Machine Learning in Human Resource Analytics: Promotion Classification using Data Balancing Techniques

Prof. Adel Ismail Al-Alawi
Management and Marketing
Department
University of Bahrain
Sakhir, Bahrain
adel.alalawi@gmail.com

Muneera Salem Albuainain
Management and Marketing
Department
University of Bahrain
Sakhir, Bahrain
m.albouainaini@gmail.com

2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) | 979-8-3503-7222-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETSIS61505.2024.10459566

***Abstract*— Nowadays, Machine Learning (ML) is widely used in many business fields, and numerous companies use it to enhance their Human Resource Management Systems (HRMS) systems. By incorporating HR analytics and ML techniques, companies can harness the power of ML throughout the employee life cycle – from recruitment to retirement. This study provides an overview of how ML applications are used in HRMS. Emphasizes the importance of addressing imbalanced data to improve system performance. To tackle this issue, Synthetic Minority Over Sampling Technique (SMOTE) and normalization techniques were employed in this research for promotion classification. The data used for analysis was derived from HR sources on Kaggle (Kaggle.com/datasets). The results obtained from this study were compared with a study that utilized Decision Tree (DT) and Principal Component Analysis (PCA) for promotion classification. The findings reveal that employing the Synthetic Minority Over Sampling Technique (SMOTE) and the Random Forest Classifier (RFC) yields results achieving an accuracy rate of 99%. Exploring models and applying them to diverse datasets would be beneficial for researchers. This study comprehensively reviews existing evidence on machine learning within HR analytics. It highlights how data processing techniques, like SMOTE and PCA, can effectively address imbalanced data while delivering accuracy.**

Keywords—Machine Learning, HR Analytics, Promotion Classification, Imbalanced Data, SMOTE, PCA

I. Introduction

Machine learning techniques are increasingly used to improve organizations' operations and decisions in today's fast-paced business world. One area where ML has become particularly influential is Human Resource Management Systems (HRMS). HR analytics play a role in enhancing aspects of employees' journey from the recruitment process through to retirement. Artificial Intelligence (AI) and Machine Learning (ML) can significantly assist recruitment efforts. By leveraging ML algorithms, HR professionals can make informed choices. Optimize their strategies for managing human resources based on data availability [1].

AI and digital innovation technologies are crucial to business continuity and organizational growth in today's rapidly involving business environment. Implementing these advanced techniques in the HR ecosystem will impact the HR functions in the organization. HR managers' roles are changing because AI affects business decision-making based on large amounts of data available for analytics. Furthermore, managers play a crucial role in creating a continuous learning environment as they can lead the employees in that direction [2]. AI is being broadly implemented in Human Resource Management (HRM) and HR Information Systems (HRIS) due to the growing digitization trend. AI has been incorporated into recruitment, employee performance and satisfaction, compensation and benefits analysis, best practice analysis, discipline management, and employee training and development [3]. A key role of Human Resources (HR) is to bridge the gap between emerging technology and human resource management. Human resource functions still need to be flexible despite technology taking over most tasks humans previously performed. Using AI algorithms, resumes and applications can be scanned to select appropriate potential candidates based on defined criteria, reducing the time and effort required for manual screening. AI can explore a candidate's data to predict if candidates are most likely to be selected, and this will improve recruitment quality as a result [4]. Artificial intelligence helps identify harmonic teams that can achieve an organization's strategy and goals. Also, adopting AI applications helps manage tasks, and training weak personnel benefits businesses in achieving their goals [5]. Machine learning will become more significant after recruitment, as it excels at analyzing and predicting employee satisfaction, employee behavior, employee performance, talent management, employee turnover, employee attrition, professional training status, and promotion classification. In this study, machine learning is mainly employed to analyze promotion classifications within HR analytics.

This study aims (1) to define key factors that influence promotion decisions, (2) to suggest an accurate classifier to assess these key factors, and (3) to provide insight into the impact of data balancing techniques on machine learning-driven promotion classification applications within Human Resource Management Systems (HRMS). Further, the study intends to (4) shed light on the impact of feature importance analysis on machine learning-driven promotion classifications.

