Journal of Intelligent Information Systems (2021) 56:279–302
https://doi.org/10.1007/s10844-020-00614-9

A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning

Nishant Jain[1] · **Abhinav Tomar**[1] · **Prasanta K. Jana**[1]

Received: 28 December 2019 / Revised: 27 July 2020 / Accepted: 30 July 2020 /
Published online: 29 September 2020

Abstract

Employee churn (ECn) is a crucial problem for any organization that adversely affects its overall revenue and brand image. Many machine learning (ML) based systems have been developed to solve the ECn problem. However, they miss out on some essential issues such as employee categorization, category-wise churn prediction, and retention policy for effectively addressing the ECn problem. By considering all these issues, we propose, in this paper, a multi-attribute decision making (MADM) based scheme coupled with ML algorithms. The proposed scheme is referred as employee churn prediction and retention (ECPR). We first design an accomplishment-based employee importance model (AEIM) that utilizes a two-stage MADM approach for grouping the employees in various categories. Preliminarily, we formulate an improved version of the entropy weight method (IEWM) for assigning relative weights to the employee accomplishments. Then, we utilize the technique for order preference by similarity to ideal solution (TOPSIS) for quantifying the importance of the employees to perform their class-based categorization. The CatBoost algorithm is then applied for predicting class-wise employee churn. Finally, we propose a retention policy based on the prediction results and ranking of the features. The proposed ECPR scheme is tested on a benchmark dataset of the human resource information system (HRIS), and the results are compared with other ML algorithms using various performance metrics. We show that the system using the CatBoost algorithm outperforms other ML algorithms.

Keywords Employee churn · Employee importance model · Retention policy · CatBoost algorithm · MADM method · TOPSIS

✉ Nishant Jain
nishantjain.iit@gmail.com

Abhinav Tomar
profession.abhinav@gmail.com

[1] Department of Computer Science & Engineering, Indian Institute of Technology (ISM), Dhanbad 826004, India

1 Introduction

Employees are valuable assets for any organization. However, when some employees leave an organization, its productivity, project continuity, and growth strategies are severely affected on which the image and turnover of the organization depend. This also makes other employees quitting the organization (Rashid and Jabar 2016). This, in turn, causes large expenditure on hiring new employees, which is a time-consuming and expensive task. The literature (Brown and Wilson 2007) refers to this issue as employee churn (ECn) problem, which draws significant attention for an efficient solution to remain productive and competitive in today's world. In this context, designing an effective scheme for employee churn prediction and retention policy is an important application of machine learning (ML). Over the past few years, ML algorithms have been booming, and they are successfully applied in various domain such as recommender system (Tarnowska et al. 2020), personal life event prediction (Khodabakhsh et al. 2018), stock price prediction (Xiao et al. 2019), and so on. The motivation for adopting ML algorithms for the ECn problem is threefold. Firstly, the organization does not have sufficient resources to predict employee churn manually. Secondly, the data for ECn prediction is now available in considerable quantity, which should be appropriately utilized to make a necessary decision (Ghasemaghaei and Calic 2019). Thirdly, the available dataset is not erratic, i.e., the dataset is being continuously updated.

In past years, quite a few ML-based solutions have been proposed for ECn and its related problems. For example, Saradhi and Palshikar (2011) described a customer churn problem and presented a case study for building and comparing predictive employee churn models. Tarnowska and Ras (2018) discussed the solution for customer attrition problem based on actionable knowledge. Fan et al. (2012) proposed a scheme based on hybrid clustering analysis for predicting trends in employee turnover for technology professionals. Punnoose and Ajit (2016) proposed a method for turnover prediction, which is based on eXtreme Gradient Boosting (XGBoost). Similarly, Gao et al. (2019) presented a weighted quadratic random forest algorithm for employee turnover prediction. However, all of these studies overlook the categorization of the employees into different classes such as those who contribute or do not contribute significantly to the turnover of the organizations. Moreover, they have not considered any retention policy according to different categories of the employees for effectively addressing the ECn problem. These facts underpin the need for further research.

It is well noted that the importance of the employees in an organization can be judged based on multiple criteria, and thus it is a very challenging task to categorize them accordingly. In recent years, multi-attribute decision making (MADM) has shown great potential to deal with such problems. The MADM methods have been successfully applied to various real-world problems such as healthcare system (Mendoza-Gómez et al. 2019), marketing and business management, human resources selection task (Krylovas et al. 2017), and other areas, (Zhou et al. 2018), (Tomar and Jana 2018). Among numerous MADM methods proposed so far, the technique for order preference by similarity to ideal solution (TOPSIS) (Hwang and Yoon 1981) has gained enormous popularity due to its multi-fold advantages. It is a simple and comprehensible concept and has good computational efficiency, and the ability to measure the relative performance for each alternative in a simple mathematical form (Yeh 2002). TOPSIS evaluates the performance of alternatives through the similarity with the ideal solutions. Moreover, the one that is nearest to the positive-ideal solution and farthest from the negative-ideal solution would be the best alternative. In TOPSIS, it is necessary to determine the relative weights of the features for which several methods have been proposed. For example, Chu et al. (1979) used the weighted least square method to determine the weights for the fuzzy set. Hamed Fazlollahtabar (2010) used

the analytic hierarchy process (AHP) (Saaty 1987) to ranking the automobile seat comfort based on consumer preferences. Entropy weight method (EWM) (Wang and Lee 2009) is frequently used in the literature as it is prone to information entropy or attribute data heterogeneity. Therefore, it is ideal for distinguishing alternatives. In contrast to other weighting methods (Fazlollahtabar 2010; Saaty 1987), the EWM also decreases the effect of false or artificial attributes' information, subjectively used by the users.

Motivated with the aforementioned evidence, we propose, in this paper, an MADM and ML-based scheme for the employee churn problem and retention policy. We refer this scheme as employee churn prediction and retention (ECPR). The ECPR is based on TOPSIS that integrates EWM as a weight estimation method. At first, we formulate a novel accomplishment-based employee importance model (AEIM) in order to group the employees into three categories based on the accomplishment parameters. The AEIM integrates an improved version of EWM with TOPSIS for quantifying the importance of employees. We then predict the class-wise employee churn using the CatBoost algorithm and compare the results with state-of-the-art ML algorithms. Finally, a class-wise employee retention strategy is provided using a permutation-based feature importance method based on prediction results.

To the best of our knowledge, we are the first to design an MADM based scheme for the ECn problem that jointly addresses the following issues: 1) It identifies valuable employees by categorizing them, 2) It predicts category-wise employee attrition (aka churn) based on ensemble methods, and 3) It highlights the causing factors of attrition so as to strategize better retention plans.

The rest of this paper is organized as follows. Section 2 reviews the previous researches, emphasizing the employee churn problem. The working process of the proposed ECPR scheme is explained in Section 3. Section 4 presents the research evaluation methodology along with performance metrics. The experimental results and findings are summarized in Section 5. Finally, Section 6 concludes this paper.

2 Related works

The existing literature on the ECn problem has gained significant attention from researchers. Several pioneering research works have been carried out on ECn prediction in human resource analysis. The volume of the ECn research based on the theoretical hypothesis has increased dramatically since the last century due to the dynamic and competitive market policy (Hom et al. 2017). Many researchers found the vital relationship between the satisfaction level of the employees and their continuation with the organizations (Harter et al. 2002). Employees with excellent job satisfaction levels present fewer rates of absenteeism, have significant contributions to the organization, and are extremely amenable to continue serving the organization (Morrow et al. 1999). Job satisfaction is influenced by personal characteristics like perception, cognitive ability, demographic variables, expectations, sense of achievement as well as environmental factors (Frederiksen 2017). These studies primarily focus on the correlation between job satisfaction and employee churn. However, these schemes provided the solution for the ECn problem on the ground of theoretical assumption.

Recently, the machine learning algorithms aspire to extract beneficial information and hidden knowledge from the past available dataset. Some of the endeavours to predict ECn involved the use of machine learning and data mining techniques (Sikaroudi et al. 2015). To date, there has been rising attention to ECn prediction using ML in tourism, health care,

trade, and industrial systems worldwide. Here, we review only relevant research works that use ML algorithms to solve the ECn problem.

