Who's Next: Evaluating Employee Churn in Retail using Machine Learning algorithm CHAID

Meenu Chaudhary[1], Loveleen Gaur[2], Amlan Chakrabarti[3]
[1]Amity International Business School, Amity University, India
[2]Amity International Business School, Amity University, India
[3]A.K. Choudhury School of IT, University of Calcutta, India

Abstract: Researchers have undertaken numerous studies on employee churn, but retail employers are facing challenges to retain employees. Academicians have emphasized the factors promoting employee churn and have also incorporated ML algorithms to predict employee churn. But the research related to employee churn prediction in the retail sector is missing. The data was collected from 446 retail employees using a questionnaire. This research attempts to demonstrate the performance of ML algorithm i.e., CHAID for predicting retail employee churn using real-time data. The performance assessment of CHAID is based on accuracy, confidence matrix, CHAID decision tree, and AUC. The deployment of efficient employee churn strategies is essential for businesses to survive in today's competitive marketplace, given that lowering employee churn increases productivity and profitability.
Keywords: Employee churn, CHAID, retail, machine learning

Introduction

Employee Churn (ECn) has an undesirable impact on retail firms resulting in decreasing productivity, profitability, and sustainability. Bothma and Roodt [1] discovered that ECn in the retail sector increased the operational budget by 9%. Several research indicates that ECn has an adverse effect on an organization's long-term viability as losing essential employees results in revenue loss, knowledge loss, and productivity loss [2] [3]. Leaders must retain key workers in a competitive market in order to continue delivering long-term value for the stakeholders.

Business practices have demonstrated that traditional models with complex components and repetitious evaluations cannot adapt to the rapidly changing market environment. Nonetheless, big data and artificial intelligence give HRM new vigor. Intelligent HRM is still a relatively new field with limited research, waiting to be explored and expanded. This can be supported by the number of research in this field over the past years. The publications on ECn and machine learning (ML) have been accelerated post-2017 with a maximum of thirty-eight relevant research in 2022. This is in contrast to traditional HRM, which favors qualitative methodologies and psychological understanding. Due to advances in IT, academics have also investigated several ML approaches to improve HRM outcomes. Various ML algorithms such as decision trees, support vector machines, logistic regression, neural networks, naive Bayes, and K-nearest neighbor have been compared in most research papers. Most studies have been conducted in the field of higher education [4], IT [5], healthcare, software [6], and manufacturing [7]. Besides concentrating on classification and prediction abilities, numerous scholars have made significant attempts to better comprehend which characteristics are most influential in forecasting ECn. In data mining applications, these characteristics often bear equal weight, therefore it is advantageous to obtain a deeper grasp of their significance.

Numerous tree-based research measured feature significance by evaluating the decreasing impurity per node-split in decision trees. In addition, customized ML and sensitivity analysis have been utilized to comprehend relative feature significance. Various research have also developed classification guidelines or displayed the classification technique providing more knowledge and confidence in the application of ML approaches.

Researchers have undertaken numerous studies on ECn, but retail employers continue to suffer with retaining top employees. Organizations struggle to manage and predict employee churn. Academicians have emphasized the factors promoting ECn and have also used ML for predicting employee churn. But the research related to employee churn prediction using ML algorithm in the retail sector is missing. The objective of this paper is to assess ECn using best-performing ML for retail employees.

Literature review

Using the keywords 'employee churn', 'employee turnover', and 'retail'; around sixty-four research articles were screened. Out of which fifty-four were found relevant for the study. The authors have included the research papers and articles whose impact factor is greater than one. Peterson [8] introduced the concept of integration and the new ECn model to the literature on employee churn. The results highlight the significance of management development in establishing and maintaining an organizational culture that encourages the retention of managerial staff. The author evaluated the correlation between ECn and customer satisfaction ratings across two hundred seventy five fast-food convenience outlets run by two separate retail companies. The clients' perception of service speed decreased when part-time ECn increased, indicating a negative link between the two variables [9]. Pandey et al [10] used an analytical hierarchy process to identify the variables that may affect retail employees' retention. The hierarchical model consists of three criteria, seven sub-criteria, and three possibilities i.e., work environment, managers, and employees, autonomy, socialization, fairness, support, behaviour, and attitude, management by objective, engagement, and management by instruction.