In this study, the following research questions will be addressed:

RQ1. What are the key factors that influence promotion decisions? What are the best machine-learning techniques for capturing and utilizing these factors for accurate classification?

RQ2. How can data balancing techniques, such as Synthetic Minority Oversampling Technique (SMOTE) and normalization, be effectively applied in machine learning-based promotion classification within Human Resource

Management Systems (HRMS), and what are their effects on improving classification accuracy and resolving imbalances?

RQ3. What impact could explainable AI techniques, such as feature importance analysis, have on machine learning-based promotion classification?

RQ4. What are the long-term impacts of machine learning-based promotion classification in HRMS on organizational performance and employee satisfaction?

II. LITERATURE REVIEW

A. *Factors influencing promotion decisions*

Various factors influence the promotion decision. Some organizations focus on performance and percentage fulfillment of KPIs, while others tie it to education level, especially professional certification. Kaewwiset et al. [6] stated that Employees' capacity can be improved through training and development, which are essential to professional development. In general, professional development programs are organized according to the individual's background, goals, and work experience, as well as the organization's objectives and job requirements. A proper classification of each employee is crucial to promoting personalized training in the professional development process.

B. *ML classifiers for Promotion Classification*

Personalized training and promotion classification using AI and ML are relatively new topics since research is limited. According to Şahinbaş [7], promotions have a positive, significant, and beneficial impact on employee performance. The study used SMOTE and Random Oversampling (ROS) techniques to reduce the negative effect of imbalanced data. Random Forest Classifier (RFC) achieves the highest accuracy percentage for forecasting employee promotion with 98.42% accuracy. According to Kaewwiset et al. [6], the significance of training and development in professional development cannot be overstated. Promoting employees reasonably within an organization motivates them and contributes to business continuity. A significant extrinsic motivation for many employees is promotion.

In addition, it maintains the commitment, obligation, and loyalty to the firm while contributing to the continuity of employee performance. It is also valuable to an organization as a reward and performance control mechanism. Various factors are considered, including superiority, performance, capabilities, age, awards, training, and organizational dedication of the personnel who will be promoted.

The results of this study were compared using Decision Tree (DT), Support Vector Machine (SVM), and Random Forest Classifier (RF). Another study by Kaewwiset and Temdee [7] used DT with PCA as a dimensionality reduction technique; however, the classification accuracy of their model needed improvement. This study focuses on Machine Learning in promotion classification, although several studies studying various HR functions and themes were screened [6-18]. In most screened studies, Random Forest Classifiers (RFCs) were extensively used and were more accurate than other used classifiers. The findings are summarized in Table I.

TABLE 1. ML MODELS AND BALANCING TECHNIQUES USED IN SCREENED PUBLICATIONS

Prediction Field	***Publication***	***ML Models***	***Accuracy***	***Balanced Technique***
Employee Turnover	Sikaroudi et al. (2015)	RF	90.6%	N/A
	Gao et al. (2019)	RF	92.80%	WQRF
	Wang and Zhi (2021)	Stacker -top5-RF	99.33%	N/A
	Shanthakumara and BCS (2022)	NB LR RF	91% 91% 93%	N/A
Employee Attrition	Liu et al. (2020)	RF RF RF	83% 60% 83%	ROS SMOTE ADASYN
	Alsaadi et al. (2022)	LR DT RF	78% 97% 98%	FS
	Krishna and Sidharth (2022)	RF	99.472%	SMOTE
Talent Management	Stephanie and Sarno (2019)	KNN C4.5 SVM	87.37% 93.87% 94.62%	N/A
	Nurajijah et al. (2022)	GTB	81%	ROS FS
Personalized Training	Kaewwiset et al. (2021)	DT SVM RF	93.84% 94.03% 96.02%	SMOTE
Job Satisfaction and Employee Behavior	Gupta et al. (2023)	RF	97%	N/A
Promotion Classification	Kaewwiset and Temdee (2022)	DT	91.25%	PCA
	Şahinbaş (2022)	RF	98.42%	ROS