Ren et al. (2011) described the ECn problem for the travel agencies and analyzed the solution in terms of the strategic model. Sexton et al. (2005) proposed employee turnover prediction based on neural network. However, they still have the scope of improvement concerning the retention policy and the prediction accuracy of employee attrition. Chien and Chen (2008) proposed the system to improve the retention rate based on effective personnel selection. For this, they use association rule along with a decision tree to set up needful rules to select a human resource. However, they have not considered employee categorization and category-wise retention strategy.

Although a very few research studies focus on retaining valuable employees (Budhwar et al. 2006), it is not beneficial for the long-term smooth functioning of the organizations. The reason is that the cost of recruiting new employees is far more as compared to keeping the existing employees. Our proposed scheme is inspired to make a retention policy for each class of employees.

It is clear from Table 1 that most of the works have been carried out on the prediction only, and a few works have been done on the retention. Besides, none of the existing schemes has considered the employee categorization based on their importance, while it is an essential component for efficiently solving the ECn problem. We consider all the aforementioned issues together for designing a complete scheme for effectively addressing the ECn problem with due emphasis on employee categorization based on their importance.

Table 1 Comparative summary of relevant research works that use ML algorithms to address the ECn problem

Literature	Dataset categorization	Prediction model	Retention model	ML model(s)	Application area
Gao et al. (2019)	No	Yes	No	Improved RF	Generalized
Sexton et al. (2005)	No	Yes	Yes	NN and MGA	Generalized
Chien and Chen (2008)	No	No	Yes	DT and AR	Specialized
Fan et al. (2012)	No	Yes	Yes	ANN and SOM	Specialized
Koh and Goh (1995)	No	Yes	No	LR, ANN, and RF	Generalized
Punnoose and Ajit (2016)	No	Yes	No	XGBoost	Specialized
Dolatabadi and Keynia (2017)	No	Yes	No	ANN and SVM	Generalized
Yigit and Shourabizadeh (2017)	No	Yes	No	DT, LR, SVM, KNN, RF, and NBC	Generalized
Frierson and Si (2018)	No	Yes	No	LR and KMSF	Specialized
Proposed	Yes	Yes	Yes	LR, SVM, DT, XGBoost, RF, and CatBoost	Generalized

Abbreviations:

DT → Decision Tree; *RF* → Random Forest; *NN* → Neural Network;
AR → Association Rule; *LR* → Logistic Regression; *SOM* → Self-Organizing Map;
CatBoost → Categorical Boosting; *MGA* → Modified Genetic Algorithm; *SVM* → Support Vector Machine;
ANN → Artificial Neural Network; *KNN* → k-Nearest Neighbors; *NBC* → Naive Bayesian Classification;
KMSF → Kaplan Meier Survival Function; *XGBoost* → eXtreme Gradient Boosting;

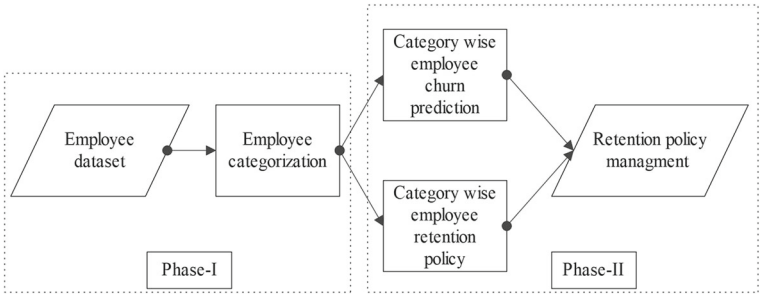

Fig. 1 The working flow map for the proposed ECPR scheme

Note that the employee categorization based on multiple factors (or accomplishments) is solved as a multi-attribute decision-making (MADM) problem.

3 Proposed ECPR scheme

Here, we first present an outline of the proposed scheme. The overall process is shown in Fig. 1. It comprises of two phases. In the first phase, we propose an accomplishment-based employee importance model (AEIM) for categorizing employees into three classes based on their importance in terms of productivity. The categorization is done by applying improved entropy weight method (IEWM) on accomplishment attribute values of each employee (called decision matrix) followed by TOPSIS (See Fig. 2 for details). In the second phase, we apply various machine learning algorithms on each employee category (generated in the first phase) for employee churn prediction and retention policy (See Fig. 3 for details). To illustrate the concept, each of these phases is explained step-wise on the human resource information system (HRIS) dataset, collected from the Kaggle Website (2017). The dataset has 14,999 instances and ten attributes (aka features). Two attributes are of float type, six are of integer type, and two are of a categorical type. Details of the attributes are shown in Table 2.

3.1 Overview of EWM and TOPSIS

This section presents an overview of EWM and TOPSIS. The working process of EWM is based on Shannon's entropy as explained in Shannon (2001). The entropy is a quantified way of uncertainty in terms of probability theory. This concept is wholly adapted to calculate the relative intensities of contrast models to reflect the average intrinsic information forwarded for decision making. Wang et al. (2009) extended Shannon's concept of entropy measurement as a method of weighting. Entropy weight is a parameter that describes how different alternatives approach each other concerning a particular attribute. The data source with lower probability values is considered to carry more information than that with higher probability values. That is to say, the former data are more important than the latter data and deserve higher weight (Lu and Yuan 2018). Entropy weights are used as input to the other MADM methods (such as TOPSIS) in order to rank the alternatives.

TOPSIS, first proposed by Hwang et al. (1981), is one of the classical ranking methods that can evaluate multiple alternatives with the same attributes. It is a simple and understandable concept with excellent computational efficiency and ability to measure the relative

