International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Low Complexity Early Employee Attrition Analysis Using Boosting and Non-Boosting Ml Techniques

G. Pratibha[1]*, Dr. Nagaratna P. Hegde[2]

Submitted: 27/05/2023 **Revised**: 07/07/2023 **Accepted**: 26/07/2023

Abstract: Every company, regardless of location, industry, or size, struggles with the problem of employee turnover and attrition. Predicting employee turnover is one of the top priorities for the human resources departments of many businesses because it is such a significant challenge. Employee turnover costs organizations a lot of money. In this research, we implemented multiple machine-learning methods to create a model that predicts employee attrition. Among them, the CatBoost algorithm is incorporated to identify a suitable approach for predicting employee attrition tasks early. The primary purpose is to find a method to predict the number of employees leaving their jobs accurately. Following training, the model for predicting employee attrition is assessed using a real dataset provided by IBM Analytics. This dataset has 35 features and around 1500 samples and is used to evaluate the model. Using CatBoost, we got a high accuracy on the Kaggle dataset titled "IBM HR Analytics Employee Attrition & Performance." We recommend using a technique called "synthetic generation" to create more combined features based on arithmetic operations, which improves the accuracy and area under the curve (AUC) of the original CatBoost model. This will allow you to get the most out of the fundamental characteristics of the dataset. We achieved high accuracy of 95.84% and consumed less time of 2.15sec as related to relevant studies; this indicates that our method is effective.

***Keywords*:** *CatBoost, Employee attrition, HR analytics, Random Forest. Synthetic features*

1. Introduction

Employee attrition is one factor contributing to the number of employees leaving their jobs. It comes at a financial cost in the kind of costs for hiring and training. Whenever experienced employees need to be replaced since they have moved on to work for competing companies [1, 2]. In addition, when a person leaves their position, they take explicit and tacit knowledge with the potential to disrupt critical social ties [3, 4]. Therefore, a company should prioritize decreasing employee turnover to keep a competitive advantage over its competitors. As a consequence of this, for the benefit of the corporation, the leaders of the corporation need to gain an understanding of the primary factors that contribute to the decisions that their employees make to leave the company and then take the right actions to boost the productivity, general workflow, and overall performance of their organization after their employees have left the company.

On the other hand, a higher retention rate signifies a reduction in the costs of employing new employees and training them, in addition to individuals with more experience and more experienced people joining the company's staff over time. Therefore, organizations currently have a strong commercial interest in understanding the causes of employee turnover to prevent employee attrition. In general, it is a goal of businesses to increase their profits as much as possible. Because they have fewer contractual obligations, workers at firms where they do basic tasks might turn to

on-call, occasional, and temporary labor. However, in firms where employees execute highly specialized jobs, employee specialization and continuity of work become critical. It has been demonstrated that businesses must recognize the importance of a person's competencies and capacity to acquire new information while working. As a result of the use of artificial intelligence in human resources, businesses can now convert data into knowledge in several ways, one of which is through the implementation of predictive models. These models make it possible to make predictions about employees based on data acquired by the organization over the preceding years. As a result, significant problems are mitigated, and all HR-related tasks are optimized.

As a consequence, estimating the employee turnover rate and determining the key contributing factors contributing to attrition in the existing organizations [5]. Several studies on employee attrition exist, one of which is the study in [6], which investigated the factors that impact employee turnover in the information technology division. They used a questionnaire to collect data from all 300 IT staff members in the division. They used a strategy based on fundamental percentages, the chi-square test, and the correlation coefficient technique. They concluded that there is no connection between

[1]* *Research scholar, Dept. of CSE, JNTUH, Hyderabad, Telangana, India*
[2] *Professor, Dept. of CSE, Vasavi College of Engineering, Telangana, India*
* *Corresponding Author Email: pratibhaphd123@gmail.com*

factors that influence the working conditions of employees and those that contribute to employee retention in the information technology industry.

