*Journal on Artificial Intelligence*
DOI: 10.32604/jai.2022.034064
***Article***



# Ensemble Classifier-Based Features Ranking on Employee Attrition

**Yok-Yen Nguwi***

Nanyang Business School, Nanyang Technological University, 639798, Singapore
*Corresponding Author: Yok-Yen Nguwi. Email: yokyen@ntu.edu.sg
Received: 05 July 2022; Accepted: 06 August 2022

**Abstract:** The departure of good employee incurs direct and indirect cost and impacts for an organization. The direct cost arises from hiring to training of the relevant employee. The replacement time and lost productivity affect the running of business processes. This work presents the use of ensemble classifier to identify important attributes that affects attrition significantly. The data consists of attributes related to job function, education level, satisfaction towards work and working relationship, compensation, and frequency of business travel. Both bagging and boosting classifiers were used for testing. The results show that the selected features (nine selected features) achieve the same result as the full features. The selected features are age, income, working years, source of employment, years since last promotion, salary hike, and business travelling frequency. These features were selected using ensemble classifiers. Satisfaction on work and relationship do not appear to be significant attributes in attrition from ensemble classifier's results.

**Keywords:** Machine learning; ensemble classifier; attrition; prediction

## 1 Introduction

Technology plays dual role as both an enabler and disruptor in this era of digitalization. It enables organization to be productive by moving some repetitive tasks from employees to technology platform. In the meantime, it is seen as disruptive by some as employees embrace the changes in the way they work. Human Resource (HR) technologies encompass a wide range of systems to support HR functions like talent acquisition, onboarding, training, attendance tracking and monitoring, payroll and leave processing. Most of which are human-in-the-loop where human involves a large part of decision making and connection to other systems for further processing. One recent trend is on the use of Artificial Intelligence to assist HR functions in reducing load involving large amount of data.

Artificial Intelligence refers to system or machine that can perform tasks which demonstrates human like intelligence. These tasks include decision making, language ability, recognition of objects, and problem-solving ability. There are 2 broad categories of AI: the general AI and the narrow AI. General AI artifacts exhibits most characteristics of human intelligence while narrow AI displays one or a few identifiable characteristics of human intelligence in machine. Machine learning is one way of developing Artificial Intelligence product, but it is not the only way. Machine learning is

a way developing training algorithm for machine to learn from data. It differs from conventional programming where the logic was embedded using standard routine instructions. Machine learning focuses on programming how the machine should respond when given a set of input and output data. The idea is to enable the machine to slowly makes more intelligent decision. Artificial Intelligence includes the system that learns and the product that executes the learning ability. Both are often used interchangeably. AI is often used to encapsulate the related technologies that automate business processes to reduce mundane repetitive tasks. Machine learning is often used to refer to the techniques used to develop AI products.

Narrow AI exists for a wide range of specific problems. Specialized AI system is trained to exhibit human intelligence on defined tasks. AI is seen as an indispensable part of "Future of Work" [1,2]. Balliester et al. [3] presented a literature review and defined future of work with five dimensions on the outcomes of Future of Work which will impact the way we work: job quality, social protection, wages inequalities, social dialogue and industrial relations, and number of jobs. A proactive approach [4–7] to align capability of AI presents is at the centre of discussion that will bring greater fruition. In all existing frameworks, decision making workflow is designed by human for human in the loop decision making. AI system should be designed to augment and not replacing human contribution [5].

Human Resource Management (HRM) is mapped to supporting business activity from Porters Value Chain [8] where competitive advantage of a firm can be achieved through excelling in its business primary or supporting activities. The HRM manages the workforce with the objective of achieving the desired working environment that drives business efficiency. It encompasses the whole lifecycle of employment from recruitment process, planning for suitable training for current employees, to the departure of staffs. The emphasis of driving business objective in the early days had led to adversarial relationship between employers and labours [9]. Taylor [10] thus advocated the use of scientific management to resolve the conflict between managers and workers systematically. Modern HRM perceives employees as assets to an organization where the right management can bring success to a business. Some recent works attempt to use machine learning approach for different aspects of HRM. Hila et al. [11] examined the turnover issues using machine learning approach. Pessach et al. [12] presented a work to predict successful placement of a candidate in a specific position at pre-hiring stage. Garg et al. [13] presented a review of machine learning applications in human resource management. There is a lack of works related to selecting useful features for machine learning to perform prediction. This work is targeted towards finding the useful features prior to classification tasks.

