

AN AI BASED SOLUTION TO ASSESS EXECUTIVE EMPLOYEE TURNOVER RISK IN THE APPAREL INDUSTRY

**R. D. S. Lakmal
2024**



An AI based Solution to Assess Executive Employee Turnover Risk in the Apparel Industry

**A Thesis Submitted for the Degree of Master of
Business Analytics**

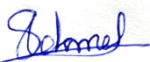


**R. D. S. Lakmal
University of Colombo School of Computing
2024**

DECLARATION


Name of the student: R. D. S. Lakmal
Registration number: 2020/BA/021
Name of the Degree Programme: Master of Business Analytics
Project/Thesis title: An AI based Solution to Assess Executive Employee Turnover Risk in the Apparel Industry

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	Date (DD/MM/YYYY)
	28/09/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	Supervisor 2	Supervisor 3
Name	Prof. M. G. N. A. S. Fernando		
Signature			
Date	28/09/2024		

I would like to dedicate this thesis to my parents and family who have continuously supported me to come to the stage where I am today.

ACKNOWLEDGEMENTS

I sincerely thank my supervisor Prof. M. G. N. A. S. Fernando, for helping me with suggestions, ideas, guidance and also the time and effort from the supervisor throughout the project. I also thank my family, friends, work colleagues and everyone who have supported me to make this effort a success.

ABSTRACT

Employee turnover, particularly among executive and above cadres, presents a significant challenge in the Sri Lankan Apparel industry, impacting organizational stability and productivity. This thesis aims to address this issue by predicting the likelihood of employees leaving their current positions within the next year and estimating the probable time frame for such turnovers.

Initially, a descriptive analysis of executive and above employee behavior within a major apparel manufacturing company revealed a concerning trend of increasing turnover among long-term and skilled employees over the past five years. Factors such as prolonged tenure within a single grade correlated positively with turnover rates, while employees with diverse job roles exhibited lower turnover tendencies. These findings underscored the urgent need for proactive measures to mitigate turnover risks. Transitioning to predictive modeling, we formulated turnover prediction as both a binary classification problem to ascertain turnover possibility and a multi-classification problem to predict turnover horizons. Leveraging supervised machine learning techniques and publicly available employee data from LinkedIn, we trained models to forecast turnover events. The XGB Classifier emerged as the most effective algorithm, achieving accuracies of 81% for turnover possibility prediction and 75% for turnover horizon estimation. Key features influencing turnover likelihood included the frequency of internal promotions, tenure, job durations, and educational qualifications. These insights emphasize the importance of continuous monitoring of such variables to preempt turnover events effectively. Furthermore, we developed a user-friendly interface to facilitate easy access to turnover risk scores and timelines based on employee LinkedIn profiles.

In considering future research directions, we propose integrating internal data sources to enhance model accuracy and exploring additional variables such as salary and employee feedback. Moreover, the analysis can be extended to encompass internal turnovers and industry-specific turnover patterns, offering tailored insights for diverse organizational contexts. Additionally, incorporating social network analysis and exploring turnover prediction from a company-wide perspective present promising avenues for further investigation. Overall, this study provides valuable insights and tools to proactively manage employee turnover in the apparel industry and beyond.

LIST OF PUBLICATIONS

TABLE OF CONTENTS

DECLARATION.....	I
ACKNOWLEDGEMENTS.....	III
ABSTRACT.....	IV
LIST OF PUBLICATIONS.....	V
TABLE OF CONTENTS.....	VI
LIST OF FIGURES	VII
LIST OF TABLES	VIII
CHAPTER 1: INTRODUCTION.....	1
1.1 PROBLEM STATEMENT AND MOTIVATION.....	2
1.2 PROJECT OBJECTIVES	2
1.3 SCOPE OF THE STUDY.....	3
1.4 STRUCTURE OF THE STUDY	3
CHAPTER 2: LITERATURE REVIEW.....	4
2.1 A LITERATURE REVIEW	4
2.2 INFLUENTIAL FACTORS OF EMPLOYEE TURNOVER.....	5
2.3 RELATED STUDIES AND FINDINGS ON EMPLOYEE TURNOVER PREDICTION	5
2.4 DATA COLLECTION APPROACHES IN RELATED STUDIES	6
2.5 DATA LIMITATION AND GAPS IN RELATED STUDIES	7
CHAPTER 3: METHODOLOGY.....	8
3.1 SCRAPING RAW DATA.....	8
3.2 EXTRACTING, PRE-PROCESSING AND ENCODING JOB EXPERIENCE DATA	10
3.3 EXTRACTING, PRE-PROCESSING EDUCATION EXPERIENCE DATA	12
3.4 EXTRACTING, PRE-PROCESSING EMPLOYEE PROFILE DATA	14
3.5 COMBINING DATA AND GETTING THE FINAL DATASET	15
3.6 DEVELOPING AI MODEL FOR TURNOVER POSSIBILITY PREDICTION	17
3.7 DEVELOPING AI MODEL FOR CHURN HORIZON PREDICTION	21
3.8 DEVELOPING UI.....	24
3.9 METHODOLOGY SUMMARY	25
CHAPTER 4: EVALUATION AND RESULTS.....	26
4.1 TURNOVER BEHAVIOR OF THE EXECUTIVE AND ABOVE EMPLOYEES:	26
4.2 EVALUATION OF THE PREDICTION MODELS:	31
CHAPTER 5: CONCLUSION AND FUTURE WORK	34
5.1 CONCLUSION:	34
5.2 FUTURE WORK:	36
APPENDICES	I
REFERENCES.....	III

LIST OF FIGURES

Figure 1: 3rd party scraping API dashboard.....	8
Figure 2: Raw data sample for employee data extracted through API.....	9
Figure 3: Raw data sample for company data extracted through API.....	9
Figure 4: Raw data sample for education institute data extracted through API.....	9
Figure 5: Job experience data sample extracted from experience Json file.....	10
Figure 6: Extracted Job experience data after filtering.....	10
Figure 7: Job experience data after categorizing the job level	11
Figure 8: Company profile data after merging with the job experience data and removing unwanted columns	11
Figure 9: Encoded job experience data.....	11
Figure 10: Job experience data sample extracted from experience Json file.....	12
Figure 11: Education experience data sample extracted from experience Json file	12
Figure 12: Education experience data after categorizing the course level	12
Figure 13: Education institute profile data after merging with the education experience data and removing unwanted columns	13
Figure 14: Encoded education experience data	13
Figure 15: Encoded education experience data	13
Figure 16: Employee profile data	14
Figure 17: Encoded employee profile data.....	14
Figure 18: PCA Chart for the Dimensionality Reduction	17
Figure 19: Correlation Matrix of the selected variables	17
Figure 20: Turnover Probability Prediction Model Accuracy Chart.....	19
Figure 21: Turnover Probability Prediction Model Precission Chart.....	19
Figure 22: Turnover Probability Prediction Model Recall Chart.....	20
Figure 23: Turnover Horizon Prediction Model Accuracy Chart.....	22
Figure 24: Turnover Horizon Prediction Model Precission Chart	22
Figure 25: Turnover Horizon Prediction Model Recall Chart.....	23
Figure 26: Interface to enter the LinkedIn profile name	24
Figure 27: Interface where a LinkedIn profile name entered and run	24
Figure 28: Output interface.....	24
Figure 29: Methodology Summary.....	25
Figure 30: Employee turnover count and percentage out of total	26
Figure 31: Employee turnover grade wise.....	27
Figure 32: Employee turnover grade wise trend	27
Figure 33: Employee turnover number of years in the company wise.....	28
Figure 34: Employee turnover number of years in the company wise trend.....	28
Figure 35: Employee turnover number of jobs held wise	29
Figure 36: Employee turnover number of jobs held wise trend	29
Figure 37: Feature correlation matrix.....	30
Figure 38: Model 1 confusion matrix	31
Figure 39: Model 2 confusion matrix	31
Figure 40: Turnover Probability Prediction Model Feature Importance.....	32
Figure 41: Turnover Horizon Prediction Model Feature importance.....	33

LIST OF TABLES

Table 1: Job data encoding logics.....	11
Table 2: Job data encoding logics.....	13
Table 3: Feature List.....	16
Table 4: Model performance	18

CHAPTER 1:

INTRODUCTION

Employee turnover which is a regular challenge every company is facing, has become worse in Sri Lanka during the last few months with the current economic situation of the country. When we look at the apparel industry specifically, it has become very critical due to the opportunities available within Sri Lanka and other South Asian countries with regards to apparel sector, which will cause organizations to lose their highly skilled employees, resulting a considerable impact to the smooth running of operation in the organization.

Due to the importance of this, employee withdrawal, in the form of turnover, has sustained the interest of personnel researchers, behavioral scientists, and management practitioners during the last many decades [1]. When an unexpected turnover request is raised in a company, significant effort is required to search for a replacement and there is a risk that the operation of the company will be disrupted if a suitable replacement is not found. [2]. Organizations invest a lot on their employees in terms of induction, training, developing, maintaining and retaining them in their organization. Therefore, managers at all costs must minimize employee's turnover [3].

Since this has become critical topic that has been studied over several decades by a number of researches, there are many studies done in order to minimize the employee turnover or to identify the turnover risk beforehand. In most of the studies, researchers have used internal HR data of companies through questionnaires or company database itself.

But the limitations of this approach are that the data in the human resources management systems are highly private and can hardly be shared among different companies. Also, the time span of the data is more relevant to the employee's stay at the current working place and the overall work life journey of the employee might not be available. Also, with questionnaire approach, responses from the individuals can be bias on questions like job satisfaction, manager support. Also, the willingness to answer or giving the correct answer can result in accuracy issues in the dataset. This has made it difficult to conduct long-term career evaluations of employees, so most of the research is based on employee personal factors and organizational factors [4].

Due to these limitation in manual or company data extraction, here we have looked at an approach is to obtain publicly available data with regards to the individuals. Here our main target data source would be LinkedIn. LinkedIn provides a database of dynamic and massive scale resumes. The application of data mining techniques can help in the analysis and classification of professional profiles [5].

With this approach we plan to observe an unbiased and uniform behavior of data with regards to the target population and come up with a solution or a tool which will ultimately provide the user with the turnover risk of an employee and also the time span the employee is probable to stay in the organization. This can be used for ongoing HR practices and also can be used as a risk assessment when recruiting new employees and for headhunters or requirement agents to target skilled employees that are probable to move.

1.1 Problem Statement and Motivation

As per the literature and initial descriptive analysis, it is evident that employee turnover has become worse in the Sri Lankan apparel industry during recent years due to many socio-economic factors[6]. This has caused a considerable impact to the smooth running of operation in the organization.

Although a few studies have been done on this problem, they have focused mainly on factory level employees which are easily replaceable and have looked only at HR data from companies which limits the visibility of the entire career span of the employees or questionnaires which is very probable to be biased and imbalance.

Based on the previous studies, there are no significant studies that has focused on executive and above professional employees who are very skilled and difficult to replace and also not looked at the publicly available data of employees to cover the entire career journey of the employee with continuous data access. Our objective is to address this gap in the research domain.

1.2 Project Objectives

- To analyze and identify the turnover behavior of the executive and above employees in Sri Lankan apparel sector (considering one of the top 3 companies which covers more than 50% of the population).

- To identify the main factors which would cause an employee to change a job with the publicly available data of the employee.
- To provide a tool to companies and 3rd parties to identify the turnover chance of employees in advance and take necessary actions using AI Techniques.

1.3 Scope of the Study

Our study is mainly focused on the executive and above employee cadre in Sri Lankan apparel sector and their publicly available data on social network platforms, mainly LinkedIn. With regards to the apparel companies in Sri Lanka, we will be considering the top 4 companies which covers 90% of the population. The number of profiles available for these 4 companies in LinkedIn is more than 17,000.

We have presumed 90% of the profiles are up to date and this will be validated after the initial data extraction. Only complete profiles were considered for the study and this will be one limitation on the study. But when compared with the amount of data available in LinkedIn, it will be an inconsiderable proportion relative to the population size.

1.4 Structure of the Study

We will be using LinkedIn as the main data source and will be using 3rd party available APIs to extract data. After cleansing the data, we will develop multiple data mining, pattern recognition or machine learning model/s or algorithms and then select the most accurate, reliable and most interpretable model or algorithm after analyzing the results.

As the final deliverable, we will develop a solution or GUI for the end users where the user can input the required LinkedIn profile/s and get the 2 main outcomes which are employee turnover risk score and probable turnover timeline.

CHAPTER 2: LITERATURE REVIEW

2.1 A Literature Review

Employee turnover is the rotation of workers around the labor market; between firms, jobs and occupations; and between the states of employment and unemployment [1]. It is one of the most significant problems an organization can encounter throughout its lifecycle, as it is difficult to predict and often introduces noticeable voids in an organization's skilled workforce [2].