The research presented in [7] aims to estimate employee turnover based on five different categories of attrition data. These folks utilize the 'Apriori' Association Rule Algorithm and the 'C5.0 Decision Tree Algorithm. Consequently, the algorithm functions more effectively when C5.0 is operated in association with the association rule algorithm as opposed to when C5.0 is used on its own. This study analyzes the causes or motives that drive an employee to quit. The Human Resources department's responsibility is to implement timely and appropriate measures, such as enhancing the working environment or providing productivity incentives. More concentration on significant features leads to turnover and offers an accurate categorization based on statistical data analysis. We use the dataset as a beginning point in our investigation. By conducting an association analysis on the heatmap containing 35 different qualities, we identify the qualities that strongly correlate with why an employee departs the organization. However, we intend to use machine learning models to determine the probability that a particular individual will resign from the organization.

The following are the primary contributions and organization of the paper: We explain the background work of employee attrition systems-based models in section 2. Section 3 proposed work. The section 4 contains the results and analysis. Finally, in Section 5, concludes the paper.

2. Background and Related work

Many studies have established the effectiveness of human resource management (HRM) in working scenarios, production and management, and developing productivity links [8]. Furthermore, the data suggest that effective human resource management influences productivity, impacting several business models [9]. However, the majority of studies concentrate on analyzing and monitoring customers and the behaviors they engage in [10]. Still, they must discuss a company's employees' most valuable assets.

In [11], the authors researched the primary elements that play a role in an employee's decision to leave a company and determine whether or not a particular worker would decide to leave the organization. After training, the model for predicting employee attrition is assessed using an IBM dataset. This dataset consists of 35 attributes and around 1500 samples in total, and it is used to evaluate the model. The findings are presented using traditional metrics, and the Gaussian Nave Bayes classifier was shown to have the best results when applied to the provided dataset. The overall false negative rate for the total data is 4.5 percent, making it the best recall rate of 0.54. This is because it evaluates how effectively a classifier can locate every affirmative case.

When investigating the factors influencing employee turnover rate, the authors of [12] employed a dataset of employee records to research. They studied fast-moving consumer goods (FMCG) data by applying DL-based predictive models and ML-based models. This shows that deep learning is superior to these approaches, and however, findings are validated by using a regression model in conjunction with an analytical hierarchy process (AHP). Both models evaluate the significance of the variables and generate weights, allowing the findings to be validated even further. According to the results, ML and DL have shown a greater accuracy rate in forecasting customer attrition (91.6%) than any of the other models. Algorithmic approaches include random forest and gradient boosting (82.5% and 85.4%, respectively). These findings may be helpful to managers of human resources (HR) in an organizational workplace environment.

In [13], the authors constructed a methodology to predict employee attrition and offer firms chances to resolve issues and enhance retention. The supervised machine learning algorithm, support vector machine, created the predictive model (SVM). The employee's work status was included in the archival employee data obtained from three IT companies in India's human resource databases. This employee data contained 22 input features. According to the confusion matrix accuracy test performed, the proposed approach achieved an 85% accuracy and a precision of 86.5%.

In [14], the authors researched to compare the effectiveness of several machine learning strategies, including the Decision Tree (DT) classifier, the Support Vector Machines (SVM) classifier, and the Artificial Neural Networks (ANN) classifier, in addition to the procedure for selecting the most appropriate model. They used the IBM dataset to evaluate and contrast the efficacy of various machine-learning methods. The preprocessing of the dataset utilized in this comparative research consists of multiple forms, including data exploration, visualization, purification and reduction, transformation, discretization, and feature selection. These methods were used to prepare the dataset.

Over fitting concerns are addressed in this work by applying processes involving parameter tweaking and regularization. To achieve optimization objectives, these tactics are put into practice. However, compared to the other categorization models examined, this model's accuracy percentage of 88.87% was the lowest of any model discussed. The following results are: the DT model is in second place, and the ANN model is in third place.

In [15], the authors developed tree-based binary classification models to forecast employee turnover based on corporate culture and management features. A data collection of applications that were anonymously submitted through Glassdoor's online portal is coupled with the information that is publicly available regarding company reviews to match the needs. As a result, one can assess the probability of an employee leaving a company while looking for a new position by employing statistical models such as decision tree (DT), random forest (RF), and gradient-boosted tree (GB).