Through open-ended interviews and survey responses, the author conducted a thematic analysis utilizing NVIVO for a field study of 18 employees [11]. The study is based on the 301 retail workers to analyze the turnover intentions. McCartney et al [12] revealed a substantial positive correlation between job satisfaction (JS) and workload, compensation and company support (CS). Relationships with co-workers did not effect job satisfaction. Additionally, JS has a substantial detrimental impact on employee turnover intention.

Considering the adverse effects of employee churn, it has apprehended the interest of scholars and practitioners by using ML algorithms for predicting employee churn post-2014. Three factors prompt the adoption of ML algorithms to solve the ECn problem. First, the organization lacks the necessary resources to manually estimate staff turnover. Second, a substantial availability of the amount of data for ECn prediction, which should be exploited effectively to make a conclusion [13]. Thirdly, the available dataset is regularly updated, thus it is not inconsistent. Studies have been conducted in the telecommunication and IT sectors. However, substantial research on retail employees is missing. Existing research primarily focuses on using datasets available on different platforms online for building models for employee churn analysis.

Table 1: Summary of some prominent researchers in the area of employee churn and machine learning

Authors	Dataset	Journal	Impact factor	Application area
Jain et al [14]	Kaggle dataset	Journal of Intelligent Information Systems	2.94	Generalized
Najafi et al [15]	IBM HR dataset	Mathematics	2.592	Generalized
Cai et al [16]	Employee data from online professional social network websites	IEEE Access	4.34	Generalized
Chaudhary et al [17]	Kaggle dataset	Intelligent Automation and Soft Computing	3.54	Generalized
Jain et al [18]	Kaggle dataset	SN Applied Sciences	2.11	Generalized
Avrahami et al [19]	Archival data	International Journal of Manpower	3.15	Generalized
Srivastava & Eachempati [20]	Real-time data from the FMCG Sector Company	Journal of Global Information Management	3.03	FMCG
Khera, S. N., & Divya [21]	Archival employee data	Vision: The Journal of Business Perspective	1.528	IT

As mentioned in table 1, the existing literature primarily relies on the generalized datasets available on different platforms. Najafi [22] introduced a three stage approach pre-processing, processing, and post-processing for churn analysis. In the logistic regression model, the coefficient of each feature reflects the feature's predictive value for attrition. The authors used the GBDT method and the LR algorithm to fit a model of ECn influencing characteristics. This study executes the ECn forecast in contemporary firms, providing a useful reference for companies seeking to reduce staff turnover [23]. Four classifiers were used to show the relationship between attributes illustrated through a correlation matrix and heatmap [24]. Using a sample size of 1469, the authors compared state-of-the-art solutions for the suggested ML methods. This research intends to compare several ML techniques [25]. Comparison of the entropy weight method and ML algorithms have been conducted to employee reviews and responses were used to gauge employee satisfaction, as the majority of research has focused on applying LDA for customer evaluations [26].

The Entropy weight method and C5 classifier determine that job satisfaction is the most noteworthy predictor. However, EWM is limited to provide relative weightage, whereas C5 reassures the overall classifier's accuracy value and also evaluates the prediction accuracy for individual employees. [27]. The discrepancy between the number of hours worked and mentioned in agreement is the most significant predictor of an employee's intent to leave, indicating that, to increase employee retention, the employee should be given the number of hours specified in the contract. Varied shift duration and continuity of services and patients appear to link with lower churn, according to secondary data [28]. The authors describe an IoT-enabled predictive technique to assess staff attrition rate and the elements that can be used to mitigate the issue in enterprises. In order to identify potential churners, the authors use filter-based algorithms to assess features and classify firms [29]. Another study utilizing real data from a big pharmaceutical business and a mix of quantitative and qualitative methodologies. The hybrid approach can find the causes of churn, allowing retention plans to capitalize on the benefits of both techniques [30]. Although studies are available on EC using various ML algorithms, there is no existing literature on churn analysis of retail employees using ML. Further it is related approached also used for different applications [31-35].