The paper is organized as follows: Section 2 describes the datasets containing employees' profiles, compensation, and their work-related satisfaction level. The association between variables are tabulated to look at the relationship between variables. Section 3 proposes the use of ensemble classifiers to select top features to predict if an employee is likely to resign. Both bagging and boosting ensemble classifiers were used. Finally, conclusion of this paper is drawn in Section 4.

## 2 Datasets

This section presents the background information about the dataset. The dataset gathered information on employees' personal information, employment satisfaction, work performance, and employment status in a company. This dataset consists of 8687 records (N = 8687) and 30 features. The data contains 139 missing data in different features, those are removed as part of the data cleaning process. The clean-up data has 8596 records (N = 8596) in total. Employee status is kept in column "Attrition", the status can be current employee (N = 7318), resigned (N = 1333), or terminated

(N = 32). There are insufficient data to provide details on the number of employees at the start and end of period, we consider all current employees in the computation of attrition rate. The voluntary attrition rate is currently at 18.2% while involuntary attrition rate is 0.4%. Due to the high voluntary attrition rate, this work focuses on finding out the reasons behind voluntary resignation using machine learning approach.

The feature and the type of feature from this dataset are listed in Table 1. Among the 30 features, 12 are numeric variables, 9 ordinal variables, and 9 nominal variables. The data is arranged into two groups: one group who resigned voluntarily and another group of current employees. We present the relationship between relevant features and employment status from Figs. 1–3. Fig. 1 denotes the difference in proportion for resignation group and current employee group on business travel requirements, gender, work overtime, and the departments they belong to. Occasional travel (travel_trarely) makes up a large proportion for current employees' group as opposed to resigned group. The proportion for gender is similar for both groups. From the third graph of Fig. 1, we can see a larger proportion of current employee group does not work overtime. The last graph shows more headcounts for current employee group in Research & Development department.

**Table 1:** Features and type of features

| Feature | Type of feature |
|---|---|
| Age | Numeric |
| Attrition | Nominal {Voluntary Resignation, Current employee, Termination} |
| BusinessTravel | Nominal {Travel_Rarely, Travel_Frequently, Non-Travel} |
| Department | Nominal {Sales, Human Resources, Research & Development} |
| DistanceFromHome | Numeric |
| Education | Nominal {1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'} |
| EducationField | Nominal {Life Sciences, Human Resources, Marketing, Other, Technical Degree, Medical} |
| EnvironmentSatisfaction | Ordinal {1 'Low', 2 'Medium', 3 'High', 4 'Very High'} |
| Gender | Nominal {Female, Male} |
| JobInvolvement | Ordinal {1 'Low', 2 'Medium', 3 'High', 4 'Very High'} |
| JobLevel | Ordinal {1, 2, 3, 4, 5} |
| JobRole | Nominal {Sales Executive, Manager, Human Resources, Research Scientist, Manufacturing Director, Laboratory Technician, Healthcare Representative, Sales Representative, Research Director} |
| JobSatisfaction | Ordinal {1 'Low', 2 'Medium', 3 'High', 4 'Very High'} |
| MaritalStatus | Nominal {Single, Divorced, Married} |
| MonthlyIncome | Numeric |
| NumCompaniesWorked | Numeric |
| OverTime | Nominal {Yes, No} |
| PercentSalaryHike | Numeric |
| PerformanceRating | Ordinal {1 'Low', 2 'Medium', 3 'High', 4 'Very High'} |

(Continued)