Employee turnover can actually be subdivided in 3 buckets: involuntary turnover (induced by the company), voluntary turnover (employee resignation) and retirements [3]. These may include job changes within a single employer and leaving one firm to take a job in another firm. In either case, there is usually the intention to grow and increase in skills, responsibility, and remuneration, and/or improve the "fit" between employee skills and desires and job requirements. [4]

When employees leave an organization, they carry with them invaluable tacit knowledge which is often the source of competitive advantage for the business. In order for an organization to continually have a higher competitive advantage over its competition, it should make it a duty to minimize employee attrition [4]. When an unexpected turnover request is raised in a company, significant effort is required to search for a replacement and there is a risk that the operation of the company will be disrupted if a suitable replacement is not found. [5].

Organizations invest a lot on their employees in terms of induction, training, developing, maintaining and retaining them in their organization. Therefore, managers at all costs must minimize employee's turnover [1]. Turnover causes many different types of costs for organizations. Costs of turnover are divided to direct costs such as advertising the position, replacement, recruitment and selection, temporary staff and management time, and indirect costs are morale related costs, pressure on remaining staff, costs of learning, product/service quality and organizational memory. It has been proved that 15-30 percent of turnover costs are direct and about 70-85 percent of turnover costs are hidden costs such as lost productivity and opportunity. [6]

2.2 Influential Factors of Employee Turnover

The desirability of movement can be characterized by job satisfaction, salary growth, promotions, and organization's commitment, while the ease of movement can be characterized by the job market availability, unemployment rate and personal skill levels etc. [5]. A high level of labor turnover can also be caused by many factors such as: inadequate wage levels leading to employees moving to competitors, poor morale and low levels of motivation within the workforce, recruiting and selecting the wrong employees in the first place, meaning they leave to seek more suitable employment, A buoyant local labor market offering more (and perhaps more attractive) opportunities to employee, poor organization and lack of development [4].

An early review of voluntary turnover studies has found that the strongest predictors for voluntary turnover were age, tenure, pay, overall job satisfaction, and employees' perception of fairness. But other studies have also stressed the importance of job performance, job characteristics (role, seniority in role...), enhanced individual demographic characteristics (age/experience, gender, ethnicity, education, marital status), structural characteristics (i.e., team size and performance) and geographical factors. Finally, some studies have also set an emphasis on salary, working conditions, job satisfaction, supervision, advancement, recognition, growth potential, burnout etc. [3].

2.3 Related Studies and Findings on Employee Turnover Prediction

The A comparative study involving accuracy and memory utilization of selected algorithms for predicting employee turnover was conducted by Rohit Punnoose et al. [7]. The authors collected the data from the in-formation system used by the human resource department of a retailer with global operations and data from the Bureau of Labor Statistics. Several classification algorithms were applied, namely, extreme gradient boosting (XGBoost), logistic regression, Naïve Bayesian, RF, linear support vector machine, linear discriminant analysis and KNN. The authors have found that XGBoost exhibits the best performance regarding the accuracy and memory utilization. Jain et al. [8] have carried out research to predict turnover rate using XGBoost. They have found that age, gender, marital status, years at the company, job satisfaction, and distance from home have the most significant effects on turnover among all attributes in the dataset.

Numerous algorithms; namely, logistic regression, gradient boosting classifier, support vector machine, and RF were applied to the IBM dataset prepared by IBM data scientists [9], [10]. After applying the RF classifier, fifteen features were found to be more significant in deciding whether employees quit their jobs or not. XGBoost was found to have the highest performance (with an AUC of 0.84596) among all the applied algorithms. Zhao et al. [2] have evaluated the performance of ten supervised machine learning algorithms; namely, RF, gradient boosting trees, XGBoost, support vector machines, decision tree, neural networks, linear discriminant analysis, Naïve Bayesian, logistic regression, support vector machines and KNN on numerous HR datasets. The authors have found that XGBoost is the most reliable algorithm among all the applied algorithms. Zhang et al. [11] have attempted to find out the most important factors that lead to employee turnover. The authors have found an essential correlation between department and work. Also, they have found that the gender of employees significantly affects turnover. A logistic regression algorithm was applied for predicting the turnover with an accuracy of 87.2%.

Sisodia et al. [12] have carried out an investigation to find out the reasons causing the employee turnover by building models using machine learning algorithms to forecast employee turnover. They have found that the main reasons causing high employee turnover rates are time spent with the company, workload, and promotion. The used machine learning algorithms for building the models were decision tree, support vector machine, Naïve Bayesian, KNN, and RF. The accuracy, precision, F-score, recall, specificity, and FPR of the models were compared. In terms of accuracy, F-score, and precision, RF performed better and in terms of recall, the decision tree was better.

2.4 Data Collection Approaches in Related Studies

In most of the studied carried out with regards to the topic, researchers have collected data of companies through questionnaires or company HR systems [5],[3],[4],[6],[13]. Human resources data is highly confidential and rarely shared between companies. As a result, some conventional statistical methods relied on employee data collected over a short period of time within an organization. As a result, conducting long-term career assessments of employees has become difficult, and the majority studies are based on employee self and organizational factors [14].

As a solution for this, few of the researches in the recent years have looked at obtaining data from social networks such as LinkedIn, Maimai, and Viadeo are examples of websites with extensive social information about employees and a longer time frame [14]–[16]. Most of the studies which have looked in to social network data are based on China and they have mainly focused on a much more common and popular platforms in China which is similar to LinkedIn [14]–[16].

LinkedIn is the largest professional social network in the world, and currently has 300 million plus users in more than 200 countries and territories[15]. For the social network related data sets, common variables which were looked at are city, state, current job position, number of undergraduate and graduate courses, time spent in undergraduate and graduate education, time in the current job position, time in the previous job position, number of languages the person speaks, number of declared skills, number of connections to other persons, number of organizations in which the person worked.[15]

2.5 Data Limitation and Gaps in Related Studies

It can be identified that almost all researches that have been conducted so far, uses classification methods for turnover prediction. Mere identification of churners from no churners is not sufficient to tackle the employee turnover problem [17]. For this study we have chosen apparel sector employees since it plays significant role in the economic development of the country. Thus, currently the industry faces a high employee turnover continuously and this problem has been seen over past few years.[13]

CHAPTER 3: METHODOLOGY

As the initial data source, we have selected LinkedIn as our primary data source due to its comprehensiveness and completeness in terms of public employee data. We have presumed that over 90% of the profile details of the professionals are accurately and comprehensively updated.

3.1 Scraping Raw Data

As per the previous studies and also based on the initial research on data extraction from LinkedIn, we are using a 3rd party paid API called “ScrapeOps” to extract data from Linked profiles which are filtered as Sri Lankan, in apparel industry and executive and above level. Figure 1 shows the API dashboard from the 3rd party solution website.

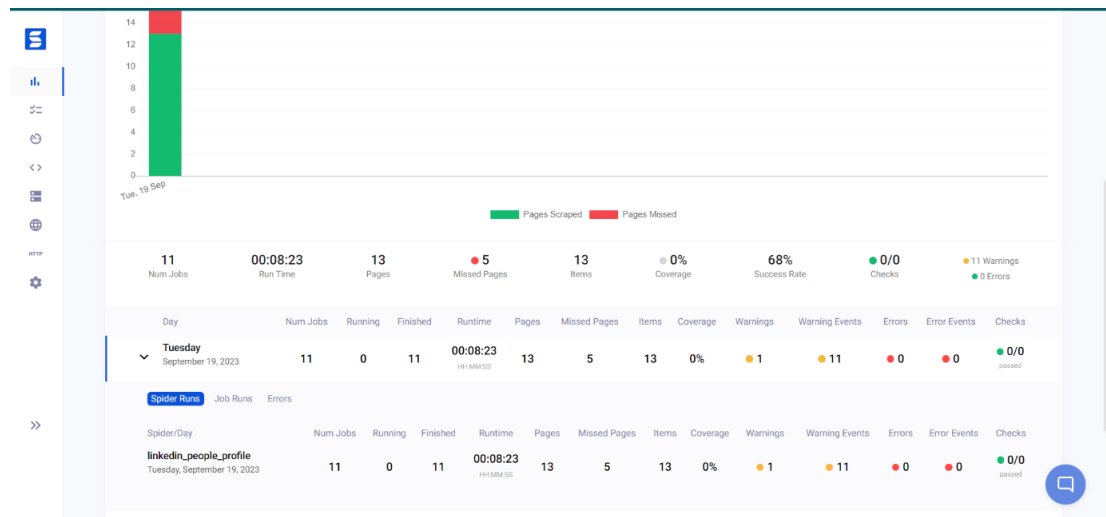


Figure 1: 3rd party scraping API dashboard

Similar to the employee profiles, we had to extract company profiles and educational institute profiles in order to create the required variables below figures show samples of the raw datasets scraped for each using the API.

profile	reidhoffman
url	https://www.linkedin.com/in/reidhoffman/
name	Reid Hoffman
description	Entrepreneur. Product and Business Strategist. Investor. Podcaster.
location	United States
followers	
connections	
about	All aspects of consumer internet and software. Focus is on product development, innovation, busi
experience	[{'organisation_profile': 'https://www.linkedin.com/company/greylock-partners', 'location': 'Menl
education	[{'organisation': '', 'organisation_profile': 'https://www.linkedin.com/school/university-of-oulu/', 'course_details
profile	sagara-lakmal-3b5634103
url	https://www.linkedin.com/in/sagara-lakmal-3b5634103/
name	Sagara Lakmal
description	Manager - Data Analytics & Governance
location	Sri Lanka
followers	
connections	
about	
experience	[{'organisation_profile': 'https://cz.linkedin.com/company/mas-holdings', 'location': '', 'descriptor
education	[{'organisation': '', 'organisation_profile': 'https://cz.linkedin.com/school/ucsc-lk/', 'course_details

Figure 2: Raw data sample for employee data extracted through API

name	MAS Holdings	
summary	Change Is Courage	
industry	Apparel & Fashion	
size	10,001+ employees	
founded	1987	
name	Dialog Axiata PLC	
summary	Dialog Axiata is Sri Lanka's premier connectivity provider.	
industry	Telecommunications	
size	1,001-5,000 employees	
founded	1993	

Figure 3: Raw data sample for company data extracted through API

Name	University of Kelaniya Sri Lanka
Founded	1875
Summary	The University of Kelaniya is committed to provide high quality education and to conduct
Industry	Higher Education
size	1,001-5,000 employees
Name	University of Colombo School of Computing
Founded	2002
Summary	Pioneer of ICT education in Sri Lanka
Industry	Higher Education
size	51-200 employees

Figure 4: Raw data sample for education institute data extracted through API

3.2 Extracting, Pre-Processing and Encoding Job Experience Data

Here the next step is to extract the job experience data from the experience Json file available in employee dataset. For this we have used the available python libraries for the data splitting and extraction.

Profile	organisation_profile'	location	Designation	start_time	end_time
sagara-lakmal-3b5634104	https://linkedin.com/company/mas-holdings	Colombo, Western, Sri Lanka	Manager - Data Analytics & Governance	Jan-23	present
sagara-lakmal-3b5634105	https://linkedin.com/company/ndbbank	Colombo, Western, Sri Lanka	Associate Manager - Data Modelling	Jun-21	Jan-23
sagara-lakmal-3b5634106	https://linkedin.com/company/dialog-axiata-plc	Colombo, Western, Sri Lanka	Senior Executive - Data Visualization	Oct-19	Jun-21
sagara-lakmal-3b5634107	https://linkedin.com/company/mas-holdings	Colombo, Western, Sri Lanka	Executive - Data Analytics	Apr-18	Sep-19

Figure 5: Job experience data sample extracted from experience Json file

In this analysis we are assuming we are now at the start of the January 2023 and predict the turnover during the year 2023. Therefore, we have dropped the records where start date is in 2023 and changed the end date to 1st of January 2023 for the end dates which are in 2023.

Profile	organisation_profile'	location	Designation	start_time	end_time
sagara-lakmal-3b5634105	https://linkedin.com/company/ndbbank	Colombo, Western, Sri Lanka	Associate Manager - Data Modelling	Jun-21	Jan-23
sagara-lakmal-3b5634106	https://linkedin.com/company/dialog-axiata-plc	Colombo, Western, Sri Lanka	Senior Executive - Data Visualization	Oct-19	Jun-21
sagara-lakmal-3b5634107	https://linkedin.com/company/mas-holdings	Colombo, Western, Sri Lanka	Executive - Data Analytics	Apr-18	Sep-19

Figure 6: Extracted Job experience data after filtering

Next step is to identify the job level from the designation. We are trying to categorize the available designation texts to 10 level as below.

1. Intern
2. Junior Executive
3. Executive
4. Senior Executive
5. Assistant Manager
6. Manager
7. Deputy General Manager
8. General Manager
9. Director
10. Chief Officer

To do this we are using NLP text matching techniques to map the keywords to the job level dictionary that we have developed.

sagara-lakmal-3b5634104	https://linkedin.com/company/ndbbank	Jun-21	Jan-23	5
sagara-lakmal-3b5634105	https://linkedin.com/company/dialog-axiata-plc	Oct-19	Jun-21	4
sagara-lakmal-3b5634106	https://linkedin.com/company/mas-holdings	Apr-18	Sep-19	3

Figure 7: Job experience data after categorizing the job level

Next task is to merge the company profile data to the job experience data and remove the unwanted columns. Here we have used a simple outer join to do the task.