Research Methodology

The responses were collected from 543 employees in the retail sector. However, only 446 responses were used due to inaccurate and missing values. Out of 495 employees, 295 are males and 151 are females. The authors used IBM SPSS Modeler 18.4 version and IBM SPSS Statistics version 28.0.1. The information collected from the respondents is converted into data and then entered in IBM SPSS Statistics 28.0.1 after deriving the weighted score for job satisfaction, job involvement, work-culture/environment satisfaction, employee relationship satisfaction, and work-life balance.

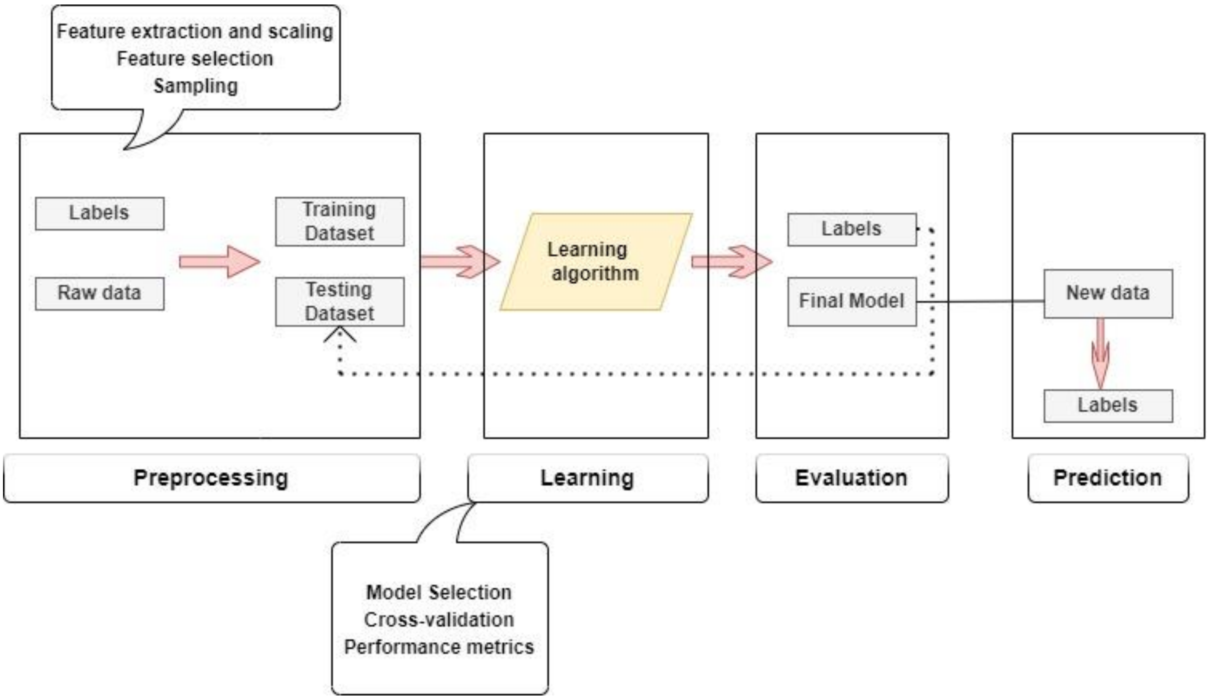

Figure 1: Research framework

Responses are divided into train and test datasets in the ratio 80:20; wherein patterns are recognized for the target variable i.e., churn. The test dataset validates the accuracy of the model. In the modeling stage, the auto-classifier node is connected to the partition node to choose the best-performing model.