C. *Data Balancing Techniques*

Imbalanced data can be considered a significant bias in machine learning (ML). Models may ignore the minority class in some instances. To address this issue, various data balancing techniques have been explored. Seven of these techniques were studied in this study, and after comparison, two balancing techniques alongside different classifiers were used. The seven techniques are weighted quadratic random forest algorithm (WQRF), Random Oversampling (ROS), SMOTE (Synthetic Minority Oversampling Technique), Adaptive Synthetic Sampling (ADASYN), Feature Selection (FS), Data normalization and Principal Component Analysis (PCA).

Gao et al. [10] propose WQRF, built on the established random forest algorithm mixed with data characteristics. By determining the F-measure of each tree and proposing weighted voting, the WQRF can unravel the enigma of unbalanced data and sort the features to shrink dimensionality. As a result of applying this technique, the RF classifier achieved 92.80% accuracy. Another data balancing technique is Random Oversampling (ROS), which involves duplicating selected minority samples until their numbers match those of the majority class. While this technique can improve the model's performance on minority classes, it should be noted that it may also increase the risk of overfitting. Researchers adopted this technique in three of the screened studies [7][13][17]. Compared to ROS, SMOTE (Synthetic Minority Oversampling Technique) is a data balancing technique that oversamples random selection from a class. Then, an updated sample is created in a way that closely resembles the original. Using SMOTE reduces the likelihood of overfitting, although it may require computational resources. This technique has been used by Kaewwiset et al. [6], Liu et al. [13], and Krishna and Sidharth [15]. Also, this technique has been used in this study, and it resulted in attaining high accuracy. Liu et al. [13] also tested a Random Forest Classifier (RFC) accompanied by Adaptive Synthetic Sampling (ADASYN), which can be identified as a variation of SMOTE that tackles overfitting issues; the accuracy attained was 83%. Feature Selection (FS) technique is a commonly used balancing technique, too; this technique involves identifying and eliminating redundant features from the data. This technique helps to reduce the feature dimensions, which will result in enhancing the performance of ML models. Alsaadi et al. [14] applied FS with an RF model and got a high accuracy.

Data normalization is the second data balancing technique used in this study. This technique helps scale features within the dataset to follow a scale or range. This process helps bring consistency to feature sizes, facilitating comparisons and analysis within machine learning models. In this study, the normalization technique helps to boost the model's accuracy. The final balancing technique studied in this study is Principal Component Analysis (PCA), which provides a simple feature extraction algorithm using an unsupervised algorithm. It is used to reduce the dimensionality of the data, which solves the imbalanced data issue and enhances the model's accuracy. Kaewwiset and Temdee [8] applied the PCA with the DT classifier and got convenient results using the same data set in this study.

D. Feature Importance Analysis

Feature importance has been used in this study to understand better the factors influencing employee promotion decisions. Classifiers such as Decision Trees (DT) and Random Forests (RF) are recommended to select the most essential features. According to Alsaadi et al. [14], the DT classifier has the highest value considering the sensitivity of the measured parameters, and it is an efficient technique compared to other mining techniques.

E. Long-term impacts of ML-based promotion classification

Promotion classification will enhance HR analytics by integrating AI into the existing Human Resource Management System (HRMS) and provide decision-makers with a broad overview of the factors to be evaluated in the promotion process. ML classifiers will empower HRMS by identifying the desirable skills each employee needs to attain to be promoted. Professional development will be enforced when this knowledge is available, and many organizational problems will be solved. Kaewwiset et al. [6] state that professional development processes are nowadays an important part of organizations' recruitment processes to meet specific business objectives. As part of the training and development process, many factors can arise, particularly in organizational problems; the training and development process can be costly, time-consuming, and challenging.

III. DATASET

This study uses an open dataset downloaded from Kaggle.com in two .csv files. The training file has 54,808 records, and the testing has 23,490 with 14 features. Table II. shows the details of the features.