Profile	organisation_profile	Start Time	End Time	Job Level	industry	size	founded	Headquarters
sagara-lakmal-3b5634104	https://linkedin.com/company/ndbbank	Jun-21	Jan-23	5	Financial Services	1,001-5,000 employees	1993	Colombo, Western, Sri Lanka
sagara-lakmal-3b5634105	https://linkedin.com/company/dialog-axiata-plc	Oct-19	Jun-21	4	Telecommunication	1,001-5,000 employees	1993	Colombo, Western, Sri Lanka
sagara-lakmal-3b5634106	https://linkedin.com/company/mas-holdings	Apr-18	Sep-19	3	Apparel & Fashion	10,001+ employees	1987	Colombo, Western, Sri Lanka

Figure 8: Company profile data after merging with the job experience data and removing unwanted columns

Next task is to encode the non-numeric data into the numeric format and creating numeric variables. For that we have used below logics.

New Column	Logic	Output
start_recency_months	today-start date	n months
end_recency_months	today-end date	n months
duration	start_recency_months- end_recency_months	
apparel_industry	if Apparel & Fashion then 1 else 0	binary
company_size	1,000 - employees	1
	1,001-5,000 employees	2
	5,001-10,000 employees	3
	10,001+ employees	4
company_age_years	2023-founded	n years
sri_lankan	if "Sri Lanka included" then 1 else 0	binary

Table 1: Job data encoding logics

Profile	organisation_profile	job_level	duration_months	start_recency_months	end_recency_months	apparel_industry	company_size	company_age_years	sri_lankan
sagara-lakmal-3b5634104	https://linkedin.com/c	5	19	19	0	0	2	44	1
sagara-lakmal-3b5634105	https://linkedin.com/c	4	20	39	19	0	2	30	1
sagara-lakmal-3b5634106	https://linkedin.com/c	3	17	57	40	1	4	36	1

Figure 9: Encoded job experience data

Next task is to derive the below 4 variables from the encoded dataset.

- Is this job a company change from the previous one (1/0)
- Cumulative number of distinct companies worked in
- Is the job a level up from the previous one (1/0)
- Is this job a lateral movement from the previous one (1/0)

Profile	job_level	duration_months	start_recency_months	end_recency_months	apparel_industry	company_size	company_age_years	sri_lankan	company_change	cum_no_of_companies	level_up	L_Move
sagara-lakmal-3b5634104	5	19	19	0	0	2	44	1	1	3	1	0
sagara-lakmal-3b5634105	4	20	39	19	0	2	30	1	0	2	1	0
sagara-lakmal-3b5634106	3	17	57	40	1	4	36	1	-1	1	-1	-1

Figure 10: Job experience data sample extracted from experience Json file

3.3 Extracting, Pre-Processing Education Experience Data

Next step is to extract the education experience data from the education Json file available in employee dataset. For this we have used the available python libraries for the data splitting and extraction.

Profile	organisation_profile	course_details	start_time	end_time
sagara-lakmal-3b5634103	https://linkedin.com/school/ucsc-lk/	Master of Business Analytics	2021	present
sagara-lakmal-3b5634104	https://linkedin.com/school/university-of-kelaniya-sri-lanka/	BSc in Management & Information Technology, sr	2013	2018
sagara-lakmal-3b5634105	https://linkedin.com/school/cima/	Part Qualified Accounting and Business/Managem	2012	2013
sagara-lakmal-3b5634106	https://linkedin.com/school/anandacollege/	GCE Advance Level Combined Mathematics	2003	2011

Figure 11: Education experience data sample extracted from experience Json file

Next step is to identify the course level from the course details. We are trying to categorize the educational qualifications texts to 6 level as below.

0. **Other** - courses which will not fall into any of the below categories
1. **High school** - ex: school, college, convent, vidyalaya
2. **Diploma**- ex: diploma, dip
3. **Bachelor** - ex: bachelor, bsc, bcom, beng
4. **Masters** - ex: masters, msc, mba
5. **Post Masters** - ex: phd, pfor, mphil

To do this we are using NLP techniques to tokenize and match the keywords and then match the categories and extract the data.

Profile	organisation_profile	course_details	course_level	start_time	end_time
sagara-lakmal-3b5634103	https://linkedin.com/school/ucsc-lk/	Master of Business Analytics	4	2021	present
sagara-lakmal-3b5634104	https://linkedin.com/school/university-of-kelaniya-sri-lanka/	BSc in Management & Information Techno	3	2013	2018
sagara-lakmal-3b5634105	https://linkedin.com/school/cima/	Part Qualified Accounting and Business/Ma	0	2012	2013
sagara-lakmal-3b5634106	https://linkedin.com/school/anandacollege/	GCE Advance Level Combined Mathematic	1	2003	2011

Figure 12: Education experience data after categorizing the course level

Next task is to map the education institute data to the education experience data and remove the unwanted columns. Here we have used a simple outer join to do the task.

Profile	course_level	start_time	end_time	Founded	size	Headquarters
sagara-lakmal-3b5634103	5	2021	present	2002	51-200 employees	Colombo, Western, Sri Lanka
sagara-lakmal-3b5634104	4	2013	2018	1875	1,001-5,000 employees	Colombo, Western, Sri Lanka
sagara-lakmal-3b5634105	1	2012	2013	1919	201-500 employees	One South Place, London
sagara-lakmal-3b5634106	2	2003	2011	1886	5,001-10,000 employees	Colombo, Western, Sri Lanka

Figure 13: Education institute profile data after merging with the education experience data and removing unwanted columns

Next task is to encode the non-numeric data into the numeric format and creating numeric variables. For that we have used below logics.

New Column	Logic	Output
duration	end_time – start_time	n years
start_recency_years	today-start time	n years
end_recency_years	today-end time	n years
school_size	1,000- employees	1
	1,001-5,000 employees	2
	5,001-10,000 employees	3
	10,001+ employees	4
school_age_years	2023-founded	n years
sri_lankan	if “Sri Lanka” included then 1 else 0	binary

Table 2: Job data encoding logics

Profile	course_level	duration	start_recency_years	end_recency_years	school_size	school_age_years	sri_lankan
sagara-lakmal-3b5634103	2	3	2	0	1	21	1
sagara-lakmal-3b5634104	2	5	10	5	2	148	1
sagara-lakmal-3b5634105	1	1	11	10	1	104	0
sagara-lakmal-3b5634106	1	8	20	12	3	137	1

Figure 14: Encoded education experience data

Next task is to derive the below the variable “Cumulative number of distinct schools” from the encoded dataset.

Profile	course_level	duration	start_recency_years	end_recency_years	school_size	school_age_years	sri_lankan	cum_no_of_edu_institutes
sagara-lakmal-3b5634103	5	2	2	0	1	21	1	4
sagara-lakmal-3b5634104	4	5	10	5	2	148	1	3
sagara-lakmal-3b5634105	1	1	11	10	1	104	0	2
sagara-lakmal-3b5634106	2	8	20	12	3	137	1	1

Figure 15: Encoded education experience data

3.4 Extracting, Pre-Processing Employee Profile Data

Next step is to get the remaining profile data from the employee profile dataset.

profile	name	location
reidhoffman	Reid Hoffman	United States
sagara-lakmal-3b5634103	Sagara Lakmal	Sri Lanka

Figure 16: Employee profile data

Next step is to identify the gender from the name. To do this we are using NLP techniques to tokenize and match with different name text corpuses that are readily available to identify the gender. Then we get the output as a binary output to indicate male as 1 and female as 0. Then we use a simple if else logic to identify the current country as Sri Lanka or not in binary form. Finally, we remove the unnecessary columns and get the final dataset.

profile	male	sri_lanka
reidhoffman	1	0
sagara-lakmal-3b5634103	1	1

Figure 17: Encoded employee profile data

3.5 Combining data and getting the final dataset

Next step is to combine the encoded features into a formal tabular structure which is compatible with the machine learning, pattern recognition or data mining. Table 3 shows the processed feature set we are planning to use for the model developments.

No	Type	Feature	Example
1	Personal	profile	sagara-lakmal-3b5634105
2		male	1
3		sri_lankan	1
4	Job Data	number of jobs	4
5		first job level	1
6		first job recency months	67
7		first job duration months	10
8		first company size	4
9		first company age years	36
10		first company apparel	1
11		first company sri lankan	1
12		last job level	5
13		last job recency months	19
14		last job duration months	19
15		last company size	2
16		last company age years	44
17		last company apparel	0
18		last company sri lankan	1
19		minimum job duration months	10
20		maximum job duration months	20
21		average job duration months	16
22		total job duration months	66
23		number of turnovers	2
24		number of companies	3
25		number of levelups	3
26		total level ups	4
27		average years for levelup	1.30
28		total LUs within company	2
29		total LUs outside company	2
30		number of lateral movements	0
31		duration ratio in sri lanka	1
32		duration ratio in apparel	0.41
33	Education	number of qualifications	4
34		first qualification level	2
35		first qualification recency years	2
36		first qualification duration years	3
37		first institute size	3
38		first institute age years	137

39	first institute sri lankan	1
40	last qualification level	5
41	last qualification recency years	2
42	last qualification duration years	3
43	last institute size	1
44	last institute age years	21
45	last institute sri lankan	1
46	minimum qualification duration years	1
47	maximum qualification duration years	8
48	average qualification duration years	4.25
49	total qualification duration years	17
50	number of institutes	4
51	qualification level ups	4
52	duration ratio in sri lanka	0.94
53	number of qualifications	4
54	first qualification level	2
55	first qualification recency years	2

Table 3: Feature List

3.6 Developing AI Model for Turnover Possibility Prediction

Before starting the development, we have done a principal component analysis and also checked the correlation of the considered variables and have selected 15 uncorrelated variables out of 55 we had as a dimension reduction step.

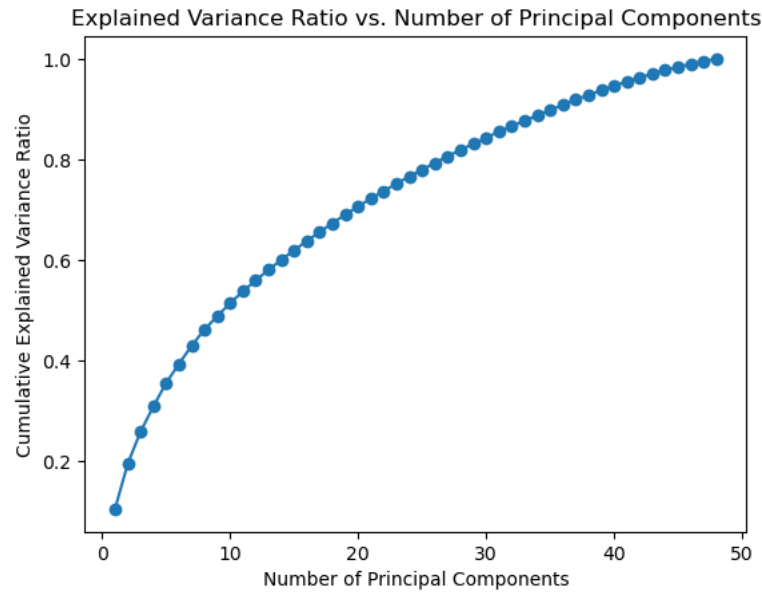


Figure 18: PCA Chart for the Dimensionality Reduction

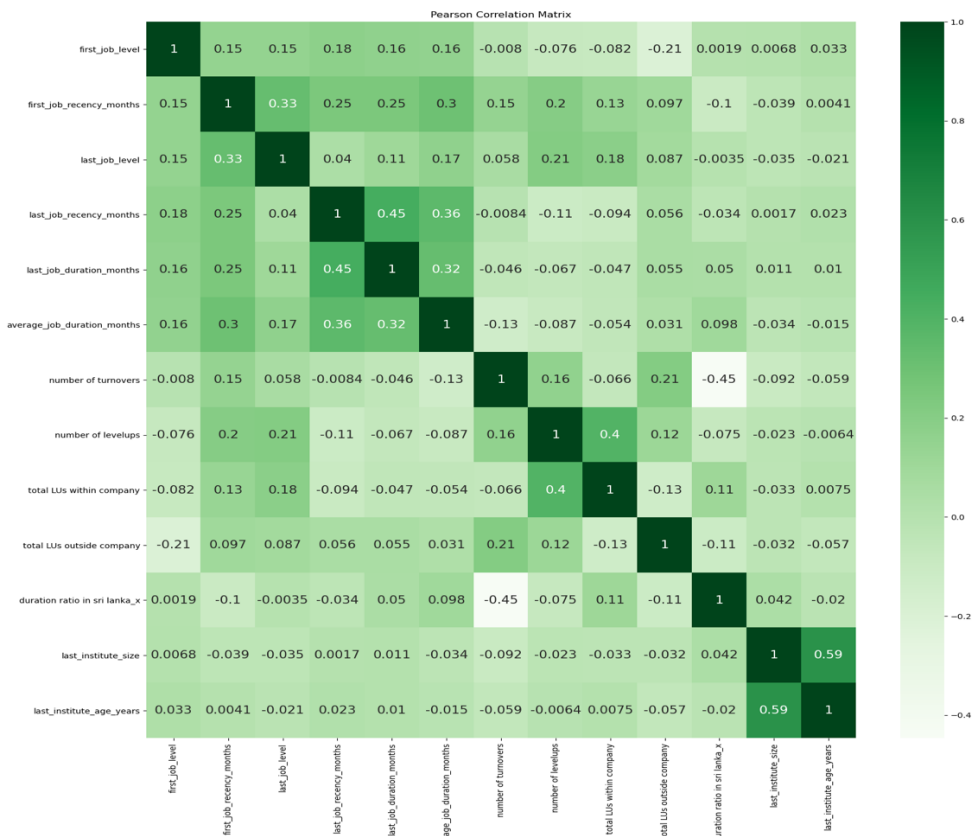


Figure 19: Correlation Matrix of the selected variables

Here we are developing a binary classification model which need to predict a binary output which is '1' for turnover expected and '0' for turnover not expected. For this we are using multiple pattern recognition and machine learning algorithms which will give the reliable and accurate results for the problem in hand. Out of the tried out different algorithms, we are planning to select and finalize with the most accurate, reliable and most interpretable model or algorithm after analyzing the results.