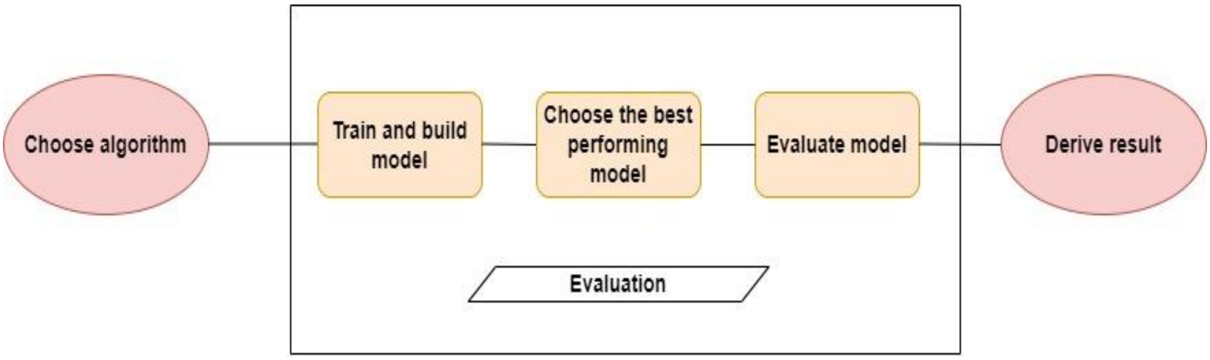

Figure 2: Methodology of the modelling stage

The quantity of training data provided improves the accuracy of ML classifier. ML algorithm aims to ensure that, whenever it is presented with a sample, the predicted outcome corresponds to the actual outcome. Each model prediction can fall into any of the four performance categories: True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN).

Table 2: Result of Data Audit node

Result analysis

The measurement scale is defined for each field using the field ops in SPSS Modeler and then further processed with a data audit node that gives the simple yet comprehensive matrix for the preparation of the data.

S.No.	Field/Feature	Minimun	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Unique	Valid
1	Age	19	59	35.65	7.98	0.452	-0.23	-	446
2	Gender							2	446
3	Marital status							3	446
4	Education	1	5					5	446
5	Education field							5	446
6	Job level	1	5					5	446
7	Job role							3	446
8	Distance from home	1	29	9.43	8.807	0.969	-0.166		446
9	Business Travel							3	446
10	Environment satisfaction	1	5					5	446
11	Job Involvement	1	5					5	446
12	Job Satisfaction	1	5					5	446
13	Relationship satisfaction	1	5					5	446
14	Work life balance	1	5					5	446
15	Monthly income	8987	37899	13224	6984.9	1.301	1.46		446
16	Percent salary hike	11	25	15.78	3.86	0.821	-0.284		446
17	Performance rating							2	446
18	Total working years	0	38	11.105	7.597	1.208	1.323		446
19	Training times last year	0	6	2.848	1.55	0.551	0.567		446

20	Year in company	0	37	7.285		1.7	3.842	446
21	Years in current role	0	17	4.487	3.66	0.777	0.178	446
22	Years since the last promotion	0	15	2.354	3.353	1.79	2.765	446
23	Worked with current manager (Years)	0	17	4.271	3.643	0.744	-0.07	446
24	Number of companies worked	0	9	3.65	3.782	1.054	0.048	446

The dataset was divided using partition node before proceeding to the modelling stage. The auto-classifier node to the partition node was connected to get the best-performing algorithm on our dataset. As a result, CHAID was the best-performing algorithm on the basis of accuracy, AUC, lift, and profit.