**Table 1:** Continued

| Feature | Type of feature |
|---|---|
| RelationshipSatisfaction | Ordinal {1 'Low', 2 'Medium', 3 'High', 4 'Very High'} |
| StandardHours | Numeric |
| StockOptionLevel | Ordinal {0, 1, 2, 3} |
| TotalWorkingYears | Numeric |
| TrainingTimesLastYear | Numeric |
| WorkLifeBalance | Ordinal {1 'Bad', 2 'Good', 3 'Better', 4 'Best'} |
| YearsAtCompany | Numeric |
| YearsInCurrentRole | Numeric |
| YearsSinceLastPromotion | Numeric |
| YearsWithCurrManager | Numeric |
| Employee Source | Ordinal {Referral,'Company Website', Indeed, Seek, Adzuna, Recruit.net, GlassDoor, Jora, LinkedIn} |

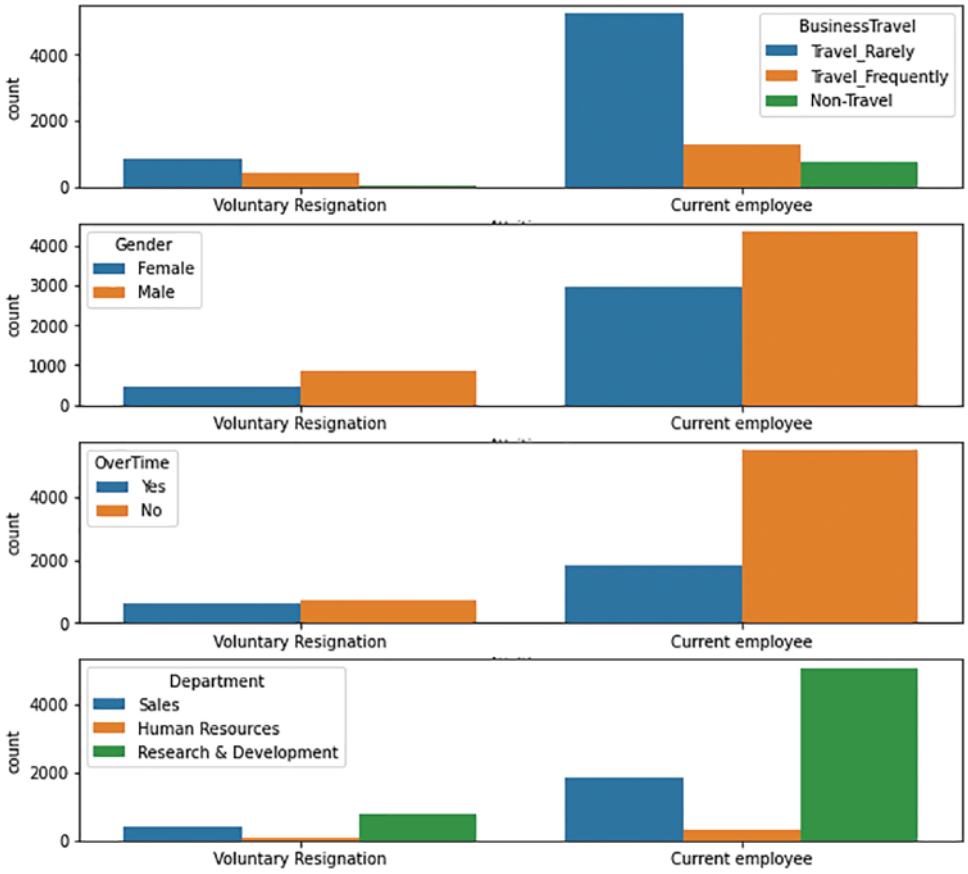


**Figure 1:** Nominal features *vs.* attrition

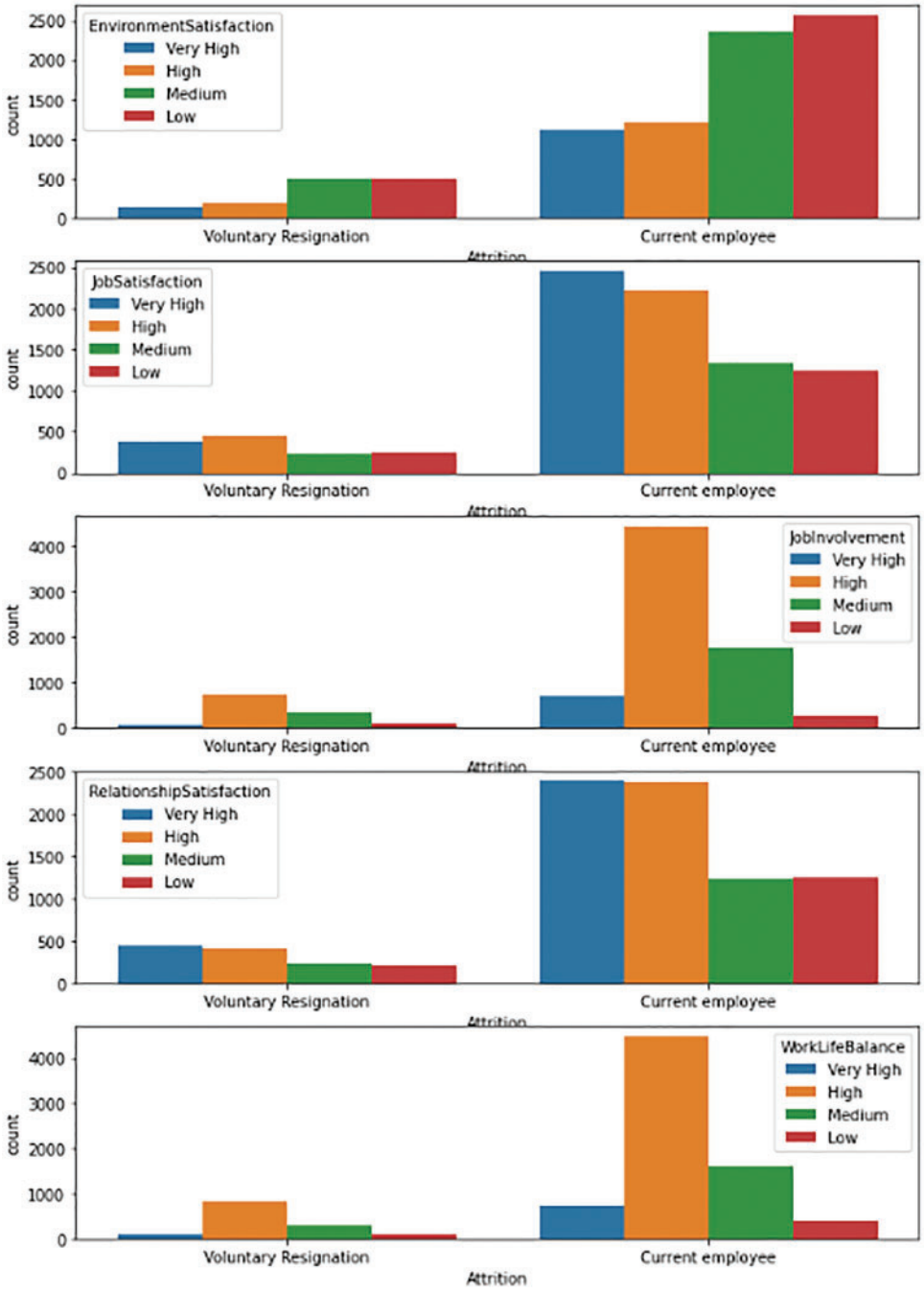


**Figure 2:** Employee's satisfaction level *vs.* attrition

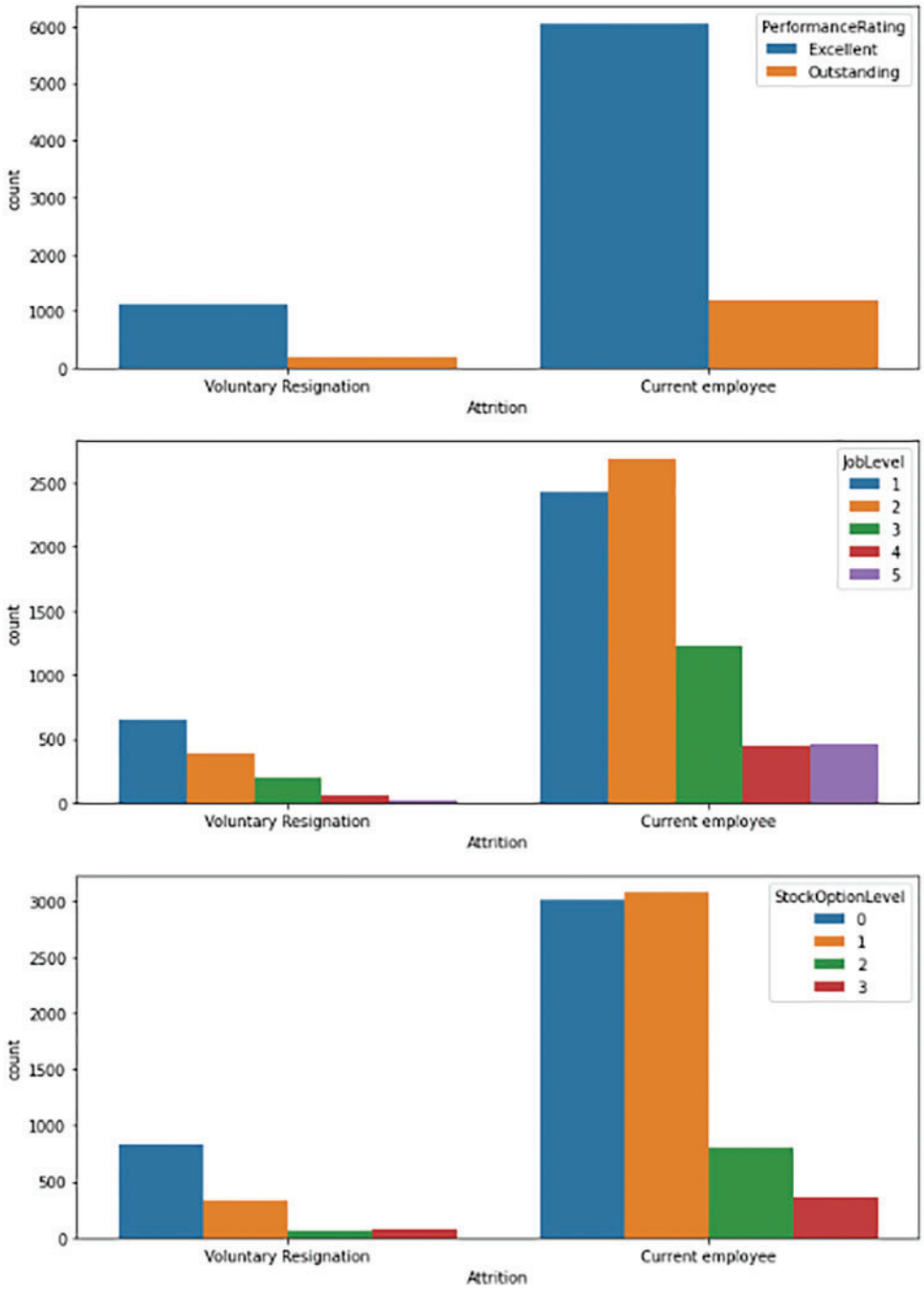


**Figure 3:** Employee's performance and rewards *vs.* attrition

Fig. 2 shows employee's satisfaction level is rated in four levels (Low, Medium, High, Very High) on surrounding factor like environment satisfaction, on job factors like job satisfaction and job involvement, on work relationship satisfaction and work life balance. The first plot from the top of Fig. 2 shows that most current and resigned employees are having low to medium environmental satisfaction, especially for resigned employees. On the other hand, job's satisfaction level is higher

with current employees showing mostly very high job satisfaction. This could illustrate that employees hired are matching the job roles that results in high job satisfaction. The next figure on job involvement further strengthens this point with high job involvement rating in both current and resigned groups. In terms of working relationship, current employees reported higher proportion of very high and high ratings. Work life balance is high for both groups as well.