Below are few of the algorithms we have tried and their performance metrics accuracy, precision and recall. Based on all the 3 KPIs, we have selected the Gradient Boosting classifier, XGB classifier and Random Forest classifier for further developments and hyper parameter tuning.

	MLA Name	Accuracy	Precision	Recall
9	GaussianNB	69.33%	37.93%	63.46%
7	XGBClassifier	82.35%	60.87%	53.85%
3	AdaBoostClassifier	80.67%	56.25%	51.92%
1	DecisionTreeClassifier	71.43%	38.03%	51.92%
4	GradientBoostingClassifier	81.09%	58.14%	48.08%
5	BaggingClassifier	81.09%	54.84%	38.46%
2	RandomForestClassifier	83.19%	65.22%	36.54%
13	RidgeClassifierCV	81.51%	64.29%	34.62%
11	LogisticRegressionCV	78.57%	51.52%	32.69%
15	Perceptron	73.11%	36.96%	32.69%
14	SGDClassifier	78.15%	38.46%	28.85%
8	BernoulliNB	71.85%	33.33%	28.85%
12	PassiveAggressiveClassifier	80.25%	26.45%	25.00%
6	ExtraTreesClassifier	80.67%	77.78%	21.15%
0	KNeighborsClassifier	71.01%	28.21%	21.15%
10	LogisticRegression	78.57%	52.94%	17.31%

Table 4: Model performance

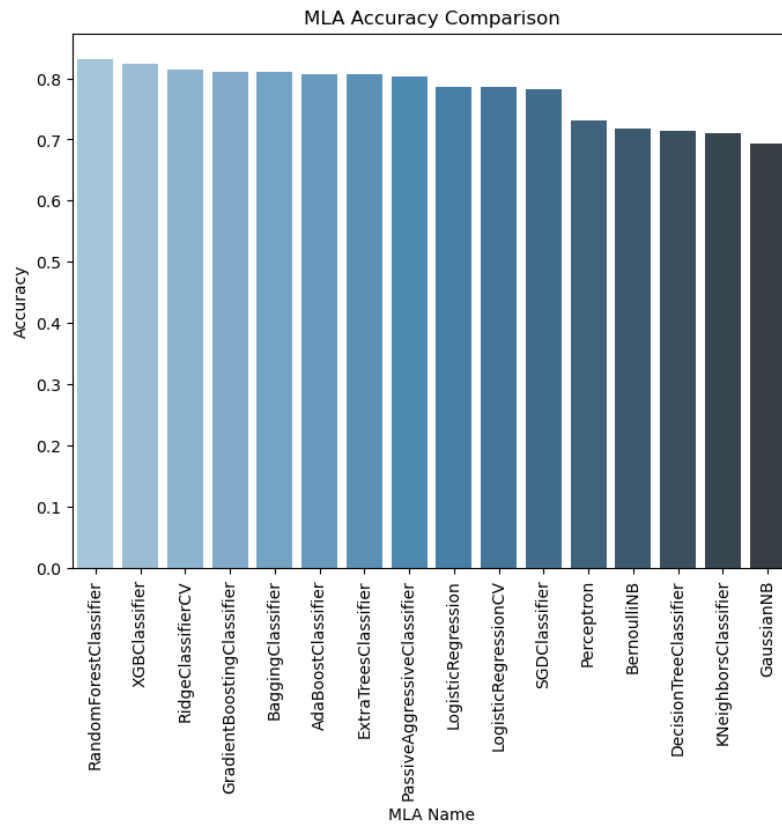


Figure 20: Turnover Probability Prediction Model Accuracy Chart

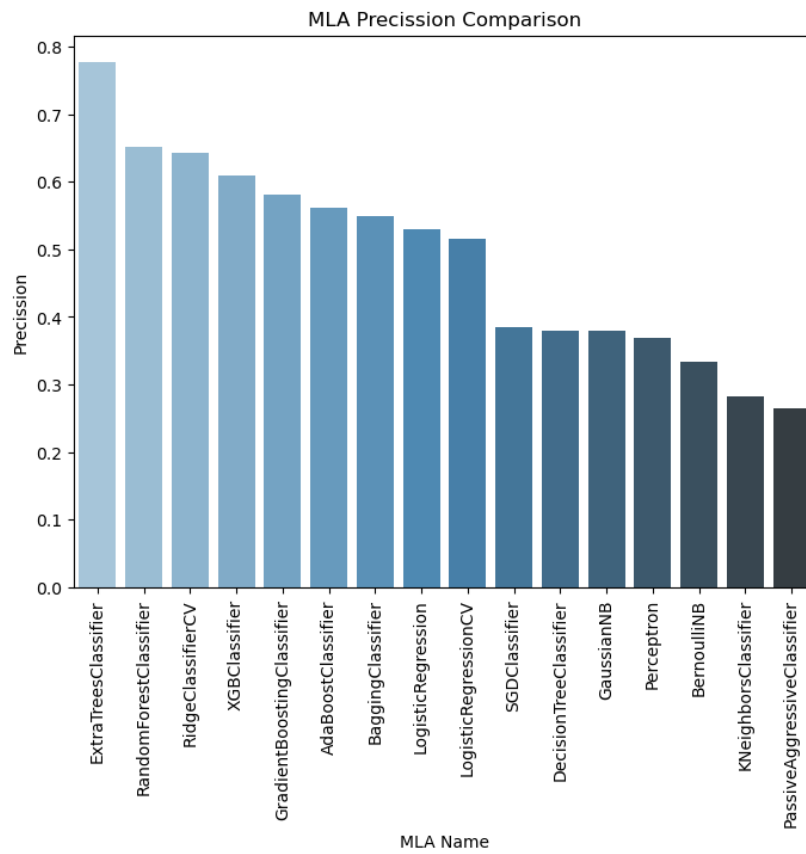


Figure 21: Turnover Probability Prediction Model Precision Chart

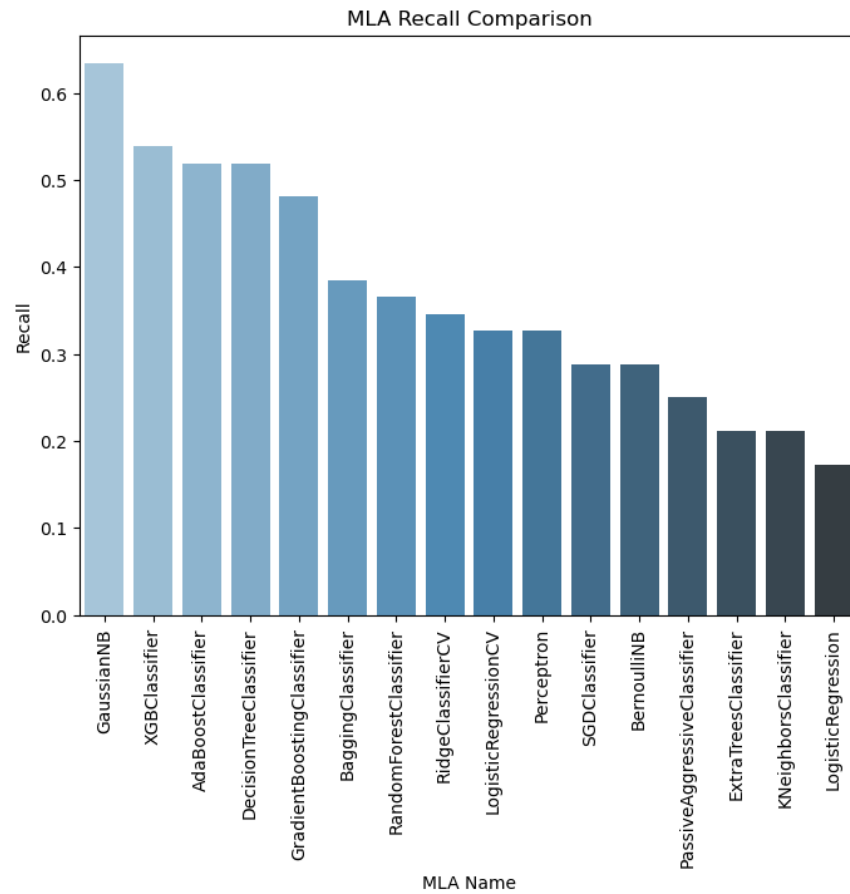


Figure 22: Turnover Probability Prediction Model Recall Chart

Based on the above algorithm analysis, we selected the below 3 models for further developments.

- GradientBoostingClassifier
- XGBClassifier
- RandomForestClassifier

Since XGB Classifier has given the best results after many scenario testings and hyper parameter tunings, we have selected the same as the finalized model.

3.7 Developing AI Model for Churn Horizon Prediction

Here we are developing a multi classification model which need to predict a ‘-1’ for no turnover expected, ‘0’ for turnover expected in the first six months and ‘1’ for turnover expected in the second six months. For this also we are using different pattern recognition and machine learning algorithms which will give the reliable and accurate results for the problem in hand. Out of the tried out different algorithms, we are planning to select and finalize with the most accurate, reliable and most interpretable model or algorithm after analyzing the results.

Below are few of the algorithms we have tried and their performance metrics accuracy, precision and recall. Based on all the 3 KPIs, we have selected the XGB classifier and Random Forest classifier for further developments and hyper parameter tuning.

	MLA Name	Accuracy	Precission	Recall
7	XGBClassifier	75.00%	75.00%	75.00%
8	BernoulliNB	75.00%	75.00%	75.00%
6	ExtraTreesClassifier	73.44%	71.88%	73.44%
2	RandomForestClassifier	73.44%	73.44%	71.88%
4	GradientBoostingClassifier	71.88%	73.44%	71.88%
11	LogisticRegressionCV	71.88%	71.88%	71.88%
0	KNeighborsClassifier	70.31%	70.31%	70.31%
3	AdaBoostClassifier	70.31%	70.31%	70.31%
10	LogisticRegression	68.75%	68.75%	68.75%
5	BaggingClassifier	68.75%	73.44%	67.19%
15	Perceptron	67.19%	67.19%	67.19%
9	GaussianNB	65.62%	65.62%	65.62%
13	RidgeClassifierCV	62.50%	62.50%	62.50%
14	SGDClassifier	62.50%	56.25%	53.12%
1	DecisionTreeClassifier	48.44%	51.56%	53.12%
12	PassiveAggressiveClassifier	57.81%	73.44%	45.31%