The models based on the decision tree and random forest approaches were the most accurate for estimating how customers would stop using a service. However, the decision to quit the organization was influenced by various other factors, including but not limited to income, the culture of the company, and the performance of top management, amongst others. In addition, the methodologies utilized in this work could be applied to data sets exclusive to a company to generate distinct attrition models.

3. Proposed Model

This work was divided into three key steps: (i) derive an improved estimate of a depth parameter that was found to have a significant relationship with employee attrition. (ii) By applying ML models to examine the reliability of the updated estimation of the depth to predict employee attrition and (iii) finally, quantitatively examine the performance of each of these models and then choose the early employee attrition prediction model for the study.

A. Description of the dataset:

IBM Analytics offers access to the HRM dataset [16]. This US-based data set comprises 35 attributes and 1500 observations. In every way, people's jobs and personal traits are linked. The variable is used to identify the feature of the dataset. For example, an employee who responds with "no" didn't quit the company, whereas "Yes" means they quit. The machine learning system will be able to learn from data collected from the current world, eliminating the need for the system to be explicitly designed. Furthermore, the output predictions will be more accurate if this training method is carried out over time and on data relevant to the problem.

B. Data Preprocessing: Before using the data for training and testing ML algorithms, we convert it into a format that could be used by taking the steps below:

1. Remove all NULL values
2. Remove column values that were not unique.
3. Remove unnecessary columns
4. Converted an unsupported UTF format to a valid UTF format.
5. Rearrange the columns according to their weight.
6. Save the filtered data as a CSV file.

Data cleaning and reduction:

Because it has 35 features, the dataset is classified as high-dimensional. Any irrelevant qualities that do not contribute to the study's aims should be deleted.

Discretization and normalization:

Feature scaling or normalizing is used during the preprocessing step of data transformation. The range of independent variables or data components can be normalized with the help of a technique known as normalization, which is a procedure. Scaling or normalizing the features can be a valuable tool for preventing attribute reliance on measurement unit selection. Scaling can be done manually. After the data had been cleaned up and reduced, which included applying the discretization method and shifting the attribute type from numerical to nominal format, the next step was to analyze the data set was then ready for analysis. Because of this method, the range of data characteristics was narrowed down to 0 and 1. Following the above findings, four (4) attributes were eliminated, leaving thirty (30) attributes. Following the regeneration of the interquartile filter, it was discovered that there were no outliers.

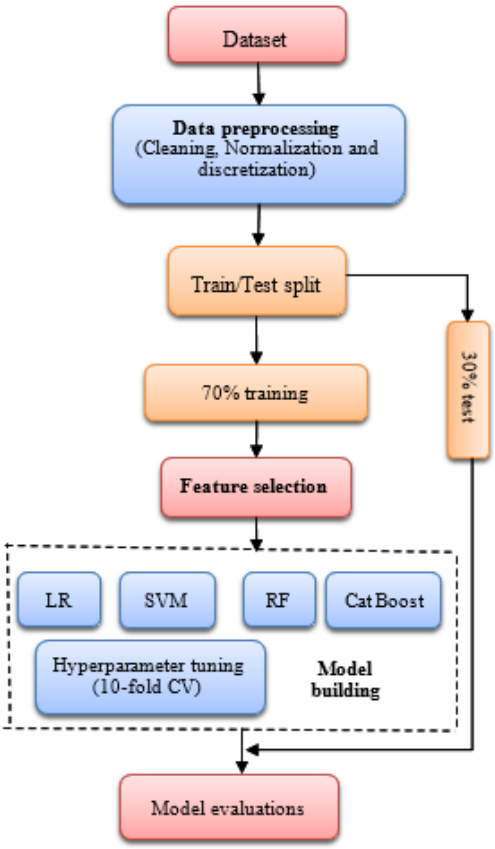

Fig 1: Flowchart of Employ attrition prediction modeling

C. Feature importance: This is picking important features before applying them to classifier models and explaining how they measure up. In general, feature importance quantifies the value that a feature (attribute) contributes to the growth, and significance is directly proportional to the number of times it appears in the decision tree. Next, the performance improvement is averaged based on the location of the leaf. When the weight of the leaf is greater, its position on the tree will be lower, bringing it closer to the treetop. Finally, every tree-weighted sum of feature importance is computed and takes the average of that value over all the boosting trees to get the feature importance.