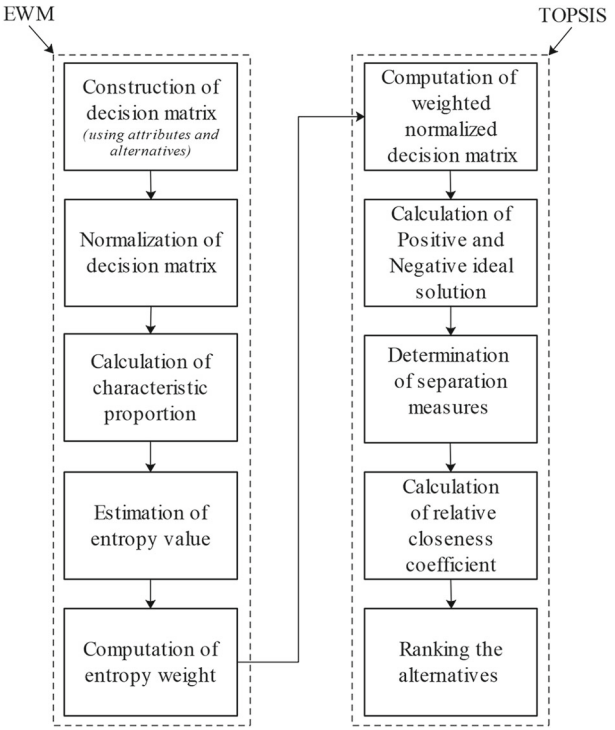

Fig. 2 The working flow graph of EWM and TOPSIS

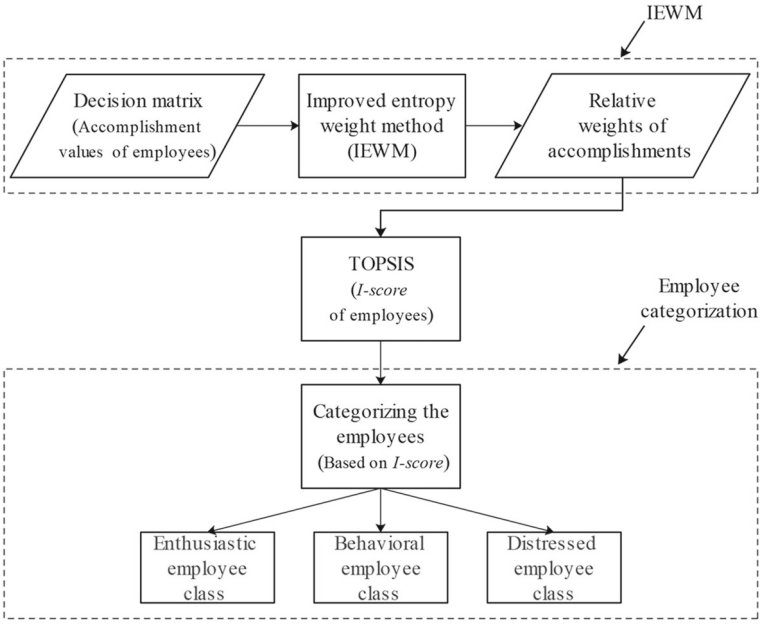

Fig. 3 The conceptual flow graph of the proposed AEIM model

Table 2 Data attributes with their data type and description

Attribute name	Data type	Description
Satisfaction_level	Float	This is the employee satisfaction level, which ranges from 0 to 1.
Last_evaluation	Float	This is the employer's measured performance, which also ranges from 0 to 1.
Number_project	Integer	How many number of projects is assigned to an employee?
Average_monthly_hour	Integer	How many average hours worked by an employee in a month?
Time_spent	Integer	This is the number of years an employee has spent in the organization.
Work_accident	Integer	If an employee has had an accident at work or not.
Left	Integer	If the employee has quit the organization or not.
Promotion_last_5year	Integer	Whether or not an employee has earned a promotion over the last five years.
Department	Categorical	Department of employment employee.
Salary	Categorical	Employee salary levels such as low, medium and high.

performance for each alternative in a simple mathematical form (Yeh 2002). The basic principle of TOPSIS is that the most desirable alternative is nearest to the positive-ideal solution and farthest from the negative-ideal solution. In other words, the positive-ideal solution is the best value solution for each alternative by maximizing the profit criteria and minimizing the cost criteria. On the contrary, the negative ideal solution is the worst value solution for each alternative by the profit criteria and minimizing the cost criteria. The step-wise working details of EWM and TOPSIS are shown in Fig. 2.

3.2 Employee categorization

According to the 'Pareto principle' (Sanders 1987), it can be stated that 80% of the profits of an organization are generated by 20% of its employees, and thus these employees are precious for the organization. Therefore, there is a need to differentiate among employees and identify the valuable ones. One way is to differentiate the employees based on different types of accomplishments that vary from organization to organization. For example, a healthcare organization may have patient satisfaction, medical work experience, etc., as the accomplishments on which the doctors can be evaluated. Table 3 shows various employees with their accomplishment types for different types of organizations. Another problem is that a large organization may have an enormous number of employees, and there may be a chance that an abundant number of employees want to quit the organization (known as attrition). Such a large number is indeed a challenge for retention because not all such

Table 3 Some examples for accomplishment types in different organizations

Type of organization	Employee type	Type of accomplishment
Healthcare	Doctor	Patient satisfaction, Medical work experience
Website design	Web designer	Hit rate, Strategy, Usability, Content
School/College	Teacher/Professor	Student passing rate, Student feedback
Restaurant	Waiter	Customer feedback, Service rating

employees are profitable for the organization. It is challenging to invest in retaining all of these employees. Additionally, holding on all the employees are not equally important for the organization. Therefore, class-wise retention is more effective and productive.

With this motivation, we propose a novel AEIM model (depicted in Fig. 3), which aims at categorizing the employees according to their importance in terms of productivity. It utilizes decision-making methods by contemplating different accomplishments of the employee, such as their satisfaction level, last evaluation, number of projects, average monthly working hours, and time spent in the organization. First, relative weights of different accomplishments are determined using an improved entropy weight method (IEWM). Next, TOPSIS is applied to categorize the employees into various classes, i.e., Enthusiastic, Behavioral, and Distressed. By enthusiastic employees, we want to mean the employees who are the most productive for the organization. Similarly, the behavioral and distressed employees are those who are average and less productive, respectively. We now discuss in details the process of employee categorization as follows.

3.2.1 Relative weights of accomplishments using IEWM

It is worth noting that for a decision-maker, it is quite challenging to quantify the importance of two employees if they have the same type of accomplishments, such as an equal number of projects done or equal time spent in the organization. However, both the accomplishment may not have equal importance according to decision-maker perception. For example, in order to judge the importance of an employee, the number of projects done by him/her may be given more importance than the time spent in the organization or vice versa. Therefore, while evaluating the employees' importance, each type of accomplishment should have different weights in order to reach the correct conclusion. In view of the above facts, we apply EWM for assigning the relative weight to various employee accomplishments. The step-wise procedure of EWM, along with a case study, is as follows. However, this may be noted that we incorporate some change in step 5 of the original EWM with the justification therein. This modification improves the accuracy of the relative weights, and hence EWM is referred to as IEWM.

Step 1: *Construct a decision matrix*: We first build decision matrix $X(= [x_{ij}])$ for n number of employees with m number of accomplishments. The structure of X is as follows.

$$X = \begin{array}{c} \\ E_1 \\ E_2 \\ \vdots \\ E_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_m \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \end{array} \tag{1}$$

where E_i denotes the employee i, $i = 1, ..., n$; A_j represents the j^{th} accomplishment, $j = 1, ..., m$, and x_{ij} is the value of accomplishment A_j for employee E_i. An example decision matrix based on the benchmark HRIS employee dataset (refer to Section 3) is shown in Table 4.

Step 2: *Normalize the decision matrix*: The number of accomplishment may be different from each other having different set of values and meanings, thereby causing inconsistent comparisons. Therefore, the decision matrix X needs to be normalized to standardize the data. Although any normalization method can be used in MADM approaches, existing

Table 4 Decision matrix based on HRIS employee dataset

	Satisfaction_level A_1	Last_evaluation A_2	Number_project A_3	Average_monthly_hour A_4	Time_spent A_5
E_1	0.68	0.53	4	197	3
E_2	0.31	0.88	7	272	4
E_3	0.91	0.86	6	262	6
E_4	0.45	0.51	2	160	3
E_5	0.82	0.87	5	283	4
E_6	0.52	0.47	3	150	2

studies suggest that most preferable method is vector normalization as it can generate most consistent results (Lu and Yuan 2018). Thus, we employ vector normalization method and calculate the normalized decision matrix $R(= [r_{ij}])$ in which normalized value r_{ij} is calculated as follows.

If the accomplishment of the employee is positive, i.e, it should be maximum for better performance, then normalization is done using the following equation.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1,\ x_{ij}\neq 0}^{n} (x_{ij})^2}},\ \text{for } i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \tag{2}$$

If the accomplishment of the employee is negative, i.e, it should be minimum for better performance, then normalization is done using the following equation.

$$r_{ij} = \frac{\frac{1}{x_{ij}}}{\sqrt{\sum_{i=1,\ x_{ij}\neq 0}^{n} \frac{1}{(x_{ij})^2}}},\ \text{for } i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \tag{3}$$

Step 3: *Characteristic proportion calculation for an employee*: As the values of an accomplishment varies with respect to the employees, we use characteristic proportion that implies the probability of an accomplishment value for a particular employee in all the other employees. We consider p_{ij} to be the characteristic proportion of A_j for E_i, defined by (4).

$$p_{ij} = \frac{r_{ij}}{\sum_{i=1}^{n} r_{ij}},\ \text{for } i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \tag{4}$$

The value of p_{ij} lies in the range [0,1].

Step 4: *Entropy value estimation for each accomplishment*: Using the characteristic proportion values, we estimate entropy value for each accomplishment according to (5).

$$e_j = -\frac{1}{\ln(m)} \sum_{i=1}^{n} p_{ij} . \ln(p_{ij}),\ \text{for } j = 1, 2, \ldots m \tag{5}$$

where e_j denotes the entropy measure of A_j for all employees having the range as [0, 1]. For a particular accomplishment j, if the difference among its values p_{ij} is higher for different i, then its entropy value e_j is small. Note that the attribute with smaller entropy value reflects larger amount of information, i.e., the attribute is more important and deserves higher weight (Lu and Yuan 2018).