Table 4: Model performance

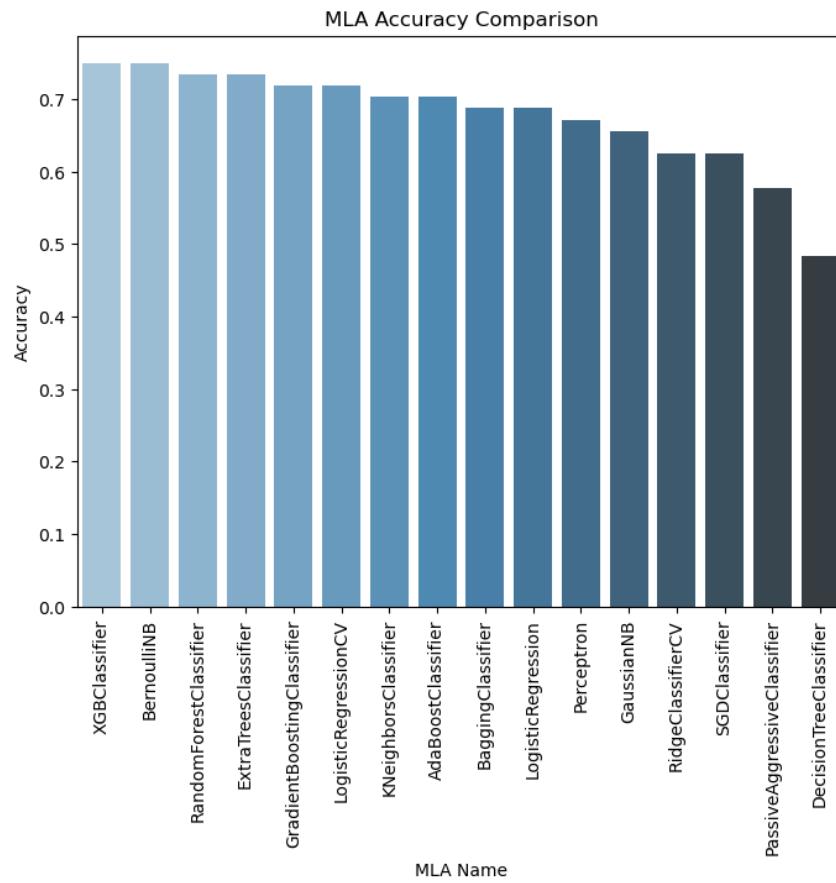


Figure 23: Turnover Horizon Prediction Model Accuracy Chart

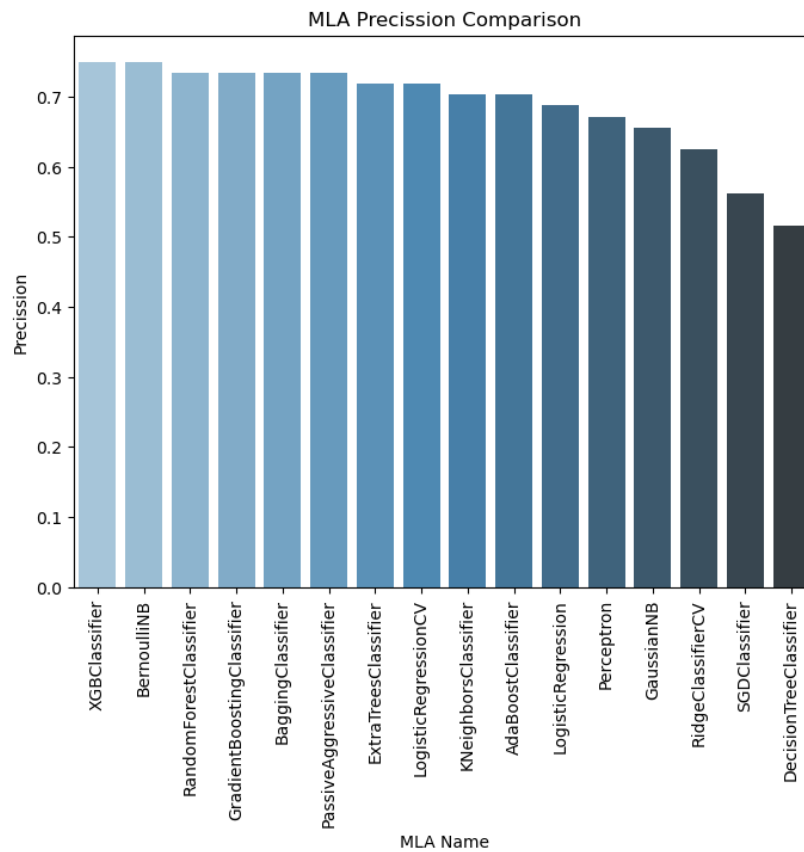


Figure 24: Turnover Horizon Prediction Model Precision Chart

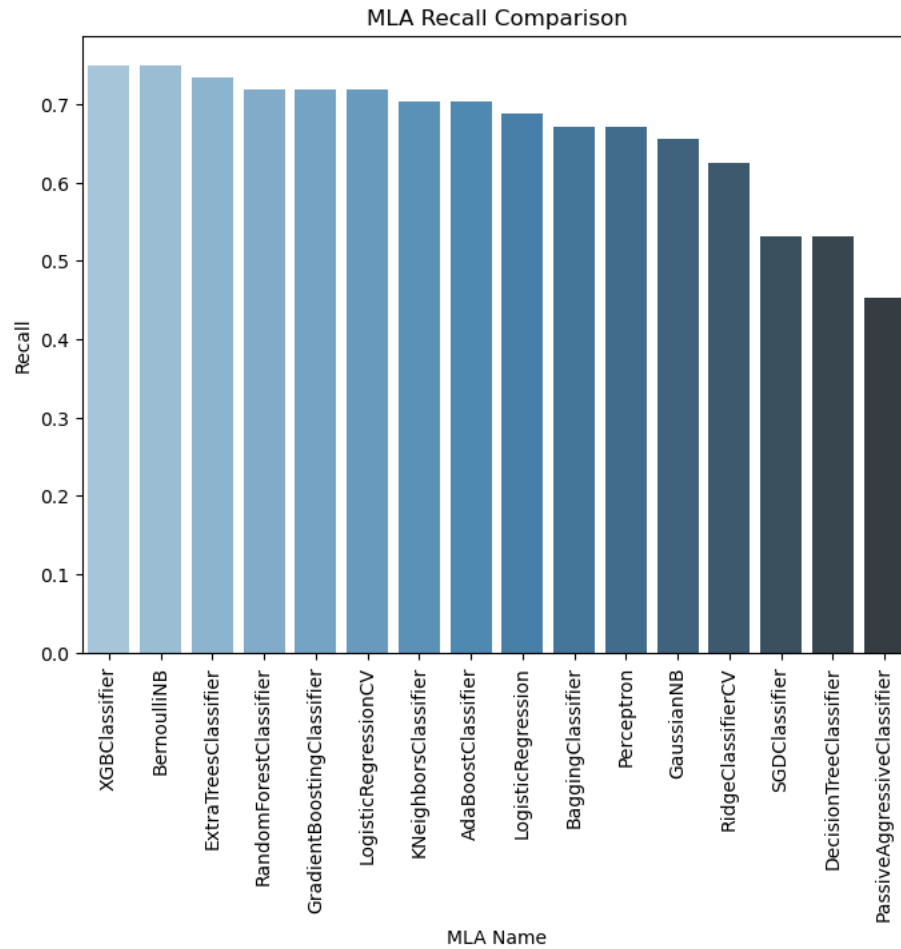


Figure 25: Turnover Horizon Prediction Model Recall Chart

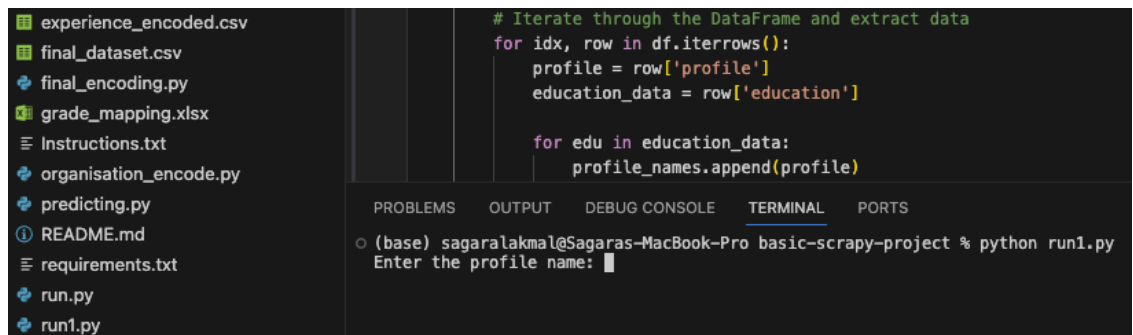
Based on the above algorithm analysis, we selected the below 3 models for further developments.

- Extra Trees Classifier
- XGB Classifier
- Random Forest Classifier

Since XGB Classifier has given the best results after many scenario testings and hyper parameter tunings, we have selected the same as the finalized model.

3.8 Developing UI

Then it is to develop a solution or UI for the end user where the user can input the required LinkedIn profile/s and get the 2 main outcomes which are employee turnover risk score and probable turnover timeline. We have used Visual Studio to create a command line interface and this can be further developed to have a graphical user interface.



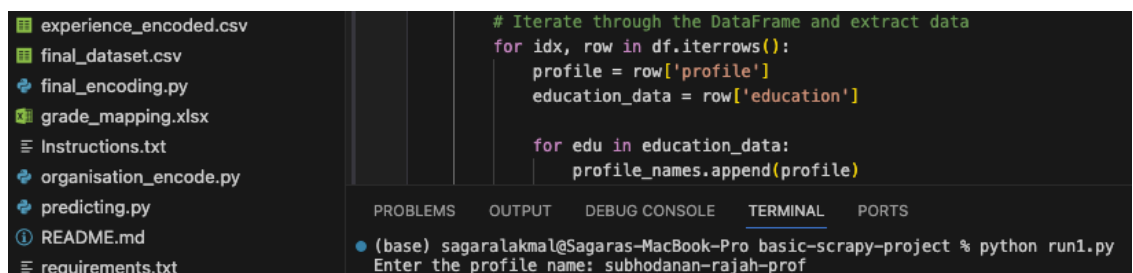
```
# Iterate through the DataFrame and extract data
for idx, row in df.iterrows():
    profile = row['profile']
    education_data = row['education']

    for edu in education_data:
        profile_names.append(profile)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

(base) sagaralakmal@Sagaras-MacBook-Pro basic-scrappy-project % python run1.py
Enter the profile name: █

Figure 26: Interface to enter the LinkedIn profile name



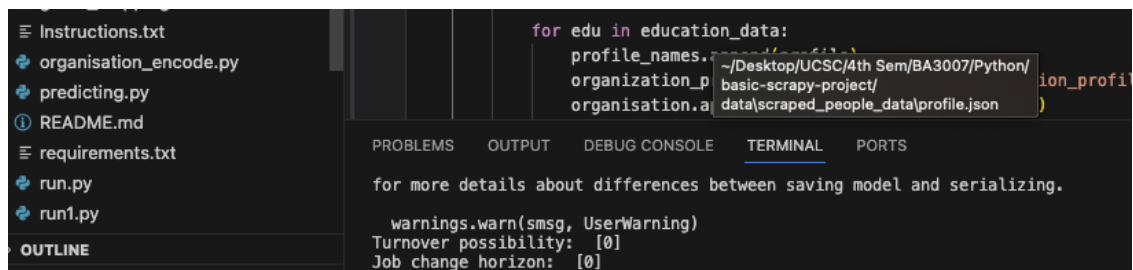
```
# Iterate through the DataFrame and extract data
for idx, row in df.iterrows():
    profile = row['profile']
    education_data = row['education']

    for edu in education_data:
        profile_names.append(profile)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

(base) sagaralakmal@Sagaras-MacBook-Pro basic-scrappy-project % python run1.py
Enter the profile name: subhodanan-rajah-prof