TABLE II. FEATURES DETAILS

#	*Feature*	*Type*	*Options*
1	employee_id	Number	N/A
2	department	Text	9 departments (Analytics, Finance, HR, Legal, Operations, Procurement, R&D, Sales and Marketing and Technology)
3	region	Text	34 different regions (branches) from region_1 to region_34
4	education	Text	3 different levels (Below Secondary, Bachelor's, and Master's above)
5	gender	Text (One Letter)	F stands for female, and M stands for male
6	recruitment_channel	Text	3 different channels (referred, sourcing and other)
7	no_of_trainings	Number	From 1 to 10 trainings
8	age	Number	From 2 to 60 years
9	previous_year_rating	Number	From 1 to 5
10	length_of_service	Number	From 1 to 37 years
11	KPIs_met >80%	Binary	1 for meeting the KPI and 0 for not meeting
12	awards_won?	Binary	1 for winning awards and 0 for not winning
13	avg_training_score	Number	Percentage (from 39% to 99%)
14	is_promoted	Binary	1 for promoting and 0 for not promoting

IV. METHODOLOGY

This paper uses a seven-step methodology, starting with data download and ending with ML model selection. (1) the process begins with downloading the open source dataset from Kaggle.com, (2) datasets are uploaded in Python and then the required libraries have been imported for data exploration and visualization, missing data has been checked to ensure that the data is clean and ready to be processed, (3) two balancing techniques were applied in this paper, normalization and Synthetic Minority Over-Sampling Technique (SMOTE) alongside four classifiers, Decision Tree (DT), Extra Trees (ET), Random Forest (RF) and eXtreme Gradient Boosting (XGB). (4) in this step, the ten most important feathers were plotted in a bar chart, which helped us to identify the key factors influencing the promotion decision. (5) the ML models were built in this phase, and then the training and testing happened in step (6), and finally, the best model was selected in step (7). Figure 1. shows the seven-process-based methodology used in this study.

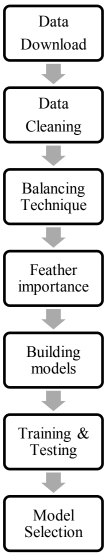


Fig. 1. Methodology

V. RESULT AND ANALYSIS

In this study, the data has been explored and analyzed using Python. The used data is imbalanced. The promoted employees class has only 8.5%, while the not promoted employees class has 91.5%. Figure 2. shows the gap between promoted and not-promoted employees. Moreover, to identify the most significant fundamental factors that impact promotion decisions, a correlation heatmap has been plotted, as shown in Figure 3. The correlation shows that the age, length of service, previous year rating, award won, and average training score have a high relation with the promotion decision. The ten most important features show that the average training score was the most critical feature, and the region was the least important feature.

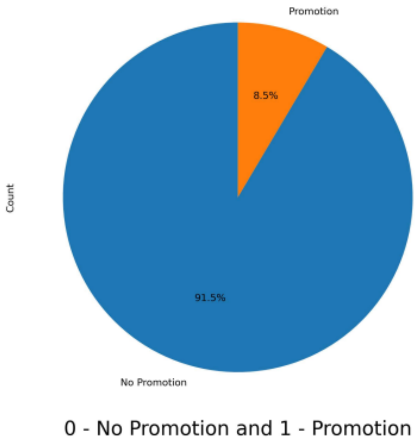
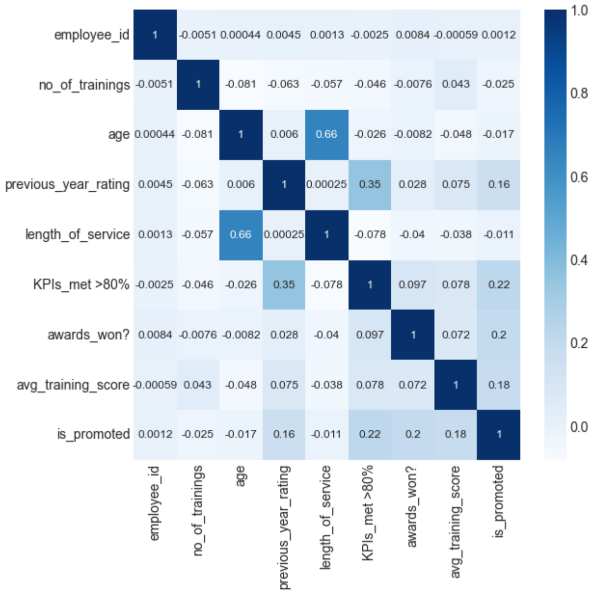