Step 5: *Computing the entropy weight for each accomplishment*: Let $w_e(j)$ be the entropy weight of the A_j, we calculate $w_e(j)$ according to (6).

$$w_e(j) = \frac{1 - e_j}{\sum_{j=1}^{m}(1 - e_j)}, \quad 0 \leq w_e(j) \leq 1, \quad \sum_{j=1}^{m} w_e(j) = 1 \tag{6}$$

According to (6), when all entropy values $e_j \rightarrow 1 (j = 1, 2, \ldots m)$, a small difference among the entropy values will bring about the change in the corresponding entropy weight. For example, let {0.9, 0.8, 0.7} and {0.99, 0.98, 0.97} be the estimated entropy vectors for $\{A_1, A_2, A_3\}$ and $\{A_4, A_5, A_6\}$ accomplishment, respectively. Here, the differences among the entropy values of the two vectors are not the same, however, they turn out to have the same entropy weight vector {0.167, 0.333, 0.5} when using (6). This is obviously improper manner to assign weights because different entropy value vectors provide different amounts of information, so they should be given different entropy weights.

To overcome this downside, we replace the expression of entropy weight in (6) as follows:

$$w_e(j) = \begin{cases} \alpha \times w_{e1}(j) + \beta \times w_{e2}(j) & \text{if } e_j < 1 \\ 0 & \text{if } e_j = 1 \end{cases}, (j = 1, 2, \ldots m) \tag{7}$$

where,

$$w_{e1}(j) = \frac{1 - e_j}{\sum\limits_{j=1,\ e_j \neq 1}^{m} (1 - e_j)}, \quad w_{e2}(j) = \frac{1/e_j}{\sum\limits_{j=1,\ e_j \neq 0}^{m} (1/e_j)} \tag{8}$$

Here, α and β are constants (such that $\alpha + \beta = 1$) and $0 \leq w_e(j) \leq 1$, $\sum_{j=1}^{m} w_e(j) = 1$. When the entropy value $e_j \rightarrow 1$ $(j = 1, 2, \ldots m)$, a small difference among the entropy values will result in a significant change of $w_{e2}(j)$. This means that $w_{e1}(j)$ and $w_{e2}(j)$ are complementary to each other. At the same time, α and β control the proportions of these two weights, i.e., when α is close to 0, $w_{e1}(j)$ will have little contribution to $w_e(j)$, and when β is close to 0, $w_{e2}(j)$ will have little contribution to $w_e(j)$. In another words, (7) and (8) can give a reasonable entropy weight regardless of the extreme entropy values.

The results obtained after applying IEWM (*Steps* 2 - 5) on the employee data (shown in Table 4) are presented in Table 5. In this paper, α and β are taken equal as 0.5 to show the equal impact of both the weights ($w_{e1}(j)$ and $w_{e2}(j)$).

Table 5 Results of IEWM computations

	Satisfaction_level A_1	Last_evaluation A_2	Number_project A_3	Average_monthly_hour A_4	Time_spent A_5
e_j	0.9666	0.9796	0.9579	0.9829	0.9687
$w_{e1}(j)$	0.2312	0.1414	0.2917	0.1186	0.2171
$w_{e2}(j)$	0.2009	0.1983	0.2027	0.1976	0.2005
$w_e(j)$	0.2161	0.1698	0.2472	0.1581	0.2088

3.2.2 Categorizing employees using TOPSIS

Here, we apply TOPSIS for categorizing employees based on the relative weights obtained in the preceding section. The step-wise procedure along with illustrations are given as follows.

Step 1: *Compute the weighted normalized decision matrix*: We multiply normalized matrix $R(=[r_{ij}])$ ((2) and (3)) by the entropy weights of accomplishments to obtain the weighted normalized decision matrix $Z(=[z_{ij}])$. The weighted normalized value z_{ij} is calculated as:

$$z_{ij} = r_{ij} \times w_e(j), \text{ for } i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m \tag{9}$$

Table 6 shows the calculated matrix Z using (9) for the employee data shown in Table 4.

Step 2: *Obtain the positive and negative ideal solution*: We define the positive ideal solution as I_j^+ and the negative ideal solution as I_j^-.

$$\begin{aligned} I_j^+ &= max\{z_{1j}, \ldots z_{nj}\} \text{ and } I_j^- = min\{z_{1j}, \ldots z_{nj}\}, \text{ for positive accomplishment} \\ I_j^+ &= min\{z_{1j}, \ldots z_{nj}\} \text{ and } I_j^- = max\{z_{1j}, \ldots z_{nj}\}, \text{ for negative accomplishment} \end{aligned} \tag{10}$$

For positive accomplishments, the larger the accomplishment value is, the better suitability of the corresponding organization. For example, an employee having higher satisfaction level or last evaluation should be more valuable for the organization. On the contrary, negative accomplishments have opposite meaning.

Step 3: *Determine separation measures for each employee*: Separation measures are calculated using widely accepted Euclidean distance method. According to existing literature (Taşabat 2019), other distance measures, such as Statistical, Manhattan, linear, spherical, Hamming, Chebyshev distance, etc., can also be applied here. The separation of each employees from positive ideal solution I_j^+ and negative ideal solution as I_j^- is calculated using (11) as follows.

$$S_i^+ = \sqrt{\sum_{i=1}^{n}(z_{ij} - I_j^+)^2} \text{ and } S_i^- = \sqrt{\sum_{i=1}^{n}(z_{ij} - I_j^-)^2} \text{ where } j \in \{1, ..., m\} \tag{11}$$

Table 6 Weighted normalized decision matrix

	Satisfaction_level A_1	Last_evaluation A_2	Number_project A_3	Average_monthly_hour A_4	Time_spent A_5
E_1	0.0924	0.0517	0.0839	0.0560	0.0660
E_2	0.0421	0.0858	0.1468	0.0773	0.0880
E_3	0.1236	0.0839	0.1258	0.0744	0.1321
E_4	0.0611	0.0497	0.0419	0.0455	0.0660
E_5	0.1114	0.0848	0.1048	0.0804	0.0880
E_6	0.0706	0.0458	0.0629	0.0426	0.0440

Table 7 Importance score (*I-score*) of employees using TOPSIS

Employee	S_i^+	S_i^-	CC_i (*I-score*)
E1	0.1051	0.0706	0.4016
E2	0.0927	0.1254	0.5751
E3	0.0219	0.1545	0.8759
E4	0.1476	0.0295	0.1665
E5	0.0620	0.1168	0.6532
E6	0.1436	0.0354	0.1978

Step 4: *Calculation of relative closeness of each employee to the ideal solution*: The closeness coefficient CC_i (between $0-1$) indicates the relative closeness of i^{th} employee which is determined as follows.

$$CC_i = \frac{S_i^-}{S_i^- + S_i^+}, \text{ where } i \in \{1, ..., n\} \tag{12}$$

Step 5: *Class-wise categorization of employees based on importance score*: Here, we consider the closeness coefficient (CC_i) values as the importance score (*I-score*) of the employees. Table 7 shows each employee's *I-score*, calculated by using 9- 12 (Steps 1-4) of TOPSIS.

Next, we categorize the employees based on their *I-score*. Let $maxR$ and $minR$ be the maximum *I-score* and minimum *I-score*, respectively. Let D be defined as $(maxR - minR)/Nclass$, where $Nclass$ is number of total employee classes. Then, *I-score* ranges are mathematically defined (shown in Table 8) based on which the employee are categorized into ($Nclass = 3$) classes (i.e., *Enthusiastic*, *Behavioral*, and *Distressed*). Table 8 shows the class-wise categorization of six employees based on their I-score (obtained in Table 7). The output of AEIM is further used in prediction and retention phases.

3.3 Prediction of churning employees followed by retention policy management

The main objective of this step is to identify the employees from each employee class that has the highest probability of quitting next followed by determining the factors causing employee attrition. The proposed ECPR scheme integrating with prediction and retention phase is shown in Fig. 4. The working process for the same can be described by the following steps.

Step 1: In this step, the original employee dataset is divided into enthusiastic, behavioral, and distressed employee dataset using AEIM model (defined in Section 3.2).