Model Construction:

Critical factors drive employee attrition to understand better why employees leave the IBM Company. This part of the section focuses on developing machine learning models to choose and organize the data. Moreover, we will build a model to predict employee attrition using the data provided.

The modeling phase includes selecting classification models for testing after reducing the features. In this case, datasets are split into 70% for the training set and 30% for the testing set, respectively.

- The training set accounts for 70% of the dataset and contains 1029 observations. This part of the dataset was set aside for the training phase so that the model could figure out what the patterns in the data were. The remaining 30% were included in the testing set.
- Similarly, 441 observations in the test set were used for evaluation purposes and to quantify errors between the estimated and actual results throughout the testing and validation phases.

Logistic Regression (LR):

The training data will be used to compute the coefficients of the logistic regression algorithm, which are called beta values. The maximum-likelihood estimation method is needed to get this done successfully. Maximum-likelihood logistic regression is based on the concept that a search technique should look for coefficient values. The idea behind this concept is that maximum-likelihood logistic regression should be used (for example, a probability of 1 if the data is the primary class). For example, a model would be produced using the optimal coefficients if it projected a value near 1 for the default class and a value close to 0 for the other class.

Support vector machine (SVM):

The elimination of exceptions will have a severely adverse impact on the training model's accuracy. Many outliers in the associated dataset remain in our investigation. To overcome this issue, a kernel-based technique known as support vector machine (SVM) was

proposed by identifying a solution to the minimization problem, the coefficients for the SVM model:

$$arg \min_{\beta_0,\beta} C \sum_{i=1}^{N} L_e(y_i - f(x_i)) + \sum_{j=1}^{p} \beta_j^2 \quad (1)$$

In eq. (1) C , L_e denotes a penalty and loss respectively. Applying the sign function to a given group of unknowns enables us to arrive at the solution parameters.

$$Z = f(y) = sign\left(\sum_{i=1}^{N} y_i p_i K(x, x_i) + c\right) \quad (2)$$

Where p_i and c are necessary components in the construction of an optimal separation hyperplane, and the following expression explains $K(x, x_i)$.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (3)$$

$\|\boldsymbol{x} - \boldsymbol{x'}\|^2$ represents Euclidean distance. So, when SVM was applied to our dataset, we realized that altering the relevant parameters was difficult and obtaining the entire model would take time. In addition, the SVM model fell short of our expectations for a good model that just needed a small amount of memory.

Random forest (RF):

The random forest approach is a widely used solution in the bagging process for addressing the challenge given by large correlations between various individual trees. To further increase prediction accuracy, we use a technique known as "feature bagging," in which each tree is considered the vote, and each variable is assigned for the class with the highest probability of occurring. Using the probability distribution, we obtain the output function. $q_t\left(b/a\right)$. of each tree:

$$Y = arg\ max \frac{1}{T} \sum_{t=1}^{T} q_t\left(b/a\right) \quad (4)$$

When the RF method is applied, it deals appropriately with noisy and intricate data. In addition, it demonstrated a high degree of prediction accuracy when contrasted with several traditional modeling situations, as mentioned before.

CatBoost focuses on category features. The gradient-boosting decision tree (GBDT) approach is used. Because the complete dataset can be used, this works well in small datasets. CatBoost corrects by adjusting the bias, which improves the method's accuracy. It performs feature categorization during the training stage rather than the step assigned to the preprocessing of features, which is its key advantage. This eliminates reliance on data sorting—the classic GBDT figures out the average label value using greedy target-based statistics instead of classification features.