Table 3: Result of the CHAID algorithm

	Overall Accuracy	Area under curve	Accumulated accuracy	Accumulated AUC
Training set	89.643	0.821	89.643	0.821
Testing set	84.259	0.771	84.259	0.771

Table 4: Coincidence value report

"Partition' = 1_Training	
Range	0.334-0.939
Mean Correct	0.762
Mean Incorrect	0.518
Always Correct Above	0.813 (38.76% of cases)
Always Incorrect Below	0.373 (1.18% of cases)
90.12% Accuracy Above	0.399
2.0 Fold Correct Above	0.523 (94.22% of cases)
"Partition' =1_Training	
Range	0.321-0.951
Mean Correct	0.76
Mean Incorrect	0.57
Always Correct Above	0.869 (22.22% of cases)

Always Incorrect Below	0.394 (2.78% of cases)
90.12% Accuracy Above	0.573
2.0 Fold Correct Above	0577 (91.76% of cases)

When AUC equals 1, the algorithm can differentiate between all positive and negative class points with precision. If the AUC had been zero, the algorithm would have predicted every negative as a positive and every positive as a negative.

Table 5: Evaluation Metrics

"Partition"	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$XF_Churn	0.95	0.9	0.82	0.62

A score of 0.5 is equivalent to a random guess. A model with a score of 0.9 would be considered excellent, but a score of 0.9999 would imply overfitting. The Gini index or Gini impurity assesses the degree or probability of misclassification of a variable when randomly selected. In the above table, the value of the Gini index is 0.9 for the training dataset and 0.62 for the test data which states that the information is randomly distributed.

Table 6: Comparing $XF-Churn with Churn and Confidence matrix of $XF-Churn

"Partition"	**1_Training**		**2_Testing**		**"Partition" = 1_Training**	**No**	**Yes**
Correct	297	87.87%	90	83.33%	**No**	259	6
Wrong	41	12.13%	18	16.67%	**Yes**	35	38
Total	338		108		**"Partition" = 2_Testing**	**No**	**Yes**

As per table 6, out of 338 sample size, 297 are correct and 41 are wrong predictions. Similarly, out of 108 set, 90 are correct predictions and 18 are wrong predictions. Churn is the target variable and $XF-Churn is the predicted value. The same can be viewed by connecting the table node to the model nugget.

The accuracy is calculated on the basis of TP, TN, FP, and FN. Considering the confidence matrix for the training set, total TP, FP, TN and FN are 38, 35, 259 and 6 respectively. For the testing set, total TP, FP, TN and FN are 6, 13, 84 and 5 respectively.

CHAID employs Person's Chi-square independence tests, which examine the relationship between two categorical variables. A statistical result suggests that the two variables are independent. At each stage of developing the CHAID tree, as explained above, Chi-square tests ensures that each node is correlated with a statistically significant predictor of the dependent variable. CHAID can also be applied in situations when the response variable is continuous. Typically, a substantial sample size is required to conduct a CHAID analysis. As we divide the overall population at each branch, minimize the number of observations, and with a small total sample size, the individual groups may become too insufficient for a credible study.