Table 8 Employee categorization into respective classes

Employee class	I-score range formula	I-score range values	Employee
Enthusiastic	$(minR + 2 * D,\ maxR]$	$(0.6395,\ 0.8759]$	E_3, E_5
Behavioral	$(minR + D,\ minR + 2 * D]$	$(0.4030,\ 0.6395]$	E_2
Distressed	$[minR,\ minR + D]$	$(0.1665,\ 0.4030]$	E_1, E_4, E_6

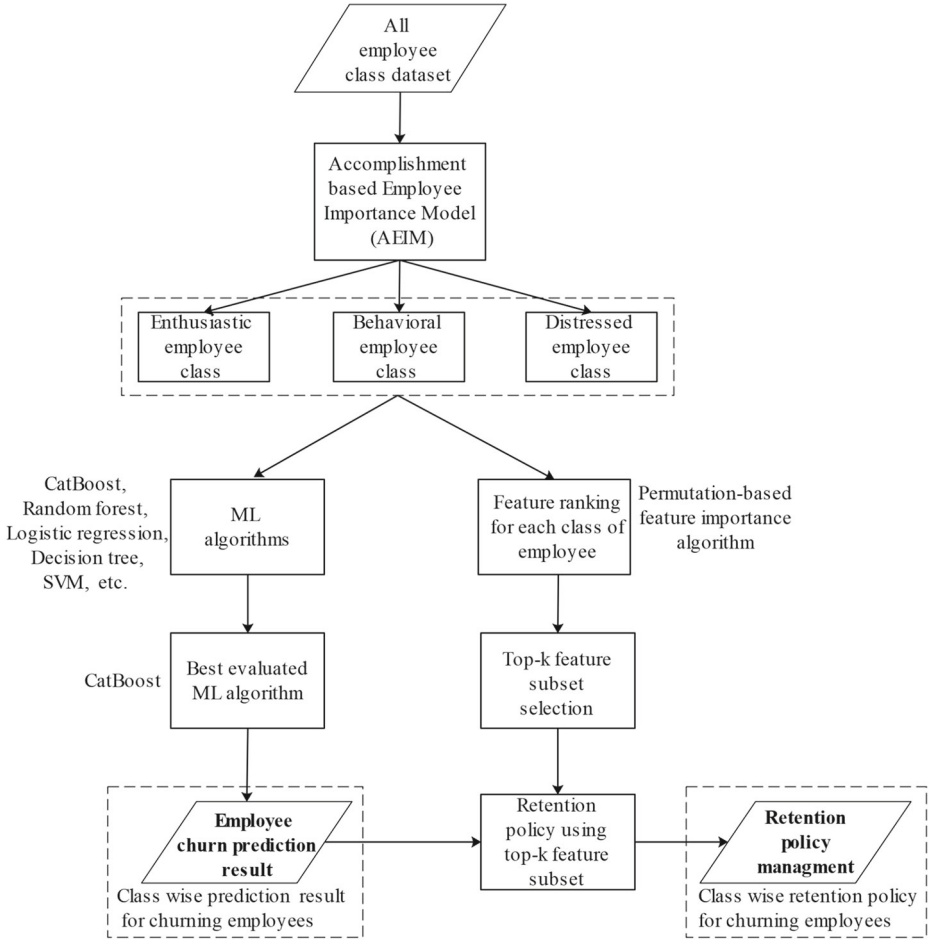

Fig. 4 The conceptual flow graph for predicting churning employees followed by retention policy management for each employee class

Step 2 : The primary goal of this step is to show the CatBoost is the accurate prediction model for the ECn problem. In this paper, we compare the result of CatBoost algorithm with three individual machine learning algorithm, namely support vector machine (SVM) (Cortes and Vapnik 1995), logistic regression (LR) (Webb et al. 2011), decision tree (DT) (Jin et al. 2009) and two tree-based ensemble machine learning algorithm, namely random forest (RF) (Breiman 2001), extreme gradient boosting (XGBoost) (Chen et al. 2015).

Step 3 : Next, we use the test dataset to estimate how well it performs on unseen data to determine the generalization error after evaluating the model best mounted on the training dataset.

Step 4 : Here, we get the class-wise employees who may quit the organization. The output of each class of the employee (i.e., *Enthusiastic*, *Behavioral*, and *Distressed*) from the prediction phase is used for the retention phase.

Algorithm 1 Permutation-based feature importance algorithm.

Input : Feature matrix (F_m), Trained model (T_m), Target vector (T_v), Error measure $\mathrm{E}(T_v, T_m)$

Calculate the original model error $E_{om} = \mathrm{E}(T_v, T_m(F_m))$.
foreach *Feature i = 1,...,n* **do**
    Generate $(F_m)^{perm}$ by permuting feature i in the data (F_m). This interrupts the association between feature i and (T_v).
    Calculate error $E_{perm} = \mathrm{E}(T_v, T_m(F_m)^{perm})$ based on the permuted data.
    Estimate the pemutation-based feature importance $(PF_{imp})^i = E_{perm} - E_{om}$.
end
Sort the feature by descending (PF_{imp}).
Output: (PF_{imp})

Step 5 : In this step, the primary goal is to find the top reasons for employee quitting. For this, first, we rank the important features causing employee attrition using permutation-based feature importance method. Feature values are shuffled randomly in this method, one column at a time, and the model's performance is measured before and afterwards. The results returned by the method reflect the changeover in a trained model's performance after permutation. Essential features are generally more susceptible to the method of shuffling, resulting in more exceptional results of significance. Fisher et al. (Fisher et al. 2018) presented the algorithm for permutation-based feature importance method. The pseudo-code for this algorithm is expressed as Algorithm 1.

The rationale behind using permutation-based feature importance method with the tree-based strategies is that they rank naturally by how well they enhance node purity. In addition to this, nodes with the most significant reduction in contamination occur at the beginning of the trees, whereas nodes with the least reduction in contamination occur at the end of the trees. Thus, we can create a subset of the essential features by pruning trees below a particular node. Next, we select the top-k important features and their importance value from each employee class.

Step 6 : Finally, using the top reasons for an employee quitting along with the result of step 4, we make the class-wise retention strategy. The goal is to highlight the prime reasons contributing to employee attrition so that an organization can make an effective decision support retention policy. Ultimately, we give the weights to the respective features according to their importance value so as to prioritize them while making a retention strategy. What needs to be pointed out is that handling the identified features depends on the nature and requirements of an organization. However, our retention strategy can be applied in the real-time environment of any organization.

4 Evaluation methodology

In this section, we discuss and present the methodology for evaluating the proposed scheme. The ECPR scheme is implemented in Python version 3.6 with the NumPy, SciPy, SciKit-Learn, Pandas, and Matplotlib libraries. We perform all the experiments on the Intel Core-i7-6500U 2.5 GHz processor and 8 GB of RAM on a 64-bit platform.

4.1 Baseline algorithms

Human resource information systems typically have categorical features. This makes the CatBoost algorithm better suited to the ECn prediction task. The prediction results are compared with that of three individual machine learning algorithm, namely support vector machine (SVM) (Cortes and Vapnik 1995), logistic regression (LR) (Webb et al. 2011), decision tree (DT) (Jin et al. 2009) and two tree-based ensemble machine learning algorithm, namely random forest (RF) (Breiman 2001), extreme gradient boosting (XGBoost) (Chen et al. 2015). It is well known that CatBoost is a modified gradient boosting decision tree algorithm (GBDT) which uses symmetric trees. It helps to decrease prediction time and error in the ECn problem, due to the inclusion of a large number of 'categorical features' (Prokhorenkova et al. 2018). Moreover, the following factors make CatBoost the best choice to use for our proposed scheme.