Fig. 2. Gap between promoted and not promoted employees

The over-sampling technique (SMOTE) and Normalization techniques are applied in this study to analyze promotion classification alongside four classifiers. The highest accuracy was achieved using the SMOTE.

Fig. 3. Correlation heatmap



Balancing technique and RF, DT and ET classifiers achieved 99%. Accuracy rates achieved for all ML models used in this study, accompanied by the normalization, and SMOTE techniques, are presented in Table III.

TABLE III. ML MODELS

Balancing Technique	***ML Model***	***Accuracy***
Normalization	DT	0.8833
	Extra Trees	0.9227
	RF	0.9339
	XGB	0.9424
SMOTE	XGB	0.9639
	RF	0.9998
	Extra Trees	0.9998
	DT	0.9998

The accuracy rate achieved using the DT classifier with SMOTE is higher than that achieved in the “Promotion Classification Using DecisionTree and Principal Component Analysis” study by Kaewwiset and Temdee (2022). A comparison is available in Table IV.

TABLE IV. DT Classifier

Studies	***Balancing Technique***	***Accuracy***
Kaewwiset and Temdee (2022)	PCA	91.25%
This study	SMOTE	99%

VI. Conclusion

This paper proposes using data balancing techniques to overcome the imbalanced data issue. The findings propose using the Synthetic Minority Over-Sampling Technique (SMOTE) accompanied by Random Forest (RFC) or Tree-based classifiers to achieve high accuracy. This study suggests that future researchers may test alternative models with more data balancing techniques; testing another data set would also be insightful. Moreover, it is recommended to test promotion classification datasets with Artificial Neural Networks (ANN) and more advanced ML models.

VII. ACKNOWLEDGMENTS

We would like to express our gratitude to Prof. Hatem Masri and Applied Science University in Bahrain (https://www.asu.edu.bh/) for offering free registration to the International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS24).

References

[1] A. I. Al-Alawi, M. Naureen, E. I. AlAlawi, and A. A. Naser Al-Hadad, “The Role of Artificial Intelligence in Recruitment Process Decision-Making,” 2021 International Conference on Decision Aid Sciences and Application (DASA), Dec. 2021.
DOI: 10.1109/DASA53625.2021.9682320

[2] A. I. Al-Alawi, S. K. Sanosi, and A. H. Althawadi, "Effects of technology and digital innovations on the human resources ecosystem," in 2021 International Conference on Decision Aid Sciences and Application (DASA), Dec. 2021, pp. 502-510.
DOI: 10.1109/DASA53625.2021.9682279

[3] A. M. Votto, R. Valecha, P. Najafirad, and H. R. Rao, "Artificial intelligence in tactical human resource management: A systematic literature review," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100047, 2021.
DOI: https://doi.org/10.1016/j.jjimei.2021.100047

[4] U. Murugesan, P. Subramanian, S. Srivastava, and A. Dwivedi, "A study of Artificial Intelligence impacts on Human Resource Digitalization in Industry 4.0," Decision Analytics J., vol. 100249, 2023.
DOI: https://doi.org/10.1016/j.dajour.2023.100249

[5] M. A. Ammer, Z. A. T. Ahmed, S. N. Alsubari, T. H. H. Aldhyani, and S. A. Almaaytah, “Application of Artificial Intelligence for Better Investment in Human Capital,” Mathematics, vol. 11, no. 3, p. 612, Jan. 2023.
DOI: 10.3390/math11030612