Figure 27: Interface where a LinkedIn profile name entered and run



```
for edu in education_data:
    profile_names.append(profile)
    organization_p = row['organization_p']
    organisation_a = row['organisation.a']
    ion_profi
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

for more details about differences between saving model and serializing.

warnings.warn(smsg, UserWarning)

Turnover possibility: [0]

Job change horizon: [0]

Figure 28: Output interface

3.9 Methodology Summary

Below is the methodology we have worked on, in a summary.

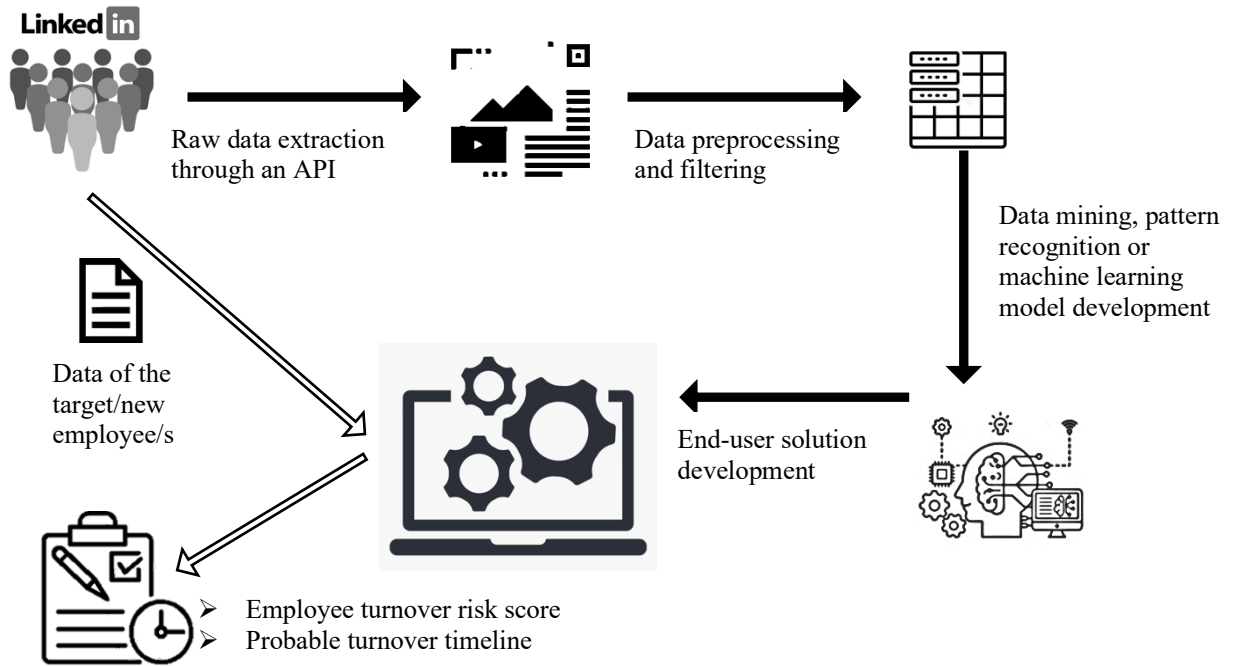


Figure 29: Methodology Summary

CHAPTER 4: RESULTS AND EVALUATION

Evaluation of this project is expected justify the problem statement that there is a critical issue of employee turnover based on the initial descriptive data analysis and then to determine the accuracy and reliability of the attrition prediction model built and assess the model's ability to generalize to new data. Below are the evaluation approaches we are planning to follow in this study.

4.1 Turnover Behavior of the Executive and above Employees:

For the initial descriptive analysis of the employee turnover behavior, we have considered an executive and above employee dataset from one of the big 3 apparel companies of Sri Lanka, which covers more than 50% of the population.

As per the analysis, executive and above employee turnover count as well as the turnover percentage out of the total count have been constantly increasing during the last 5 years. This shows that, employee turnover has become a major concern within the apparel sector and need to identify ways to minimize it.

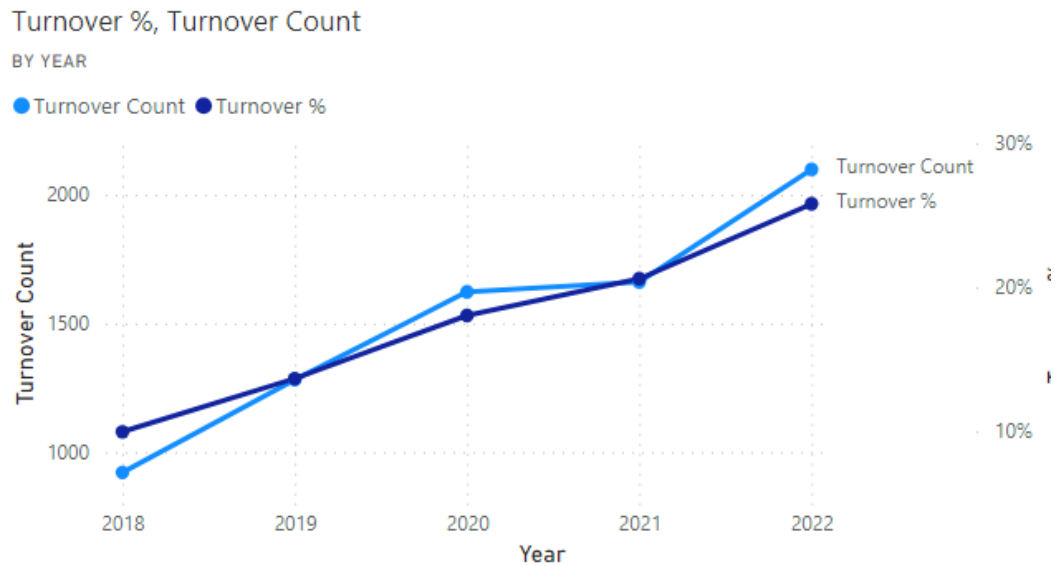


Figure 30: Employee turnover count and percentage out of total

Highest Turnover has been with the entry level employees and it has been continuously increased during the last 5 years.

Turnover Count

BY GRADE

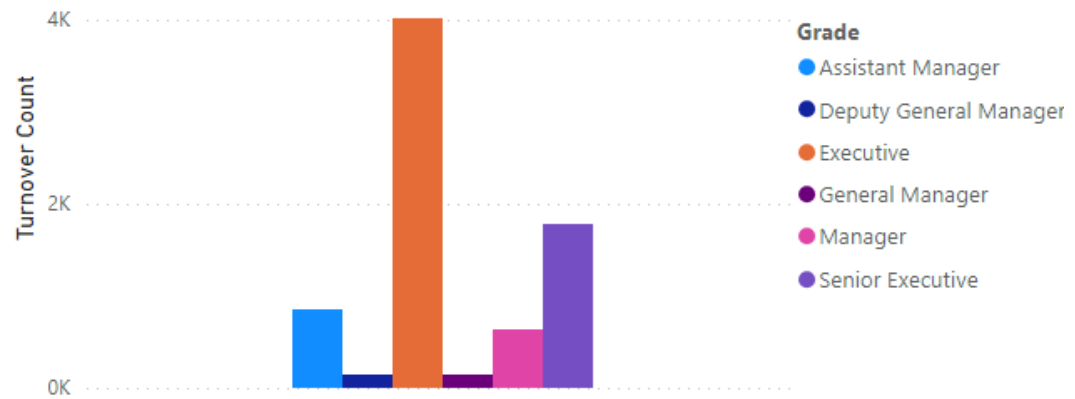


Figure 31: Employee turnover grade wise

Turnover Count

BY GRADE, YEAR

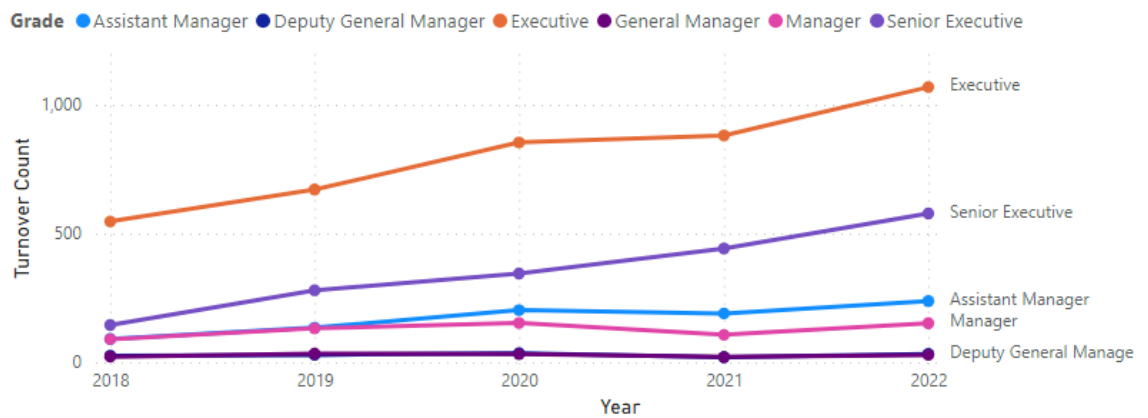


Figure 32: Employee turnover grade wise trend

When analyzing the number of years the employee has been in the company before the termination, it is evident that there is a continuous increase in the turnover of the employees who have been working in the company for 5-10 years and 10+ years, when compared with the rest. This is a very critical issue for the company because they are losing the experienced talent, not the new recruits or raw talent.

Turnover Count

BY # YEARS IN THE COMPANY

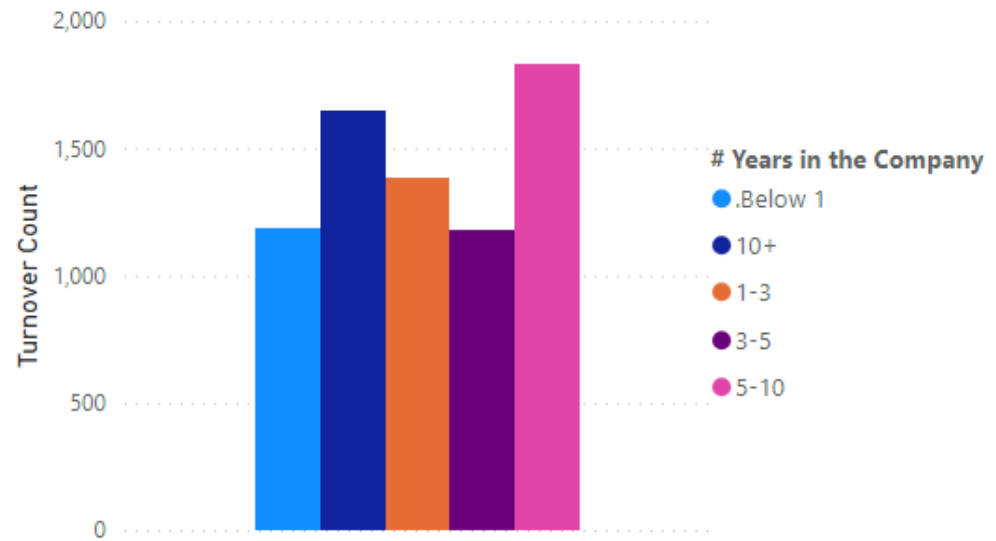


Figure 33: Employee turnover number of years in the company wise

Turnover Count

BY YEAR, # YEARS IN THE COMPANY

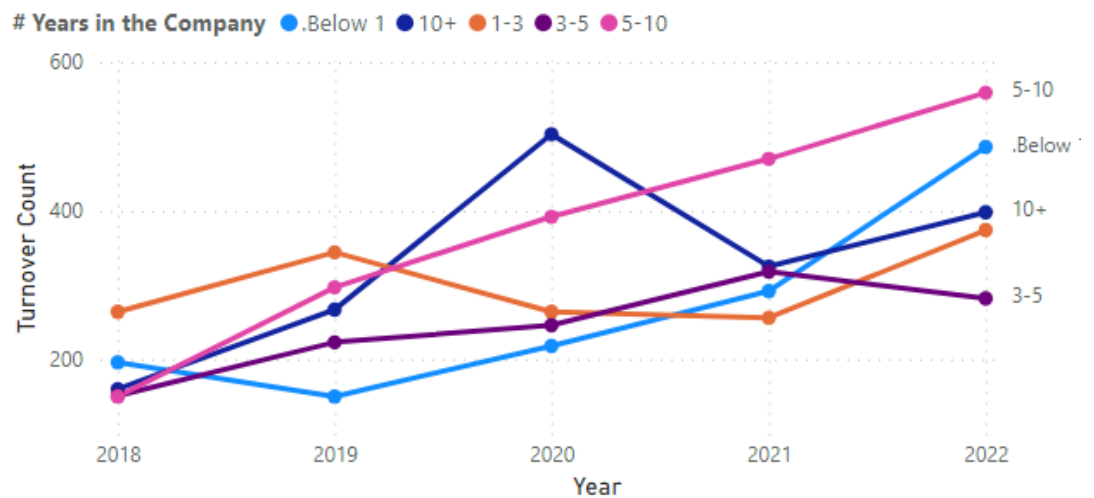


Figure 34: Employee turnover number of years in the company wise trend

When looking at the number of promotions/ level-ups an employee has got in the company by the time of turnover, it is evident that the lesser the number of level-ups, higher the turnover count.

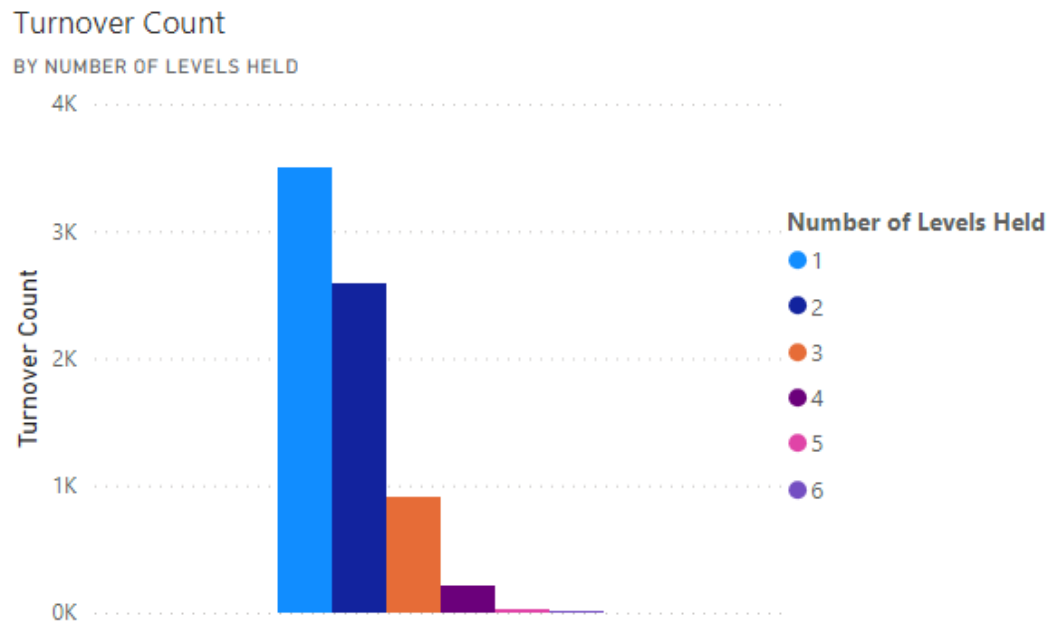


Figure 35: Employee turnover number of jobs held wise

It is also evident that more the number of years an employee stay in one grade, higher the turnover count.

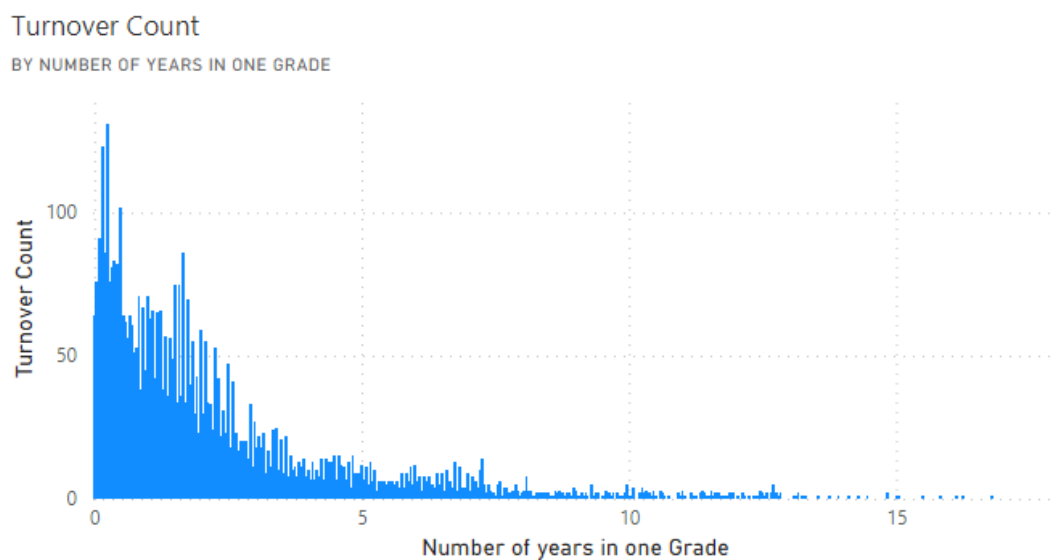


Figure 36: Employee turnover number of jobs held wise trend

We have also checked the correlation of the matrix and other than grade jump and grade count nothing is highly correlated.

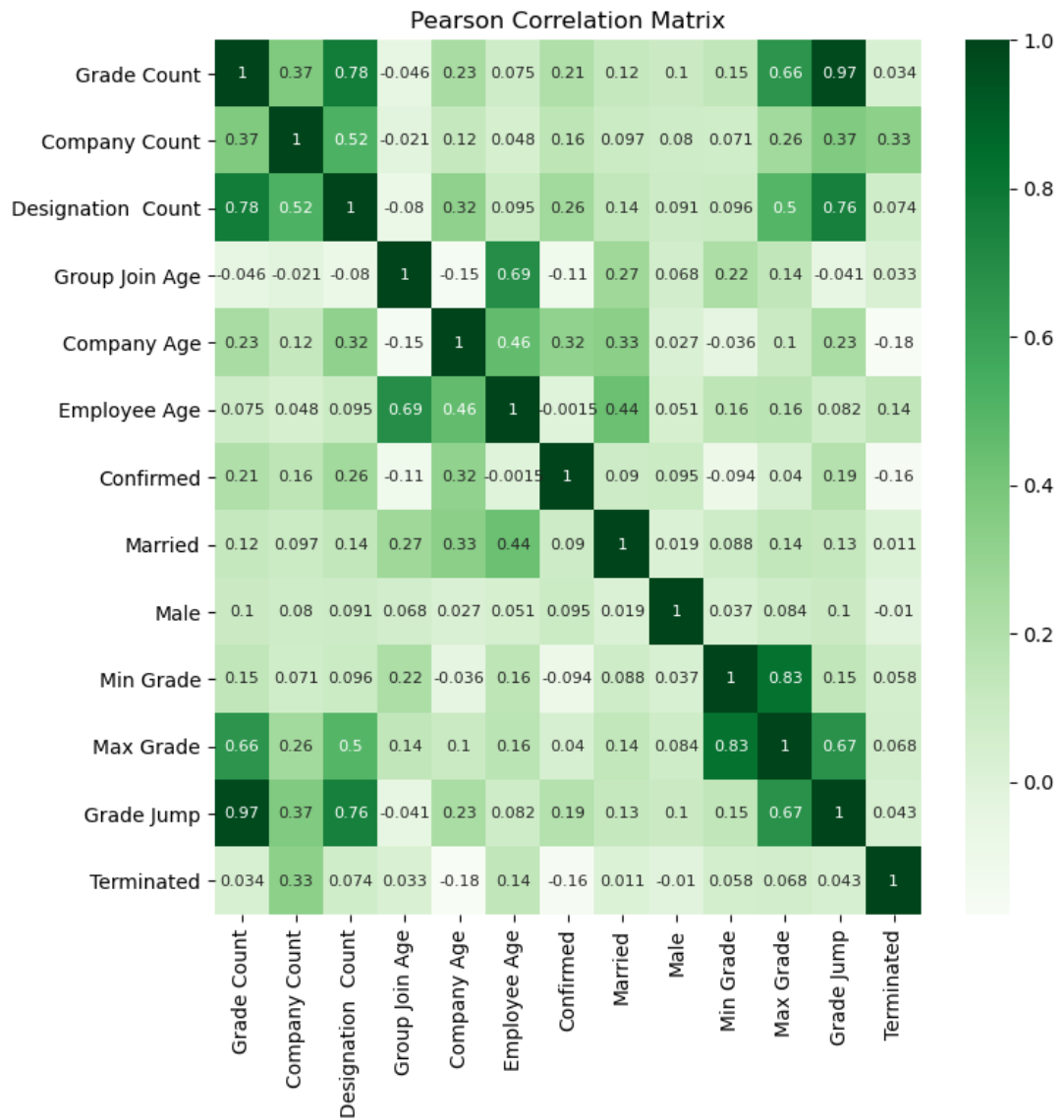


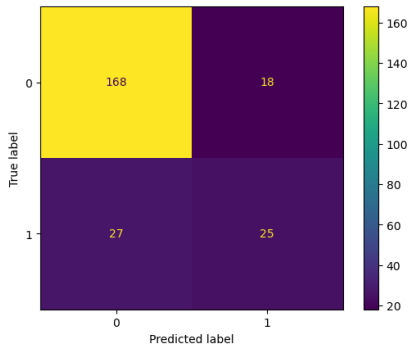
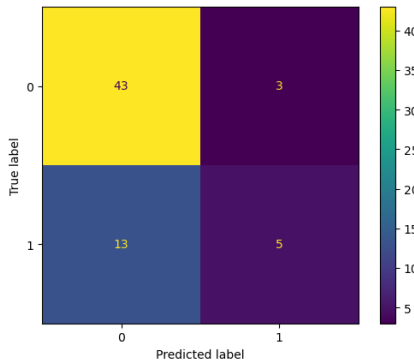
Figure 37: Feature correlation matrix

Therefore, based on the above analysis results, it is evident that there is a real critical employee turnover problem in the apparel industry and it is required to identify it in advance and take the required actions in order to minimize the turnover.

4.2 Evaluation of the Prediction Models:

For the evaluation of the two models, we have identified below standard KPIs used to evaluate classification models.

- Accuracy and Precision: Measure the overall correctness of predictions and the precision of attrition predictions.
- Recall and Sensitivity: Assess the model's ability to identify all instances of actual attrition.
- F1 Score: Balance between precision and recall, considering both false positives and false negatives.
- Area Under the ROC Curve (AUC-ROC): Evaluate the model's ability to discriminate between attrition and non-attrition cases.
- Confusion Matrix Analysis: Examine true positives, true negatives, false positives, and false negatives.

KPI	Turnover Possibility Prediction Model	Turnover Horizon Prediction Model
Accuracy	0.81	0.75
Precision	0.58	0.63
Recall	0.48	0.28
F1 Score	0.69	0.61
AUC-ROC	0.53	0.38
Confusion Matrix	 <p><i>Figure 38: Model 1 confusion matrix</i></p>	 <p><i>Figure 39: Model 2 confusion matrix</i></p>

Base on the above KPIs we can mention that overall accuracy of the 2 models is above 75% and in an acceptable level. However, precision and recall need to be improved further. For this we can,

- Try more algorithms and ML models
- Gather more data and retrain the models
- Identify and add more attributes which can improve the accuracy.

According to the turnover possibility model, out of the available features, below are the most impactful variables arranged in descending order.

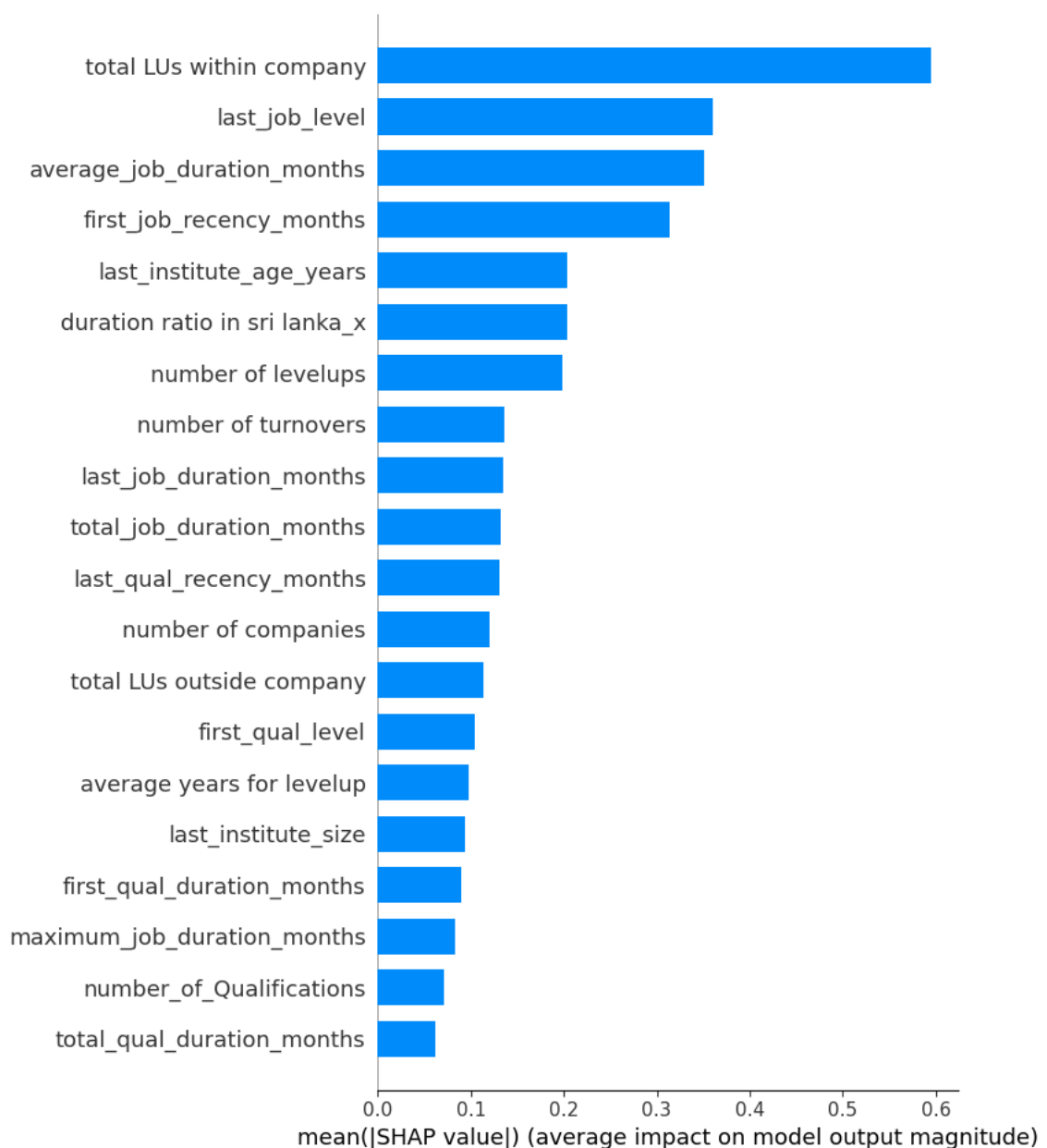


Figure 40: Turnover Probability Prediction Model Feature Importance

According to the turnover horizon model, out of the available features, below are the most impactful variables arranged in descending order.

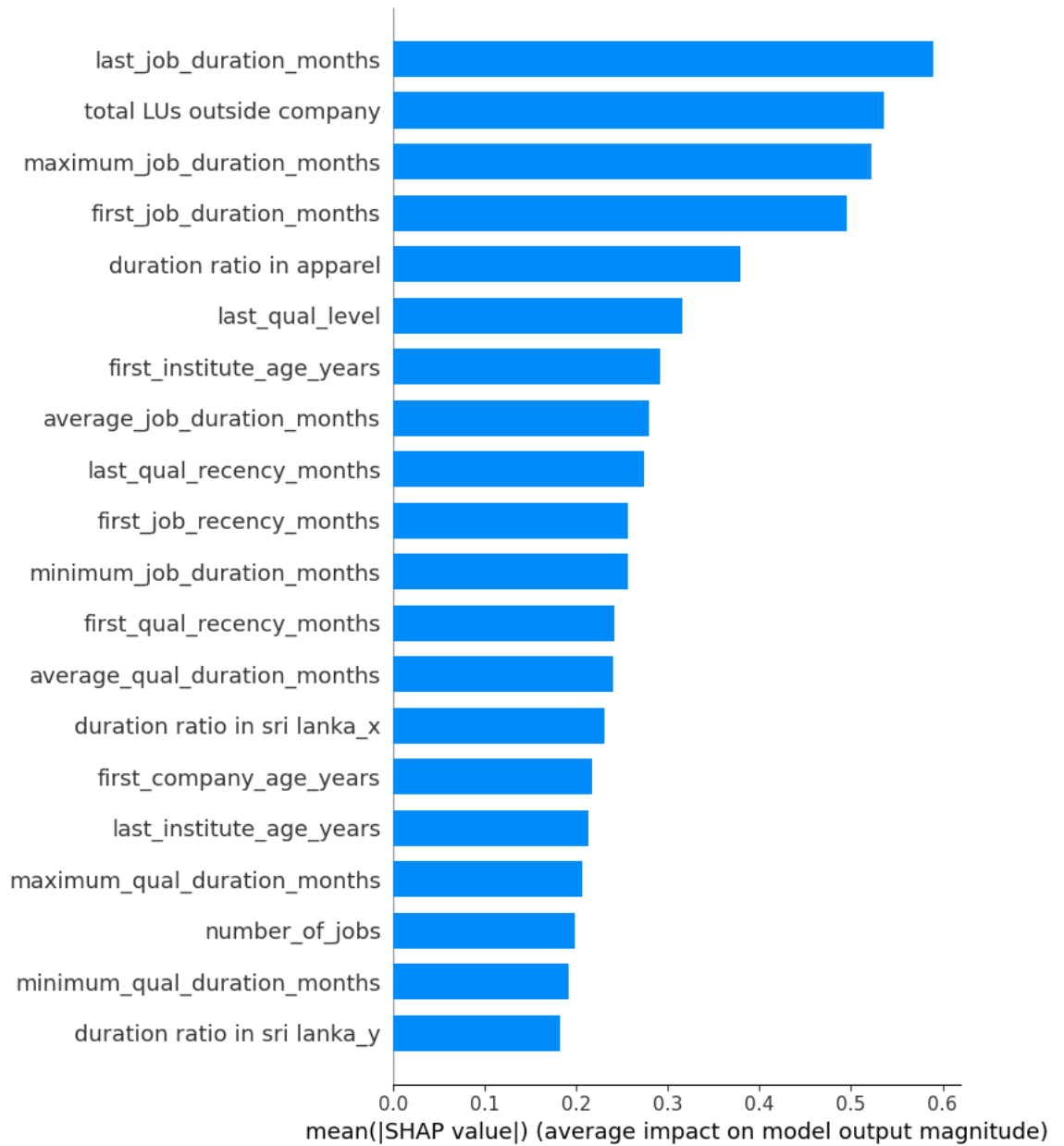


Figure 41: Turnover Horizon Prediction Model Feature importance

CHAPTER 5:

CONCLUSION AND FUTURE WORK

5.1 Conclusion:

In this thesis work, we focus on the employee turnover problem of the executive and above cadre in the Sri Lankan Apparel industry and try to predict whether the employee will leave his/her current company within the coming year and the probable time period.

To address the first research objective, we have analyzed the executive and above employee behavior of one of the biggest apparel manufacturing companies which covers more than 50% of the population and with the descriptive analysis we have identified that executive and above employee turnover actually has become a critical problem in the apparel domain during the last 5 years and the employees who are leaving the company are also not new recruits but long-term and skilled employees. We also discovered that longer the employees remain in one grade are more likely to turnover and more versatile the job roles held, employees are less likely to turnover. So with the descriptive analysis on a real company dataset, it is evident that this is a real critical problem in the apparel domain and it is required to identify it in advance and take the required actions in order to minimize the turnover.

When moving into the prediction part of the project, we have formulated the turnover possibly model as a binary classification problem which classifies the employees as who have changed the job in the last year as “1” and who continue to work for current company as “0”. For the turnover horizon model, we have used a multi classification model where we have given the labels ‘-1’ for no turnover expected, ‘0’ for turnover expected in the first six months and ‘1’ for turnover expected in the second six months.

Since the internal models built by the company, using internal employee system and survey data didn’t show reliable results, we focused on data sources which are available outside the organization and for the model training we have used publicly available employee data from LinkedIn which was scraped through a paid API and after that preprocessed and encoded as required for the model development using python.

We apply supervised machine learning algorithms for these classification tasks and we

have tried out more than 15 different algorithms. Out of all XGB Classifier algorithm gave the best result and same was selected to build the final model. Then we have tuned the hyper parameters of the model and was able to improve the model KPIs slightly. We have also tried Ensemble method to try combining multiple models, but it didn't give better results. Experimental results are evaluated with accuracy, precision, recall, F1 score and AUC-ROC metrics. For the turnover possibility model, we achieved an accuracy of 81% and for the turnover horizon model, we were able to achieve an accuracy of 75%.

As the most influential features for the turnover possibility, we have got the features such as number of level-ups in the company, level-ups outside the company, maximum job duration, last job grade, average job duration, tenure of employment, last qualification level and the time since last academic qualification obtained. Therefore it is clear that these features have a high influence in an employee taking a decision to change the job and hence should be monitored continuously. Finally, we have developed a user interface for the end user where the user can input the required LinkedIn profile/s and get the 2 main outcomes which are employee turnover risk score and probable turnover timeline. We have used Visual Studio to create a command line interface and this can be further developed to have a graphical user interface.

As the conclusion, we can state that there is a real problem of executive and employee turnover in the apparel industry. It seems that models built with internal data doesn't give much reliable results. Solution we have built is able to predict the possibility and horizon of an employee turnover and with the given user interface, any user can easily use the tool to obtain the same in a very quick time.

5.2 Future Work:

In this study we have used only publicly available data since the internal models didn't show much accurate result. As the next step we are planning to combine the external data with internal data and fine tune the model further to get a higher accuracy and prediction performance. With more personal data like salary and employee feedback, we will be able to identify more influential variables for the employee turnover.

We focus on only "external" turnovers of the employees. "internal" turnovers (intercompany position changes) can also be studied. In this study, only the employees working in the apparel industry are analyzed. However, each industry may have a different turnover pattern. As a future work, the problem can be specified as the analysis of the employees from a certain industry or certain company in terms of turnover.

We use publicly available employee and company information to extract the features and train our models. With new features, even detailed features, our models can be improved. The influence of social connections of the employees for turnover problem can be studied. Additionally, we analyze the turnovers from employee perspective. From company perspective, the problem can be stated as prediction of company turnover rate by considering employee turnovers and can be studied as a regression problem.

APPENDICES

A.1 Final Feature List

No	Type	Feature	Example
1	Personal	profile	sagara-lakmal-3b5634105
2		male	1
3		sri_lankan	1
4	Job Data	number of jobs	4
5		first job level	1
6		first job recency months	67
7		first job duration months	10
8		first company size	4
9		first company age years	36
10		first company apparel	1
11		first company sri lankan	1
12		last job level	5
13		last job recency months	19
14		last job duration months	19
15		last company size	2
16		last company age years	44
17		last company apparel	0
18		last company sri lankan	1