- *Handling categorical features:* One common technique for dealing with categorical features is one-hot encoding (Micci-Barreca 2001). However, this encoding of individual features with high cardinality contributes to algorithm inefficiency. Since algorithms offer more importance to continuous features than to dummy features, which mask the order of significance of the feature resulting in more unsatisfactory results (Pargent et al. 2019). In the CatBoost, these issues raised by one-hot encoding can be resolved by grouping the categories into a limited number of clusters before applying one-hot encoding. A popular way is to group the categories by using target statistics (TS) (Micci-Barreca 2001). Prokhorenkova et al. (2018) proposed a more effective TS strategy in the CatBoost algorithm, namely ordered TS to group the categories.
- *Feature blend*: The CatBoost produces feature combination by constructing a base tree with a root node composed of only one feature, and randomly selects the other best feature for the child nodes and represents it along with the root node feature. This feature of the CatBoost helps to reduce the dimensionality of the ECn problem.
- *Faster churn prediction*: The CatBoost algorithm utilizes oblivious trees as base predictors where the same dividing criterion is used throughout the entire tree level (Kohavi and Li 1995). These trees are balanced, so they are less likely to over-fit. Each leaf index is encoded as a binary vector in oblivious trees, with a length equal to the size of the tree. This principle is commonly used for estimating model predictions in CatBoost model evaluators, since all binaries use float, statistics and one-hot encoded functions. As a consequence, the employee's churn prediction produces better result.
- *Unbiased boosting*: The traditional GBDT techniques pose a statistical issue, namely, prediction shift during execution. In order to conquer the prediction shift problem, the CatBoost adopts ordered boosting in which different permutations are used for training different models. As a result of multiple permutations, the CatBoost demonstrates lower bias than the traditional GBDT techniques (Prokhorenkova et al. 2018). Therefore, the prediction of the employee churn is expected to be more generalized.

4.2 Performance metrics

The objective of the ECPR scheme is to maximize the positive performance of the prediction model on a test dataset for which the actual values are known. The 'Confusion Matrix' is used to evaluate the experimental results, which is a standard evaluation criterion in the area of machine learning. This provides a summary of the prediction results for a specific problem. In the case of the confusion matrix, there exist four following cases.

- Let *Tp* be the 'True positive' in which the attribute labeled as 'left' and is predicted to be 'left'.
- Let *Tn* be the 'True negative' in which the attribute labeled as 'left' and is not predicted to be 'left'.
- Let *Fp* be the 'False positive' in which the attribute not labeled as 'left' and is predicted to be 'left'.
- Let *Fn* be the 'False negative' in which the attribute not labeled as 'left' and is not predicted to be 'left'.

The detailed explanation of the metrics used for performance analysis is given as follows.

Classification Rate/Accuracy (ACC): It is defined as the total number of correctly predicted churners which is calculated as follows:

$$ACC = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{13}$$

Recall (RC): The rate of absolute churners which are correctly predicted is known as Recall, mathematically expressed as follows:

$$RC = \frac{Tp}{Tp + Fn} \tag{14}$$

Precision (PC): The rate of predicted churners which are precise is known as Precision, calculated as follows:

$$PC = \frac{Tp}{Tp + Fp} \tag{15}$$

Here, in ACC, RC, and PC, if the 'True positive' value is less then the 'False positive' value ($Tp < Fp$), then accuracy will always increase when we have a prediction rule consistently producing 'negative' output. Similarly, when ($Tn < Fn$), the same will happen when we have a rule that always gives a 'positive' output (Chicco and Jurman 2020). This limitation is called the accuracy paradox (Afonja 2017). To overcome this, we use another performance matrix, namely Matthew's correlation coefficient (MCC).

Matthew's Correlation Coefficient (MCC): It is known as the 'correlation coefficient' between the predicted and actual value. It is mathematically defined as follows:

$$MCC = \frac{Tp \times Tn - Fp \times Fn}{\sqrt{(Tp + Fp) \times (Tp + Fn) \times (Tn + Fp) \times (Tn + Fn)}} \tag{16}$$

Here, the MCC is used as a balanced measure because it includes all the four aforesaid cases of the confusion matrix. It has a range of '-1' to '1' where '-1' implies a wrong dichotomous prediction, and '1' implies correct dichotomous prediction. Thus, we can measure how well the prediction model performs.

5 Experimental results and discussion

For experimental purposes, we divide both the original and categorized dataset into training and testing datasets. Each of these datasets is divided into the ratio of 80:20 for training and testing purposes. We then obtain the prediction model by applying the training dataset

Table 9 Results of the different machine learning algorithms for all employee class dataset and new categorized dataset

Employee class	ML algorithms / Performance metrics	SVM	LR	DT	RF	XGBoost	CatBoost
	ACC	0.946	0.783	0.975	0.981	0.971	0.985
All employee	*RC*	0.951	0.781	0.981	0.984	0.972	0.989
class	*PC*	0.952	0.762	0.982	0.973	0.961	0.988
	MCC	0.853	0.751	0.933	0.941	0.921	**0.954**
	ACC	0.944	0.887	0.942	0.985	0.969	0.989
Enthusiastic	*RC*	0.951	0.892	0.983	0.991	0.971	0.996
class	*PC*	0.941	0.891	0.981	0.992	0.971	0.997
	MCC	0.867	0.722	0.946	0.974	0.925	**0.978**
	ACC	0.960	0.945	0.985	0.991	0.975	0.994
Behavioral	*RC*	0.961	0.940	0.989	0.990	0.981	0.991
class	*PC*	0.960	0.950	0.992	0.991	0.980	0.991
	MCC	0.854	0.797	0.947	0.983	0.911	**0.987**
	ACC	0.953	0.922	0.981	0.983	0.984	0.989
Distressed	*RC*	0.950	0.921	0.981	0.990	0.981	0.992
class	*PC*	0.941	0.931	0.982	0.982	0.972	0.991
	MCC	0.903	0.840	0.962	0.975	0.967	**0.981**

and evaluate it using the testing dataset. To evaluate the experimental results, we adopt the Accuracy (ACC), Recall (RC), Precision (PC), and Matthew's Correlation Coefficient (MCC) as performance measures. The results for the prediction and retention phases in the case of both the original dataset as well as the new categorized dataset are presented in the subsequent sections. The results achieved by CatBoost are shown to have better prediction performance than other existing ML algorithms.

5.1 Experimental results for prediction phase

Table 9 shows the experimental results of various individual and tree-based ensemble machine learning algorithms, which we have considered for experimental study. The learning performance are shown in terms of ACC, RC, PC, and MCC performance metrics.

The results marked in boldface for each class of employees conclude that the CatBoost algorithm provides the best prediction performance having the highest MCC. Fig. 5a, b, c, and d display the prediction performance for the dataset of 'All employee class' (without using the AEIM model), 'Enthusiastic employee class', 'Behavioral employee class' and 'Distressed employee class', respectively. From the results, we conclude that the prediction performance with the categorized dataset (using AEIM model) is better in terms of ACC and MCC as compared to those achieved without using the AEIM model.

It can be further concluded from Table 10 that prediction results of the CatBoost algorithm for categorized datasets are far better as compared to the results obtained using the original dataset. In this table, E_CatBoost, B_CatBoost, and D_CatBoost show the MCC for enthusiastic, behavioral, and distressed employee class, respectively. Whereas the Avg_CatBoost shows the average MCC of all the categorized employee classes, and All_CatBoost shows the MCC for the original employee dataset (All employee class).