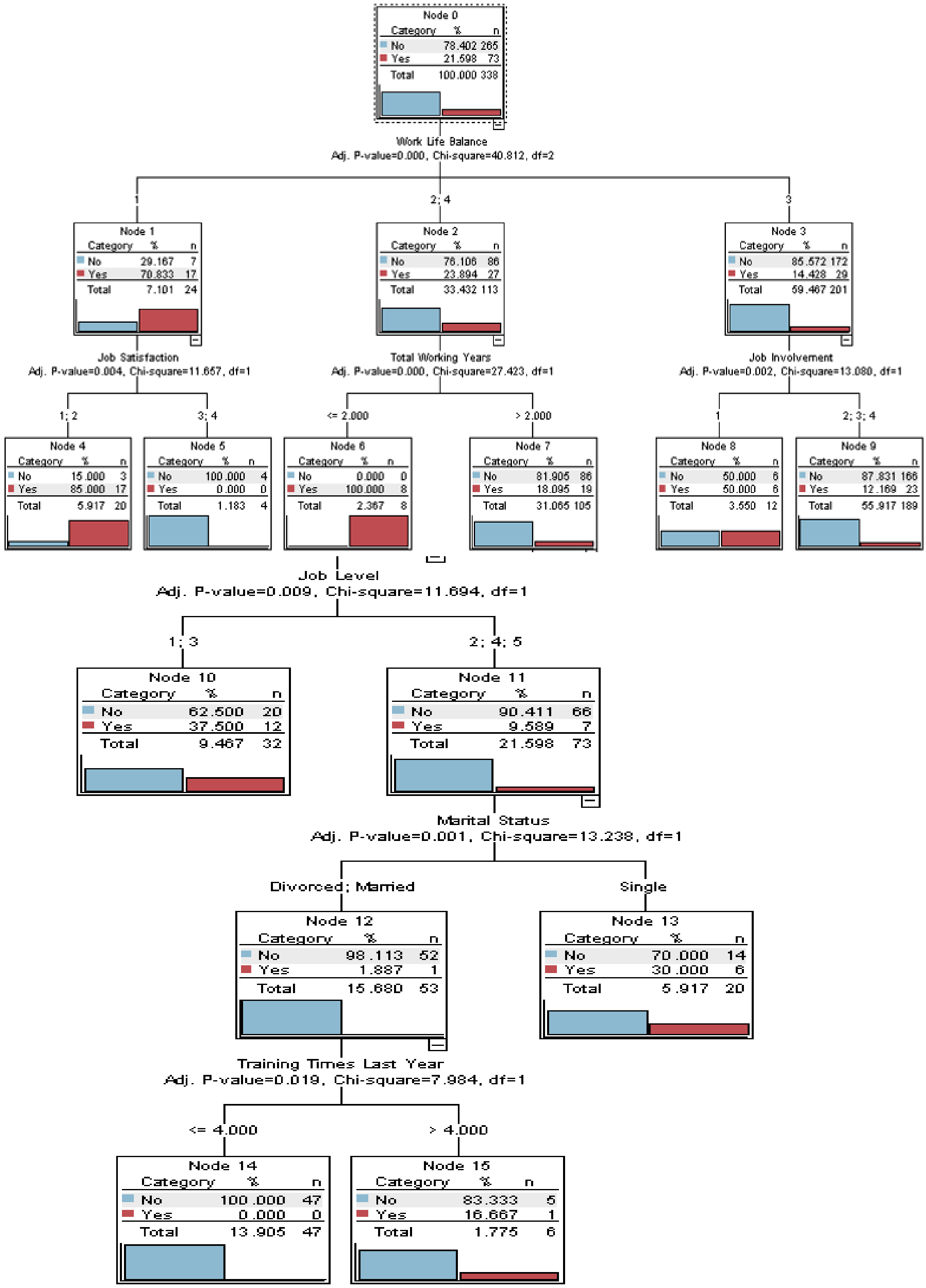

Figure 3: CHAID Tree

As shown in Figure 3, the first node 0 is the target variable 'churn' of the training dataset. The value of chi-square determines the statistical significance of the difference between expected and observed data. A low chi-square score implies that any differences between actual and anticipated data are due to random fluctuation.

Discussion

Due to the increase in e-commerce and the decline of a brick-and-mortar store, businesses require significantly fewer people on the ground than they did in the past. Only 35% of those who leave retail professions remain in the business, meaning that many former salespeople must pursue whole new careers. Typically, it costs 50 percent of that person's yearly salary to replace an entry-level employee, whereas replacing a technical or senior-level employee can cost five times of their yearly salary.

The deployment of efficient employee churn strategies is essential for businesses to survive in today's competitive marketplace, given that lowering employee churn increases productivity and profitability. When firms have a high employee churn rate, the organization's future is jeopardized since employee churn can result in decreased profits due to decreased production and knowledge. Managing ECn is crucial in the retail sector, which has a greater turnover rate than other industries. To sustain a competitive edge, an organization should prioritize lowering ECn. Therefore, corporate executives must identify the key reasons why their human resources wish to quit, and then implement preventative steps to boost their firm's overall workflow, productivity, and performance.

Data mining methodologies are useful methods for determining the employee churn rate and the factors that contribute to employee churn. It is gaining the attention of academicians and can be expounded as publications are increasing since 2018. However, it still has not gained the momentum that it should have and most publications are on online available datasets. Very few have worked on real-time datasets and almost no one has published relevant research on retail employees.