Fig. 3 visualizes the performance and rewards for both resigned and current employees. The performance rating currently only comes with two ratings: Excellent and Outstanding. It is not a good gauge to understand the performance of employee's base on this. The second plot on job level shows a disproportionate level for current employees with more junior level staffs than mid to high level employees. We can also see people from different job grade leaving the position. The last plot on stock option level denotes higher proportion of resigned group having the most basic level of stock option (level 0). This could likely to hint to insufficient rewards being given to resigned group. The current employee group has higher proportion of mid-range reward of stock option level 1.

Spearman's rank correlation coefficient can be used to find out the rank correlation between ranked values. The correlation denotes monotonic relationship as opposed to linear relationship in Pearson's correlation. The coefficient can range from −1 to +1 with positive value denotes a positive monotonic relationship when one variable increases, the other also increases monotonically. Negative value denotes negative monotonic relationship when one variable increases, the other decreases monotonically. The ranked correlation can be computed using the following formula [14]:

$$r_s = 1 - \frac{6 \sum d_i^2}{n\,(n^2 - 1)}$$

where $d_i$ denotes the difference between two ranks or each observation, and $n$ is the number of observations. The ranked correlation is computed in Fig. 4, all do not appear to have significant monotonic relationship.

Cramér's V association [15] is another measure to examine the association between ranked variables. The range of value for this association is from 0 to 1 where 0 denotes no association and 1 being perfect association. The strength of association has different threshold as the usual correlation. The strength of association is also called effect size. The association of less than 0.2 is considered weak association, between 0.2 to 0.6 is considered moderate association, and above 0.6 is considered strong association. The Cramér's V association [16] can be computed from chi-square statistic.

$$V = \sqrt{\frac{\varphi^2}{\min\,(k-1,\ r-1)}} = \sqrt{\frac{X^2}{\min\,(k-1,\ r-1)}}$$

where $\varphi^2$ denotes the phi coefficient, $X^2$ is derived from Pearson's chi-squared test, $n$ is the total observation, $k$ being the number of columns and $r$ is the number of rows.

Fig. 4 illustrates the matrix generated based on Spearman's correlation and Cramér's V's computation, all the ranked variables appear to have low association with attrition. Among the variables, StockOptionLevel has the highest relative association, which shows reward system makes a difference in attrition outcome.