19		minimum job duration months	10
20		maximum job duration months	20
21		average job duration months	16
22		total job duration months	66
23		number of turnovers	2
24		number of companies	3
25		number of levelups	3
26		total level ups	4
27		average years for levelup	1.30
28		total LUs within company	2
29		total LUs outside company	2
30		number of lateral movements	0
31		duration ratio in sri lanka	1
32		duration ratio in apparel	0.41
33	Education	number of qualifications	4
34		first qualification level	2
35		first qualification recency years	2
36		first qualification duration years	3
37		first institute size	3
38		first institute age years	137
39		first institute sri lankan	1
40		last qualification level	5
41		last qualification recency years	2

42	last qualification duration years	3
43	last institute size	1
44	last institute age years	21
45	last institute sri lankan	1
46	minimum qualification duration years	1
47	maximum qualification duration years	8
48	average qualification duration years	4.25
49	total qualification duration years	17
50	number of institutes	4
51	qualification level ups	4
52	duration ratio in sri lanka	0.94
53	number of qualifications	4
54	first qualification level	2
55	first qualification recency years	2

REFERENCES

- [1] H. Ongori, "A review of the literature on employee turnover," *African Journal of Business Management*, pp. 49–054, 2007, Accessed: Mar. 28, 2023. [Online]. Available: <https://ubrisa.ub.bw/handle/10311/1154>
- [2] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," *Advances in Intelligent Systems and Computing*, vol. 869, pp. 737–758, 2018, doi: 10.1007/978-3-030-01057-7_56/COVER.
- [3] E. Ribes, K. Touahri, and B. Perthame, "Employee turnover prediction and retention policies design: a case study," Jul. 2017, Accessed: Mar. 30, 2023. [Online]. Available: <https://arxiv.org/abs/1707.01377v1>
- [4] A. A. B, "ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS," *Information Systems & Development Informatics*, vol. 4, no. 1, 2013.
- [5] M. Teng, H. Zhu, C. Liu, C. Zhu, and H. Xiong, "Exploiting the Contagious Effect for Employee Turnover Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 1166–1173, Jul. 2019, doi: 10.1609/AAAI.V33I01.33011166.
- [6] A. Mohammad Esmacieeli Sikaroudi and A. Esmacieeli Sikaroudi, "A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *Journal of Industrial and Systems Engineering*, vol. 8, no. 4, pp. 106–121, 2015.
- [7] R. Punnoose and C. Xlri -Xavier, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 9, 2016, doi: 10.14569/IJARAI.2016.050904.
- [8] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 113–120, Nov. 2018, doi: 10.1109/SYSMART.2018.8746940.
- [9] J. Sukhadiya, H. Kapadia, and P. Mitchell D'silva 1 Student, "Employee Attrition Prediction using Data Mining Techniques".
- [10] S. Ranjitha Ponnuru *et al.*, "Employee Attrition Prediction using Logistic Regression," vol. 8, 2020, doi: 10.22214/ijraset.2020.5481.
- [11] H. Zhang, L. Xu, X. Cheng, K. Chao, and X. Zhao, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning," *ISCIT 2018 - 18th International Symposium on Communication and Information Technology*, pp. 433–437, Dec. 2018, doi: 10.1109/ISCIT.2018.8587962.
- [12] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, pp. 1016–1020, May 2018, doi: 10.1109/ICICI.2017.8365293.
- [13] A. G. N. K. Fernando, "Factors Impact on Employee Turnover with Special Reference to the Apparel Industry in Sri Lanka," *International Journal of Research and Innovation in Social Science*, pp. 2454–6186, 2019, Accessed: Mar. 31, 2023. [Online]. Available: www.rsisinternational.org
- [14] R. M. Iniyan, "Talent Flow Employee Analysis based Turnover Prediction on Survival Analysis," 2021. [Online]. Available: <http://annalsofrscb.ro3844>
- [15] A. C. C. De Jesus, M. E. G. D. Júnior, and W. C. Brandão, "Exploiting linkedin to predict employee resignation likelihood," *Proceedings of the ACM*

- Symposium on Applied Computing*, pp. 1764–1771, Apr. 2018, doi: 10.1145/3167132.3167320.
- [16] Q. Zhu, J. Shang, X. Cai, L. Jiang, F. Liu, and B. Qiang, “CoxRF: Employee turnover prediction based on survival analysis,” *Proceedings - 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, pp. 1123–1130, Aug. 2019, doi: 10.1109/SMARTWORLD-UIC-ATC-SCALCOM-IOP-SCI.2019.00212.
 - [17] D. S. Rodrigo and G. S. Ratnayake, “Employee Turnover Prediction System: With Special Reference to Apparel Industry in Sri Lanka,” *2021 6th International Conference for Convergence in Technology, I2CT 2021*, Apr. 2021, doi: 10.1109/I2CT51068.2021.9418108.
 - [18] W. H. Mobley, R. W. Griffeth, H. H. Hand, and B. M. Meglino, ‘Review and conceptual analysis of the employee turnover process’, *Psychol Bull*, vol. 86, no. 3, pp. 493–522, May 1979, doi: 10.1037/0033-2909.86.3.493.
 - [19] M. Teng, H. Zhu, C. Liu, C. Zhu, and H. Xiong, ‘Exploiting the Contagious Effect for Employee Turnover Prediction’, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 1166–1173, Jul. 2019, doi: 10.1609/AAAI.V33I01.33011166.
 - [20] H. Ongori, ‘A review of the literature on employee turnover’, *African Journal of Business Management*, pp. 49–054, 2007, Accessed: Mar. 28, 2023. [Online]. Available: <https://ubrisa.ub.bw/handle/10311/1154>
 - [21] Q. Zhu, J. Shang, X. Cai, L. Jiang, F. Liu, and B. Qiang, ‘CoxRF: Employee turnover prediction based on survival analysis’, *Proceedings - 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, pp. 1123–1130, Aug. 2019, doi: 10.1109/SMARTWORLD-UIC-ATC-SCALCOM-IOP-SCI.2019.00212.
 - [22] A. C. C. De Jesus, M. E. G. D. Júnior, and W. C. Brandão, ‘Exploiting linkedin to predict employee resignation likelihood’, *Proceedings of the ACM Symposium on Applied Computing*, pp. 1764–1771, Apr. 2018, doi: 10.1145/3167132.3167320.
 - [23] Central Bank and Sri Lanka, ‘External Sector Performance-December 2022’, 2022.