[6] T. Kaewwiset, P. Temdee, and T. Yooyativong, "Employee classification for personalized professional training using machine learning techniques and SMOTE," in Proceedings of the 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, pp. 376-379, March 2021.
DOI: 10.1109/ECTIDAMTNCON51128.2021.9425754

[7] K. Şahinbaş, "Employee Promotion Prediction by Using Machine Learning Algorithms for Imbalanced Dataset," in 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Apr. 2022, pp. 1-5 DOI: 10.1109/ICMI55296.2022.9873744

[8] T. Kaewwiset and P. Temdee, "Promotion classification using DecisionTree and principal component analysis," in 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Jan. 2022, pp. 489-492, IEEE.
DOI: 10.1109/ECTIDAMTNCON53731.2022.9720415

[9] A. M. Esmaieeli Sikaroudi, R. Ghousi, and A. 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.
DOI: 20.1001.1.17358272.2015.8.4.7.2

[10] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," Mathematical Problems in Engineering, 2019.
DOI: https://doi.org/10.1155/2019/4140707

[11] X. Wang and J. Zhi, "A machine learning-based analytical framework for employee turnover prediction," Journal of Management Analytics, vol. 8, no. 3, pp. 351-370, 2021.
DOI: https://doi.org/10.1080/23270012.2021.1961318

[12] A. H. Shanthakumara, H. T. HT, P. LV, and S. BCS, "Prediction of Employee Turnover using Machine Learning," Grenze International Journal of Engineering & Technology (GIJET), vol. 8, no. 1, 2022.

[13] L. Liu, S. Akkineni, P. Story and C. Davis, "Using HR analytics to support managerial decisions: a case study," in Proceedings of the 2020 ACM Southeast Conference, April 2020, pp. 168-175.
DOI: 10.1145/3374135.3385281

[14] E. M. T. A. Alsaadi, S. F. Khlebus, and A. Alabaichi, "Identification of human resource analytics using machine learning algorithms," *Telkomnika(Telecommunication Computing Electronics and Control)*, vol. 20, no. 5, pp. 1004-1015, 2022.
DOI: 10.12928/TELKOMNIKA.v20i5.21818

[15] K. Shobhanam and S. Sumati, “HR Analytics: Employee Attrition Analysis using Random Forest,” International Journal of Performability Engineering, vol. 18, no. 4, p. 275, 2022.
DOI: 10.23940/ijpe.22.04.p5.275281

[16] C. Stephanie and R. Sarno, "Classification Talent of Employee Using C4.5, KNN, SVM," in 2019 International Conference on Information and Communications Technology (ICOIACT), Jul. 2019, pp. 388-393.
DOI: 10.1109/ICOIACT46704.2019.8938508

[17] S. Nurajijah, D. R. Wijaya, and S. K. Sari, "Gradient Tree Boosting for HR Talent Management Application," in 2022 10th International Conference on Information and Communication Technology (ICoICT), Aug. 2022, pp. 264-269.
DOI: 10.1109/ICoICT55009.2022.9914833

[18] A. Gupta, A. Chadha, V. Tiwari, A. Varma, and V. Pereira, “Sustainable training practices: predicting job satisfaction and employee behavior using machine learning techniques,” Asian Business & Management, Jun. 2023
DOI: https://doi.org/10.1057/s41291-023-00234-5

[19] A. M. Oyelakin and R. G. Jimoh, "Towards Building an Improved Botnet Detection Model in Highly Imbalance Botnet Dataset - A Methodological Framework," Volume, vol. 3, no. 03, 2020.

[20] A. Shukla, “Feature Scaling | Normalization and Standardization in Machine Learning,” Analytics Vidhya, 24-Apr-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

[21] S. Anbalagan, “HR Analysis,” Kaggle, 2020. [Online]. Available: https://www.kaggle.com/code/shivan118/hr-analysis. [Accessed: Feb. 28, 2023].