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} . y_j}{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}}} \quad (5)$$

In eq (5), I denote the indicator function, which is the *i*-th feature of the *k*-th training sample. CatBoost begins by arranging all of the data in an unpredictable sequence, and then it assigns a score to the level of quality possessed by each category. To achieve the goal of reducing the influence of noise on data distribution, the priority weight coefficients are used in such a way that the following is described as the appropriate approach to do so:

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} . y_j + \beta_p}{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} + \beta} \quad (6)$$

Figure 2 presents an overview of the CatBoost algorithm's underlying structure. where *p* represents a past value and the weight of that initial value, CatBoost has several benefits, one of which is that it takes a greedy method to examine possible combinations. As a result, the procedure does not end up overfitting the data. When the tree is split for the first time, CatBoost does not consider any combinations. However, when the tree is split a second time, and at any future time, CatBoost considers all predetermined combinations using each category feature in the dataset. All the tree splits picked are counted as classes [17].

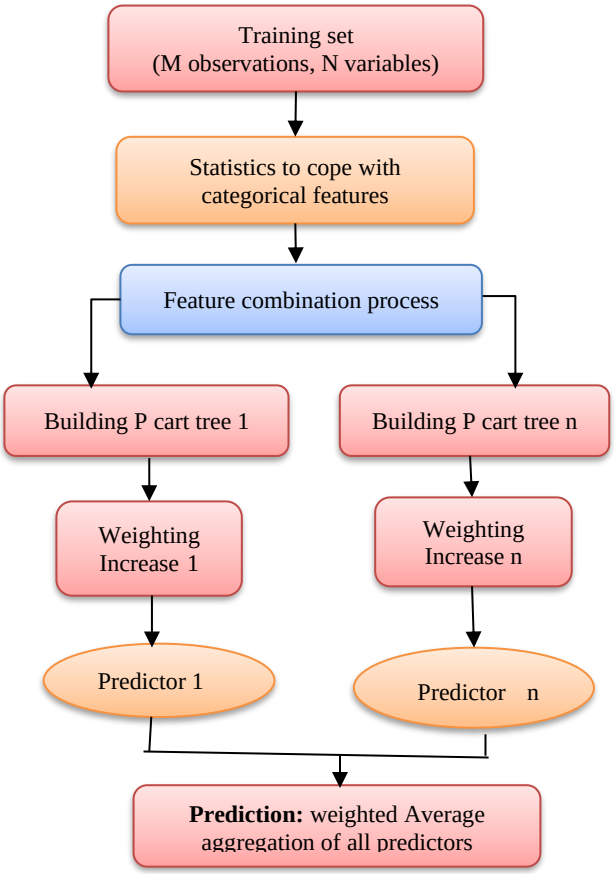

Fig 2: The structure of CatBoost

There is no combination process during the first stage of the tree's splitting phase; instead of applying to the second split. When building a new split for the tree, a greedy approach is the best way to group all the categorical features into new ones. This novel technique can be shown in the following Algorithm 1.

In unbiased boosting, the TS method is utilized while working with categorical features to translate those features into their corresponding numerical values. During this time, the distribution undergoes a shift, which reveals a different property than the distribution in its initial state. Because of this, the solution will differ, a problem that can only be avoided when discussing classic GBDT approaches. We typically employ random permutations of the training data in CatBoost to boost the algorithm's robustness. This is one of the methods we use to increase its performance. There would be no possibility of the CatBoost algorithm leading to an overfitting problem because the model would have a variety of permutations. We can train M_t in the method described above after we have independent models for each permutation of m. We must keep $O(m^2)$ estimates for model updating to build a tree.

Proposed CatBoost:

CatBoost, which outdoes during the investigation, the categorical variables served as the primary emphasis. However, traditional CatBoost is impractical when applied to a huge dataset with inadequate continuous variables. We suggested a unique technique that combines CatBoost with synthetic features to overcome this constraint. We begin by ranking features based on their contribution to the classification outcome. Then, we conduct a random arithmetic operation on the combined first two ranking features. Next, we analyzed their popularity in the produced forest to predict the probability of selection with the seed features. This distribution was understood by calculating the number of times it has been observed in the forest's trees. Finally, we can use the method described above to continue the reproduction process by utilizing the most commonly found characteristics. This way of developing new features is a creative way to eliminate the less important ones while strengthening the strongest ones, which is essential for the correctness of the model with Algorithm 2.