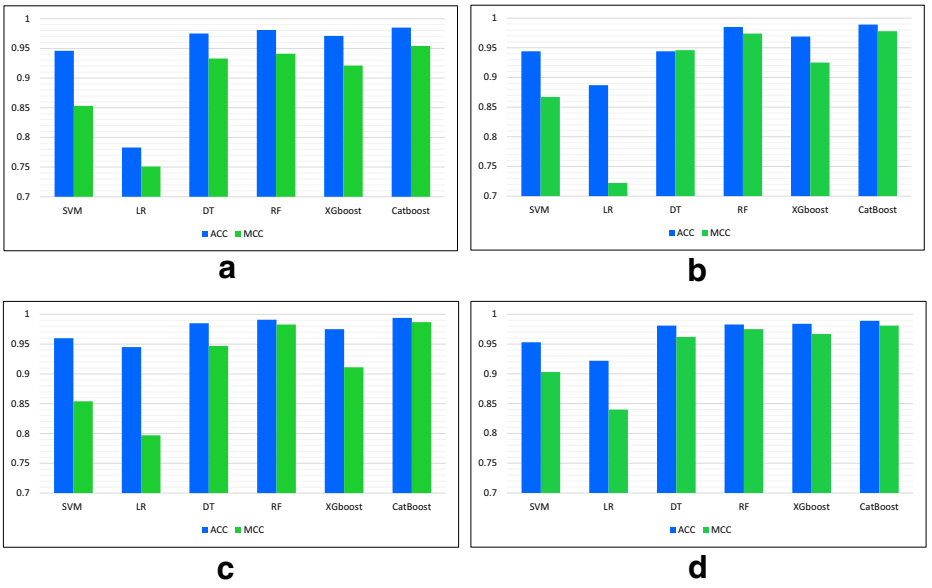

Fig. 5 Prediction performance of the different machine learning algorithms for **a** 'All employee class' (without AEIM) **b** 'Enthusiastic employee class', **c** 'Behavioral employee class', and **d** 'Distressed employee class'

Table 10 Comparative prediction performance result with 2.80% average improvement in MCC with proposed ECPR scheme

Employee class_CatBoost / Performance Metrics	E_CatBoost	B_CatBoost	D_CatBoost	Avg_CatBoost	All_CatBoost
MCC	0.978	0.0.987	0.981	0.982	0.954

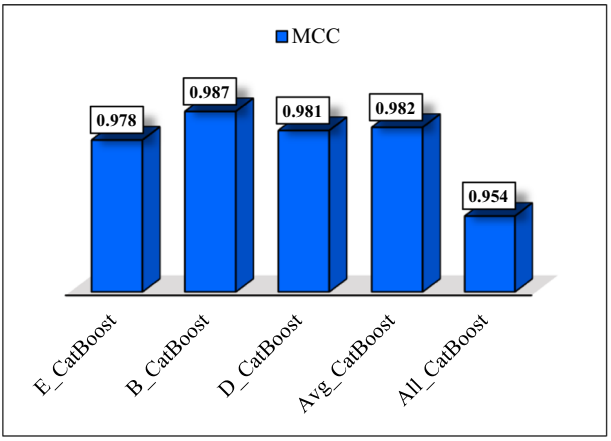

Fig. 6 Bar graph showing 2.80% average improvement in term of MCC using proposed ECPR scheme

Thus, we can summarize from Fig. 6 that using the categorized dataset achieves higher prediction accuracy over the original dataset for ECn problem. The proposed ECPR scheme with the AEIM model can perform better than directly using traditional machine learning algorithms.

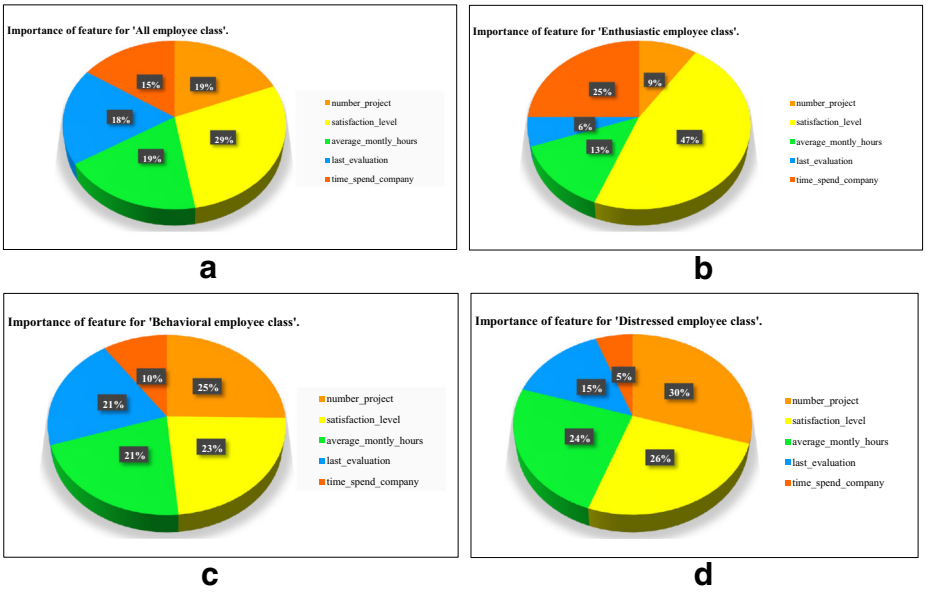

Fig. 7 Class wise primary reasons for employee quitting **a** 'All employee class' (without using AEIM), **b** 'Enthusiastic employee class', **c** 'Behavioral employee class', **d** 'Distressed employee class'

5.2 Experimental results for retention policy management

As discussed earlier, the process of finding the primary reasons for employee quitting consists of two steps. First, we rank features according to their importance using the permutation-based feature importance method. Next, we make a subset of top-k features. Then, the retention policy is designed based on the outcome of the prediction phase and the subset of the top-k features.

Figure 7a shows the top-5 vital reasons for the original dataset, which may be the prime cause of employees quitting the organization. The results show that 'satisfaction_level' has the highest (29%) percentage of correlation with the target feature ('left'). Next, Fig. 7b, c, and d show the top-5 vital reasons for enthusiastic, behavioral, and distressed employee class, respectively. It is observed from the figures that the top-5 correlated features and their respective importance values are different for each employee class. In the case of enthusiastic employees, 'satisfaction_level' feature is the most important cause of quitting, while in the case of behavioral and distressed employees, the 'number_project' feature has the highest correlation value with target feature ('left').

Thus, it can be concluded that by considering the important features of each respective employee class may help the organizations (especially for the large organizations) to make a more productive, cost-effective, and timely retention policy to stop employees from quitting their job.

6 Conclusion

In this paper, we have designed a MADM based scheme for ECn problem, referred to as employee churn prediction and retention (ECPR) scheme. The ECPR primarily sheds light on three key aspects. First, a novel accomplishment-based employee importance model (AEIM) has been proposed to categorize the employees into various classes based on their importance value by utilizing a multi-attribute decision making (MADM) approach. Second, the CatBoost ensemble machine learning algorithm has been applied to judge its potential for class-wise prediction of churning employees' attrition. Finally, the attributes responsible for employee attrition have been identified using permutation-based feature importance method so as to retain valuable employees. The benchmark dataset for the human resources information system (HRIS) has been used as an input to the ECPR scheme. The performance of the proposed scheme has been evaluated in terms of accuracy (ACC), Recall (RC), Precision (PC) and Matthew's Correlation Coefficient (MCC). The comparative results demonstrate that ECPR based on CatBoost algorithm has higher prediction accuracy rates, outperforming the direct prediction method by at least 2.80%. Note that our system is an effective tool for predicting the risk of churn in any organization, particularly in those with higher churn rates, such as telecommunications, where even small improvements in the accuracy of churn prediction models can be linked to significant financial gains. In addition, the results for the retention phase depicts different responsible attributes for attrition under different employee categories.

The results of the proposed ECPR scheme on the HRIS dataset shows its superiority over other traditional models and it is expected that the same model may produce reasonable results on other datasets having similar feature sets. However, it would be wise to test this system with other real time data from some large organizations. In the future, we plan to apply the proposed scheme to different organizations, which could help to generate real-time solutions with more promising performance.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

Afonja, T. (2017). Accuracy paradox, towards data science. https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b, Online: Stand 27. July 2020.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, D., & Wilson, S. (2007). The black books of outsourcing: How to manage the changes, challenges, and opportunities. Wiley.

Budhwar, P.S., Varma, A., Singh, V., Dhar, R. (2006). HRM systems of indian call centres: an exploratory study. *The International Journal of Human Resource Management*, *17*(5), 881–897. https://doi.org/10.1080/09585190600640976.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 04-2 pp 1–4.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.

Chien, C.F., & Chen, L.F. (2008). Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. *Expert Systems with Applications*, *34*(1), 280–290. https://doi.org/10.1016/j.eswa.2006.09.003.

Chu, A.T.W., Kalaba, R.E., Spingarn, K. (1979). A comparison of two methods for determining the weights of belonging to fuzzy sets. *Journal of Optimization Theory and Applications*, *27*(4), 531–538. https://doi.org/10.1007/bf00933438.

Cortes, C., & Vapnik, V. (1995). Support vector machine. *Machine Learning*, *20*(3), 273–297.