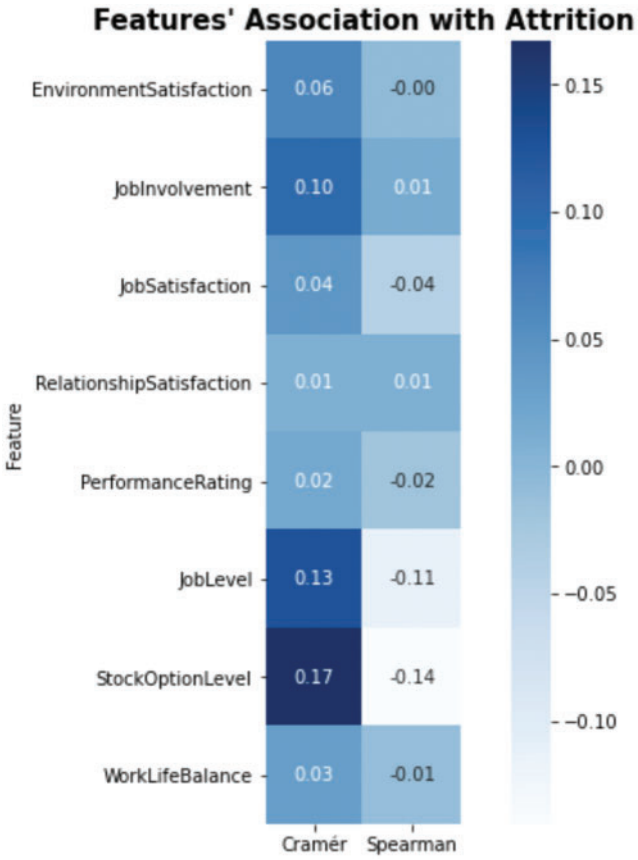


**Figure 4:** Features association with attrition

## 3 Experimental Results

This section examines the attrition data using machine learning approach. We look at the use of ensemble approach to find out the top 5 features from this dataset. Ensemble classifier combines multiple classifiers to stratify the samples or sequence the classifiers to refine classification outcome. It has achieved good success in classification tasks [13,17,18] in Human Resource Management. There are broadly 3 types of ensemble classifiers: bagging, boosting and stacking. Different base classifiers can be used in ensemble process. Bagging classifier works by stratifying the samples and perform classification on different groups of samples, a few classifiers are then set up to classify the groups of samples. The final result aggregates the performances of these classifiers. Boosting classifier works by combining weak classifiers, iteratively boost the performance of weak classifiers. The errors from earlier iterations will be focused during the refinement process, a strong classifier is eventually trained. Stacking ensemble classifier works by combining heterogenous classifiers for improved performance.

We perform features ranking based on Random Forest (bagging), Adaboost (boosting) and XGBoost (boosting). All the three ensemble rankers use decision tree as base classifier and the importance of features is explained by Gini index. Features ranking computes the scores according to its importance to classification results. The features scores are computed based on entropy of information gain. The feature ranker results are shown in Fig. 5 that shows the top 10 most important

features from the 3 rankers. Feature ranker one is Random Forest, ranker 2 is AdaBoost and ranker 3 is XGBoost. The most important feature is age as highlighted in yellow, that occurs consistently across the three feature rankers. Age represents the most important feature to predict if an employee is leaving a company. Other important features include monthly income, total working years, years at the company, percentage of salary hike, the source of hiring (Employee Source), years since last promotion, how often is the business travel, and training time in previous year.

We conducted classification using two sets of features: first with the full features, second with the selected features from the above. Table 2 shows the results of the three ensemble classifiers using full features (total 29 features), the lowest accuracy is AdaBoost with 86% accuracy while Random Forest and XGBoost both obtain 99% accuracy. Table 3 shows the results of the same ensemble classifiers with selected nine features (age, monthly income, total working years, years at the company, percentage of salary hike, the source of hiring (Employee Source), years since last promotion, how often is the business travel, and training time in previous year). Similar results are achieved despite having much lesser features: 84% for Adaboost, 99% for both Random Forest and XGBoost. There is not overfitting issue observed from the experimental results. These show that the selected features might be useful for employers to consider retaining employees in the organization, namely the compensation, the training being provided, lesser business travel, and years at the company.

| Feature Ranker 1 | Feature Ranker 2 | Feature Ranker 3 |
|---|---|---|
| Age | Age | StockOptionLevel |
| DistanceFromHome | MonthlyIncome | OverTime |
| MonthlyIncome | YearsInCurrentRole | JobLevel |
| TotalWorkingYears | BusinessTravel | TotalWorkingYears |
| YearsAtCompany | TrainingTimesLastYear | Age |
| PercentSalaryHike | YearsWithCurrManager | TrainingTimesLastYear |
| Employee Source | YearsSinceLastPromotion | BusinessTravel |
| EnvironmentSatisfaction | YearsAtCompany | JobSatisfaction |
| Education | PercentSalaryHike | JobRole |
| YearsSinceLastPromotion | OverTime | Employee Source |
| | | |
| | Voted by 3 Rankers | |
| | Voted by 2 Rankers | |

**Figure 5:** Feature ranking results

**Table 2:** Classifier results with original features

| Classifier | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Random forest | 98 | 99 | 100 | 99 |
| AdaBoost | 58 | 78 | 59 | 86 |
| XGBoost | 98 | 99 | 99 | 99 |