Decision-makers must understand the factors and their relative intensity in leading ECn; so that they can make proactive strategies to deal with them. This paper used CHAID to explore the factors/features for ECn. As shown in Figure 3, the first node 0 is the target variable 'churn' of the training dataset.

Conclusion

Although studies have been conducted for employee churn prediction using various ML algorithms, there is no existing literature on churn analysis of retail employees using ML. To bridge the gap in research, this paper focuses on the ECn prediction for retail employees. The questionnaire was used to collect data from employees and then processed using SPSS IBM Modeler. At modelling stage, the basis on the comparison of top-performing algorithms, CHAID was considered for detailed analysis as the AUC and accuracy was maximum for train and test data. CHAID tree demonstrates the chi-sqaure value at each node and also determines the factors/features that derive the prediction for ECn. CHAID concluded that work-life balance, job satisfaction, total working years, job involvement, and job level are the topmost predictors in descending order. Most literature is available on the IBM dataset and other datasets available online; however, research on the real-time dataset is scarce. Comparing the same with retail, the authors of this study did not find any relevant study. With the analyses of real-time data, our study focuses on retail employees. To retain talent within an organization, policymakers might utilize the aforementioned model to develop human resources strategies based on the predictive significance of characteristics.

Limitations and Future Scope of the study

This study on ECn analysis for retail employees using ML have tried to fill the research gap. However, the sample size can be increased in future studies to get more insight in this field. The authors have focused on FMCG, however, it can be widened in other fields also.

References

1. Bothma, F. C., & Roodt, G. (2012). Work-based identity and work engagement as potential antecedents of task performance and turnover intention: Unravelling a complex relationship. SA Journal of Industrial Psychology, 38(1), 27-44.
2. Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. Journal of management, 39(3), 573-603.
3. Eckardt, R., Skaggs, B. C., & Youndt, M. (2014). Turnover and knowledge loss: An examination of the differential impact of production manager and worker turnover in service and manufacturing firms. Journal of Management Studies, 51(7), 1025-1057.
4. Alao, D. A. B. A., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4(1), 17-28.
5. Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employees performance. Int. J. Adv. Comput. Sci. Appl. 3, 144–151 (2012)
6. Nagadevara, V., Srinivasan, V., Valk, R.: Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. Res. Pract. Hum. Resour. Manag. 16, 81–97 (2008)
7. Chang, H.Y.: Employee turnover: a novel prediction solution with effective feature selection. WSEAS Trans. Inf. Sci. Appl. 6, 417–426 (2009)
8. Peterson, S. L. (2007). Managerial turnover in US retail organizations. Journal of Management Development.
9. Hurley, R. F. (2015). An exploratory study of the effect of employee turnover on customer satisfaction. In Proceedings of the 1997 Academy of Marketing Science (AMS) annual conference (pp. 319-319). Springer, Cham.
10. Pandey, P., Singh, S., & Pathak, P. (2016). Devising retention strategy for front-end employees in retail: an application of analytic hierarchy process. International Journal of Services, Economics and Management, 7(2-4), 222-245.
11. Olubiyi, O., Smiley, G., Luckel, H., & Melaragno, R. (2019). A qualitative case study of employee turnover in retail business. Heliyon, 5(6), e01796.
12. McCartney, G., In, C. L. C., & Pinto, J. S. D. A. F. (2022). COVID-19 impact on hospitality retail employees' turnover intentions. International Journal of Contemporary Hospitality Management.

13. Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? the role of data quality and data diagnosticity. Decision Support Systems, 120, 38–49. https://doi.org/10.1016/j.dss.2019.03.008.
14. Jain, N., Tomar, A., & Jana, P. K. (2021). A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. Journal of Intelligent Information Systems, 56(2), 279-302.
15. Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics, 9(11), 1226.
16. Cai, X., Shang, J., Jin, Z., Liu, F., Qiang, B., Xie, W., & Zhao, L. (2020). DBGE: employee turnover prediction based on dynamic bipartite graph embedding. IEEE Access, 8, 10390-10402.
17. Chaudhary, M., Gaur, L., Jhanjhi, N. Z., Masud, M., & Aljahdali, S. (2022). Envisaging Employee Churn Using MCDM and Machine Learning. Intelligent Automation & Soft Computing, 33(2), 1009-1024.
18. Jain, P. K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees' attrition: a machine learning approach. SN Applied Sciences, 2(4), 1-11.
19. Avrahami, D., Pessach, D., Singer, G., & Ben-Gal, H. C. (2022). A human resources analytics and machine-learning examination of turnover: implications for theory and practice. International Journal of Manpower.
20. Srivastava, P. R., & Eachempati, P. (2021). Intelligent employee retention system for attrition rate analysis and churn prediction: an ensemble machine learning and multi-criteria decision-making approach. Journal of Global Information Management (JGIM), 29(6), 1-29.
21. Khera, S. N., & Divya. (2018). Predictive modelling of employee turnover in Indian IT industry using machine learning techniques. Vision, 23(1), 12-21.
22. Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Hashemkhani Zolfani, S. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics, 9(11), 1226.
23. Zhang, H., Xu, L., Cheng, X., Chao, K., & Zhao, X. (2018, September). Analysis and prediction of employee turnover characteristics based on machine learning. In 2018 18th International Symposium on Communications and Information Technologies (ISCIT) (pp. 371-376). IEEE.
24. Seelam, S. R., Kumar, K. H., Supritha, M. S., Gnaneswar, G., & Reddy, V. V. M. (2022, June). Comparative Study of Predictive Models to Estimate Employee Attrition. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1602-1607). IEEE.
25. Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee attrition estimation using random forest algorithm. Baltic Journal of Modern Computing, 9(1), 49-66.

26. Chaudhary, M., Gaur, L., & Chakrabarti, A. (2022, November). Detecting the Employee Satisfaction in Retail: A Latent Dirichlet Allocation and Machine Learning approach. In 2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM) (pp. 1-6). IEEE.
27. Chaudhary, M., Gaur, L., & Chakrabarti, A. (2022, April). Comparative Analysis of Entropy Weight Method and C5 Classifier for Predicting Employee Churn. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 232-236). IEEE.
28. Vergnolle, G., & Lahrichi, N. (2022). Data-Driven Analysis of Employee Churn in the Home Care Industry. Home Health Care Management & Practice, 10848223221137354.
29. Naz, K., Siddiqui, I. F., Koo, J., Khan, M. A., & Qureshi, N. M. F. (2022). Predictive Modeling of Employee Churn Analysis for IoT-Enabled Software Industry. Applied Sciences, 12(20), 10495.
30. Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B. (2022). Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking: An International Journal, (ahead-of-print).
31. Khan, A. A., & Aziz, A. (2008, April). Face recognition techniques (FRT) based on face ratio under controlled conditions. In 2008 International Symposium on Biometrics and Security Technologies (pp. 1-6). IEEE.
32. [32]Hamid, M. A., Hafeez, Y., Hamid, B., Humayun, M., & Jhanjhi, N. Z. (2020). Towards an effective approach for architectural knowledge management considering global software development. International Journal of Grid and Utility Computing, 11(6), 780-791.
33. Jayakumar, P.; Brohi, S.N.; Zaman, N. Top 7 lessons learned from COVID-19 pandemic. TechRxiv Prepr. 2020.
34. Khalil, M. I., Humayun, M., & Jhanjhi, N. Z. (2021). COVID-19 impact on educational system globally. Emerging technologies for battling Covid-19: Applications and innovations, 257-269.
35. Kaur, J., Ahmed, S., Kumar, Y., Alaboudi, A., Jhanjhi, N., & Ijaz, M. F. (2021). Packet optimization of software defined network using lion optimization. Cmc-Computers Materials & Continua, 69(2), 2617-2633.