Dolatabadi, S.H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. In *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*: IEEE. https://doi.org/10.1109/ccoms.2017.8075270.

Fan, C.Y., Fan, P.S., Chan, T.Y., Chang, S.H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications*, *39*(10), 8844–8851. https://doi.org/10.1016/j.eswa.2012.02.005.

Fazlollahtabar, H. (2010). A subjective framework for seat comfort based on a heuristic multi criteria decision making technique and anthropometry. *Applied Ergonomics*, *42*(1), 16–28. https://doi.org/10.1016/j.apergo.2010.04.004.

Fisher, A., Rudin, C., Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv:180101489.

Frederiksen, A. (2017). Job satisfaction and employee turnover: a firm-level perspective. *German Journal of Human Resource Management*, *31*(2), 132–161.

Frierson, J., & Si, D. (2018). Who's next: Evaluating attrition with machine learning algorithms and survival analysis. In *Big data – BigData 2018* (pp. 251–259): Springer International Publishing. https://doi.org/10.1007/978-3-319-94301-5_19.

Gao, X., Wen, J., Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. *Mathematical Problems in Engineering*, *2019*, 1–12. https://doi.org/10.1155/2019/4140707.

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.

Harter, J.K., Schmidt, F.L., Hayes, T.L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis. *Journal of Applied Psychology*, *87*(2), 268.

Hom, P.W., Lee, T.W., Shaw, J.D., Hausknecht, J.P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, *102*(3), 530–545. https://doi.org/10.1037/apl0000103.

Hwang, C.L., & Yoon, K. (1981). *Multiple attribute decision making*. Berlin: Springer. https://doi.org/10.1007/978-3-642-48318-9.

Jin, C., De-lin, L., Fen-xiang, M. (2009). An improved ID3 decision tree algorithm. In *2009 4th International Conference on Computer Science & Education*: IEEE. https://doi.org/10.1109/iccse.2009.5228509.

Kaggle (2017). Hr analytics dataset. https://www.kaggle.com/lnvardanyan/hr-analytics#turnover.csv, Online; Stand 04. April 2020.

Khodabakhsh, M., Kahani, M., Bagheri, E. (2018). Predicting future personal life events on twitter via recurrent neural networks. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-018-0519-2.

Koh, H.C., & Goh, C.T. (1995). An analysis of the factors affecting the turnover intention of non-managerial clerical staff: a Singapore study. *The International Journal of Human Resource Management*, *6*(1), 103–125. https://doi.org/10.1080/09585199500000005.

Kohavi, R., & Li, C.H. (1995). Oblivious decision trees graphs and top down pruning. In *Proceedings of the 14th International joint conference on artificial intelligence*, (Vol. 2 pp. 1071–1077). San Francisco: Morgan Kaufmann Publishers Inc. IJCAI'95.

Krylovas, A., Dadelo, S., Kosareva, N., Zavadskas, E.K. (2017). Entropy–KEMIRA approach for MCDM problem solution in human resources selection task. *International Journal of Information Technology &, Decision Making*, *16*(05), 1183–1209. https://doi.org/10.1142/s0219622017500274.

Lu, L., & Yuan, Y. (2018). A novel TOPSIS evaluation scheme for cloud service trustworthiness combining objective and subjective aspects. *Journal of Systems and Software*, *143*, 71–86. https://doi.org/10.1016/j.jss.2018.05.004.

Mendoza-Gómez, R., Ríos-Mercado, V., Valenzuela-Ocaña, K.B. (2019). An efficient decision-making approach for the planning of diagnostic services in a segmented healthcare system. *International Journal of Information Technology &, Decision Making*, 1–35, https://doi.org/10.1142/s0219622019500196.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor Newsl*, *3*(1), 27–32. https://doi.org/10.1145/507533.507538.

Morrow, P.C., McElroy, J.C., Laczniak, K.S., Fenton, J.B. (1999). Using absenteeism and performance to predict employee turnover: Early detection through company records. *Journal of Vocational Behavior*, *55*(3), 358–374. https://doi.org/10.1006/jvbe.1999.1687.

Pargent, F., Bischl, B., Thomas, J. (2019). A benchmark experiment on how to encode categorical features in predictive modeling. Master Thesis in Statistics, Ludwig-Maximilians-Universität München, Leopoldstr 13, 80802 München.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Advances in neural information processing systems* (pp. 6638–6648).

Punnoose, R., & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence* 5(9)https://doi.org/10.14569/ijarai.2016.050904.

Rashid, T.A., & Jabar, A.L. (2016). Improvement on predicting employee behaviour through intelligent techniques. *IET Networks*, *5*(5), 136–142. https://doi.org/10.1049/iet-net.2015.0106.

Ren, C., & Li, H. (2011). Analysis on human resource management of travel agencies. In *2011 International Conference on Computer Science and Service System (CSSS)*: IEEE. https://doi.org/10.1109/csss.2011.5974582.

Saaty, R. (1987). The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling*, *9*(3-5), 161–176. https://doi.org/10.1016/0270-0255(87)90473-8.

Sanders, R. (1987). The pareto principle: its use and abuse. *Journal of Services Marketing*, *1*(2), 37–40. https://doi.org/10.1108/eb024706.

Saradhi, V.V., & Palshikar, G.K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006. https://doi.org/10.1016/j.eswa.2010.07.134.

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M. (2005). Employee turnover: a neural network solution. *Computers & Operations Research*, *32*(10), 2635–2651. https://doi.org/10.1016/j.cor.2004.06.022.

Shannon, C.E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Sikaroudi, E., Mohammad, A., Ghousi, R., Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, *8*(4), 106–121.

Tarnowska, K., & Ras, Z. (2018). From knowledge discovery to customer attrition. In *Lecture Notes in Computer Science* (pp. 417–425): Springer International Publishing. https://doi.org/10.1007/978-3-030-01851-1_40.

Tarnowska, K., Ras, Z.W., Daniel, L. (2020). Recommender system for improving customer loyalty. Springer International Publishing, https://doi.org/10.1007/978-3-030-13438-9.

Taşabat, S.E. (2019). A novel multicriteria decision-making method based on distance, similarity, and correlation: DSC TOPSIS. *Mathematical Problems in Engineering*, *2019*, 1–20. https://doi.org/10.1155/2019/9125754.

Tomar, A., & Jana, P.K. (2018). Mobile charging of wireless sensor networks for internet of things: A multi-attribute decision making approach. In *Distributed Computing and Internet Technology* (pp. 309–324): Springer International Publishing. https://doi.org/10.1007/978-3-030-05366-6_26.

Wang, T.C., & Lee, H.D. (2009). Developing a fuzzy topsis approach based on subjective weights and objective weights. *Expert Systems with Applications*, *36*(5), 8980–8985.

Webb, G.I., Sammut, C., Perlich, C., Horváth, T., Wrobel, S., Korb, K.B., Noble, W.S., Leslie, C., Lagoudakis, M.G., Quadrianto, N., Buntine, W.L., Quadrianto, N., Buntine, W.L., Getoor, L., Namata, G., Getoor, L., Xin Jin, J.H., Ting, J.A., Vijayakumar, S., Schaal, S., Raedt, L.D. (2011). Logistic regression. In *Encyclopedia of Machine Learning* (pp. 631–631). US: Springer. https://doi.org/10.1007/978-0-387-30164-8_493.

Xiao, J., Zhu, X., Huang, C., Yang, X., Wen, F., Zhong, M. (2019). A new approach for stock price analysis and prediction based on SSA and SVM. *International Journal of Information Technology & Decision Making*, *18*(01), 287–310. https://doi.org/10.1142/s021962201841002x.

Yeh, C.H. (2002). A problem-based selection of multi-attribute decision-making methods. *International Transactions in Operational Research*, *9*(2), 169–181.

Yigit, I.O., & Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*: IEEE. https://doi.org/10.1109/idap.2017.8090324.

Zhou, M., Liu, X.B., Chen, Y.W., Yang, J.B. (2018). Evidential reasoning rule for madm with both weights and reliabilities in group decision making. *Knowledge-Based Systems*, *143*, 142–161.