**Table 3:** Classifier results with selected features

| Classifier | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Random forest | 98 | 98 | 98 | 99 |
| AdaBoost | 51 | 59 | 49 | 84 |
| XGBoost | 98 | 98 | 98 | 99 |

## 4 Conclusion

This work presented the use of ensemble classifiers on tackling attrition issue in the domain of human resource management. The work contributes to adding in the AI lens to attrition problem on picking up important features that can help to predict the departure of employees. This work adopted the use of bagging and boosting ensemble classifiers to rank the features. Results show that the use of selected features are able to predict as accurately as the use of full features. The important features are age, monthly income, total working years, years at the company, percentage of salary hike, the source of hiring (Employee Source), years since last promotion, how often is the business travel, and training time in previous year.

**Funding Statement:** The author received no specific funding for this study.

**Conflicts of Interest:** The author declares that he has no conflicts of interest to report regarding the present study.

## References

[1] T. W. Malone, "How human-computer 'Superminds' are redefining the future of work," *MIT Sloan Management Review*, vol. 59, no. 4, pp. 34–41, 2018.
[2] P. Moradi and K. Levy, "The future of work in the Age of AI," *The Oxford Handbook of Ethics of AI, 2020*, pp. 271, 2020.
[3] T. Balliester and A. Elsheikhi, "The future of work: A literature review," *ILO Research Department Working Paper No. 29*, pp. 1–62, 2018.
[4] T. A. Kochan, "Artificial intelligence and the future of work: A proactive strategy," *AI Magazine*, vol. 42, no. 1, pp. 16–24, 2021.
[5] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons*, vol. 61, no. 4, pp. 577–586, 2018.
[6] W. Wang and K. Siau, "Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda," *Journal of Database Management (JDM)*, vol. 30, no. 1, pp. 61–79, 2019.
[7] J. Howard, "Artificial intelligence: Implications for the future of work," *American Journal of Industrial Medicine*, vol. 62, no. 11, pp. 917–926, 2019.
[8] M. E. Porter, "The value chain and competitive advantage," *Understanding Business Processes*, vol. 2, pp. 50–66, 2001.
[9] R. S. *History of human resource management,* 2022. [Online]. Available: https://www.business-managementideas.com/human-resource-management-2/history-of-human-resource-management/19925.
[10] F. W. Taylor, *Scientific Management*. Routledge, United Kingdom, 2004.

[11] D. Avrahami, D. Pessach, G. Singer and H. C. Ben-Gal, "A human resources analytics and machine-learning examination of turnover: Implications for theory and practice," *International Journal of Manpower*, vol. 43, no. 6, pp. 1405–1424, 2022.
[12] D. Pessach, G. Singer, D. Avrahami, B. G. Hila, E. Shmueli *et al.,* "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decision Support Systems*, vol. 134, pp. 113290, 2020.
[13] S. Garg, S. Sinha, A. K. Kar and M. Mani, "A review of machine learning applications in human resource management," *International Journal of Productivity and Performance Management*, vol. 71, no. 5, pp. 1590–1610, 2022.
[14] Wikipedia, *Spearman's Rank Correlation Coefficient*, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.
[15] H. Cramér, "A contribution to the theory of statistical estimation," *Scandinavian Actuarial Journal*, vol. 1, pp. 85–94, 1946.
[16] Wikipedia, *Cramér's V*, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Cramér%27s_V.
[17] M. Sharma, D. Singh, M. Tyagi, A. Saini, N. Dhiman *et al.,* "Employee retention and attrition analysis: A novel approach on attrition prediction using fuzzy inference and ensemble machine learning," *Webology*, vol. 19, no. 2, pp. 5338–5358, 2022.
[18] T. Tarusov and O. Mitrofanova, "Risk assessment in human resource management using predictive staff turnover analysis," in *Int. Conf. on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*, Lipetsk, Russia, pp. 194–198, 2019.