577–596. Springer, 2022.

[11] Peter W Hom, Thomas W Lee, Jason D Shaw, and John P Hausknecht. One hundred years of employee turnover theory and research. Journal of applied psychology, 102 (3):530, 2017.

[12] R Hoppock. Job satisfaction. harper and brothers publishers. New York, 1935.

[13] Terence R Mitchell, Brooks C Holtom, Thomas W Lee, Chris J Sablynski, and Miriam Erez. Why people stay: Using job embeddedness to predict voluntary turnover. Academy of management journal, 44(6):1102–1121, 2001.

[14] Jan E Stets and Peter J Burke. Identity theory and social identity theory. Social psychology quarterly, pages 224–237, 2000.

[15] Peter W Hom, Terence R Mitchell, Thomas W Lee, and Rodger W Griffeth. Reviewing employee turnover: focusing on proximal withdrawal states and an expanded criterion. Psychological bulletin, 138(5):831, 2012.

[16] Sangita Ulhas Gorde. A study of employee retention. Journal of Emerging Technologies and Innovative Research (J ETIR), 6(6):331–337, 2019.

[17] Abreham Hunde. Factors influencing employee turnover and its effect on organizational performance: The case of harar beer factory, oromia regional states. 2019.

[18] K Balaji Mathimaran and Ananda A Kumar. Employee retention strategies–an empirical research. Global Journal of Management and Business Research: E Marketing, 17(1):17–22, 2017.

[19] Lesley Clack. Employee engagement: Keys to organizational success. The Palgrave Handbook of Workplace Well-Being, pages 1001–1028, 2021.

[20] Christopher S Reina, Kristie M Rogers, Suzanne J Peterson, Kris Byron, and Peter W Hom. Quitting the boss? the role of manager influence tactics and employee emotional engagement in voluntary turnover. Journal of leadership & organizational studies, 25(1):5–18, 2018.

[21] Anthony J Nyberg, Jason D Shaw, and Jing Zhu. The people still make the (remote work-) place: lessons from a pandemic, 2021.

[22] Denis Ushakov and Khodor Shatila. The impact of workplace culture on employee retention: An empirical study from lebanon. The Journal of Asian Finance, Economics and Business (JAFEB), 8:541–551, 2021.

[23] SHRM. Shrm customized human capital benchmarking report. 2017.

[24] Deloitte. Talent 2020: Surveying the talent paradox from the employee perspective. 2020.

[25] Bidisha Lahkar Das and Mukulesh Baruah. Employee retention: A review of literature. Journal of business and management, 14(2):8–16, 2013.

[26] Madeline Laurano. The true cost of a bad hire, August 2015.

[27] Pankaj Ajit. Prediction of employee turnover in organizations using machine learning algorithms. algorithms, 4(5):C5, 2016.

[28] I Setiawan, S Suprihanto, AC Nugraha, and J Hutahaean. Hr analytics: Employee attrition analysis using logistic regression. In IOP Conference Series: Materials Science and Engineering, volume 830, page 032001. IOP Publishing, 2020.

[29] PM Usha and NV Balaji. A comparative study on machine learning algorithms for employee attrition prediction. In IOP Conference Series: Materials Science and Engineering, volume 1085, page 012029. IOP Publishing, 2021.

[30] Stephen Taylor, Nesreen El-Rayes, and Michael Smith. An explicative and predictive study of employee attrition using tree-based models. 2020.

[31] Sandeep Yadav, Aman Jain, and Deepti Singh. Early prediction of employee attrition using data mining techniques. In 2018 IEEE 8th international advance computing conference (IACC), pages 349–354. IEEE, 2018.

[32] Sarah S Alduayj and Kashif Rajpoot. Predicting employee attrition using machine learning. In 2018 international conference on innovations in information technology (iit), pages 93–98. IEEE, 2018.

[33] Madara Pratt, Mohcine Boudhane, and Sarma Cakula. Employee attrition estimation using random forest algorithm. Baltic Journal of Modern Computing, 9(1):49–66, 2021.

[34] Guillaume Vergnolle and Nadia Lahrichi. Data-driven analysis of employee churn in the home care industry. Home Health Care Management & Practice, 35(2):75–85, 2023.

[35] Sheetal S Patil, SH Patil, Avinash M Pawar, Piyush Kumar Pandey, Swastik Sharma, and MS Bewoor. Employee churn walkthrough using knn. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), pages 1–4. IEEE, 2022.