C : Input training data,

C_{new}: synthetic features,

B: base learners count,

δ: the feature incorporating set point,

F: feature significance; $R=\{r_1, \ldots, r_s\}$: collection of terminating

1 for $s = 1,\ldots, S$ **do**

2 Train r_s

3 Eliminate *C* features with $Q_c < \delta$;

4 Accept r_s if model $Q_c > \delta$;

5 **for** $c = 1,\ldots, C_{new}$ **do**

6 Based on the feature importance *F*, sample features f_1 and f_2 are calculated.

7Sample operation is processed with arithmetic operations for f_{new};

8 Increasing sets *C* for features with f_{new}

9 **end**

10 **end**

11 **return** $R = \{r_1, \ldots, r_s\}$

When calculating the full features, we may combine several arithmetic operations, the + and / operations being the most common. We produce synthetic features using heuristics, and the significance of those features is determined by the feature significance provided through tree building. This allows our model to capture features more flexibly. Algorithm 2 describes the whole CatBoost with synthetic features approach.

Step 1: During each training iteration, a base learner provides dataset *C* features whose relevance is evaluated by how frequently they appear and is less than a particular threshold.

Step 2: Based on the base learner r_s,, the model generates a new sample distribution θ_F, indicating the feature importance. In addition, the following framework is utilized in each cycle to generate synthetic features.

Step 3: From θ_F distribution, two features, f_1, and f_2, are sampled. These features can be actual or synthetic features created in a previous step.

Step 4: This sampling operation is done based on {+, −, ∗, /} using a uniform distribution, new feature, $f_{new} = f_1 \circ f_2$, is suitably assigned.

Step 5: The creation of synthetic features is iterated as described in the previous step until the required synthetic features, C_{new}, are obtained. After that, the augmented dataset is used in the subsequent iteration to construct the r_{s+1} base model by integrating extra features.

4. Results and Discussion

This study uses a personal desktop computer as a training tool. A core i5 CPU and 8 GB RAM with graphics card are included. All the program code is done using Python. For the RF Model, the ideal settings for the CatBoost model depth are set to 15, the learning rate is set to 0.05, and the N estimators are set to 170. As a penalty, the maximum depth is 16. The kernel is RBF, the C is 7, the Gamma is 0.1, and the I1 solver is liblinear. In this study, we use two of the most frequently used accuracy estimate criteria: The most straightforward criterion for classification result accuracy is the overall accuracy (OA), which measures how well the data is detected.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (9)$$

$$F1 - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (10)$$

To be more specific, there are four significant values: True Negative (TN), False Negative (FN), False Positive (FP), and True Positive (TP).

Figure 3 shows that hourly, daily, and monthly rates do not have a significant association with any of the other factors. Nevertheless, the data presented in the figure demonstrates that there is a statistically significant correlation between monthly income (0.95), full working years (0.78), and total working years (0.77). Consequently, the hourly, daily, and monthly rates have been removed from the dataset.

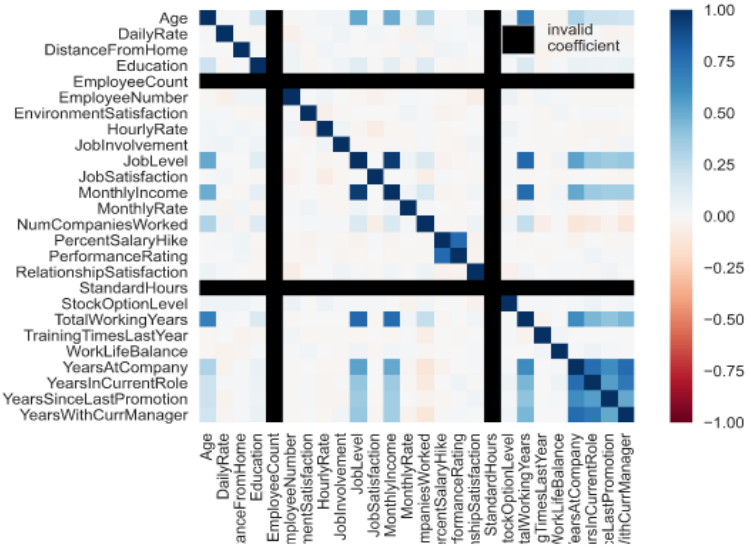

Fig 3: Correlation Matrix

According to Figure 4, the top three key variables to predict whether someone tends monthly income, age, and distance from home are the three most important factors influencing a person's decision to quit. At the same time, marriage status, married individuals, and women aged 40–50 are the demographics that are least likely to do so. Income is the key reason why individuals leave a company, which is intimately connected to people's quality of life. Individuals with more money in their pockets may be able to afford more expensive services (like medical care) and a healthier way of life. As a result, many desire to make more money and plan to quit their existing jobs to work for firms that pay well.

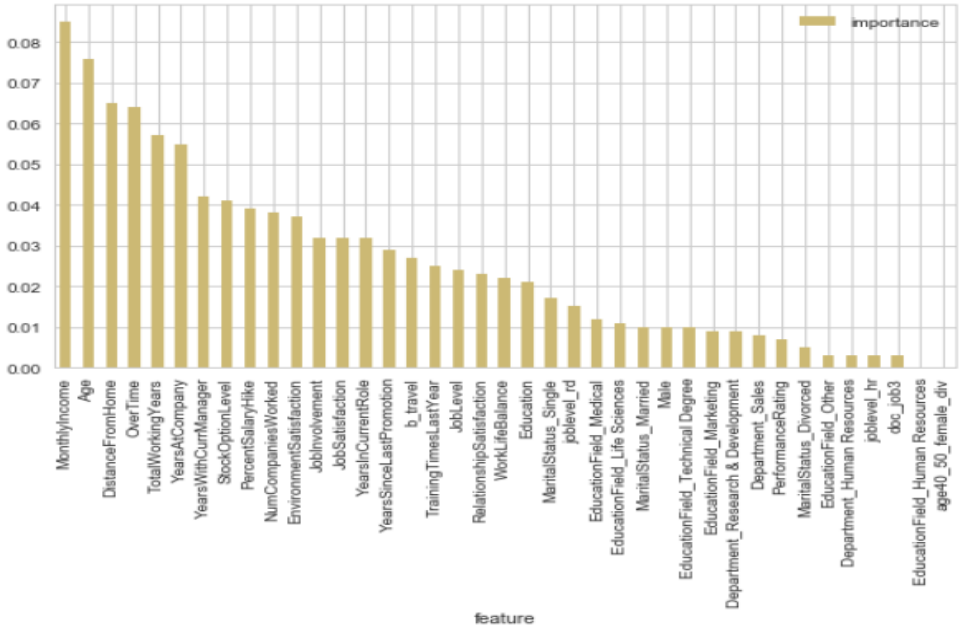

Fig 4: Important Features

We used CatBoost to choose features and detected 19 significant traits for our inquiry. In this study, we looked at four ML models lately mentioned by numerous scholars (e.g., CatBoost, LR, RF, and polynomial, radial basis function) based on SVM. The analysis focused on operational variables such as Time spent by the central processing unit (CPU) in training multiple models, the absence of a graphics processing unit (GPU), and the size of models stored in memory. In addition to that, it considers the machine learning metrics: Cross Validation Mean Score, Model Accuracy, F1-Score, Classification Precision, and Recall. Table 1 shows a comparison of performance indicators for four machine learning models.

Table 1. Comparison of performance metrics for 4 ML models

Type of Model	Accuracy	Precision	Recall	F1-score
LR	89.96	88.12	87.14	89.88
SVM	88.24	87.15	86.12	87.54
RF	88.85	87.59	86.57	89.21
CatBoost (Our work)	95.84	96.13	94.22	94.66

Fig 5 shows that our CatBoost performs better than all the other three ML approaches.

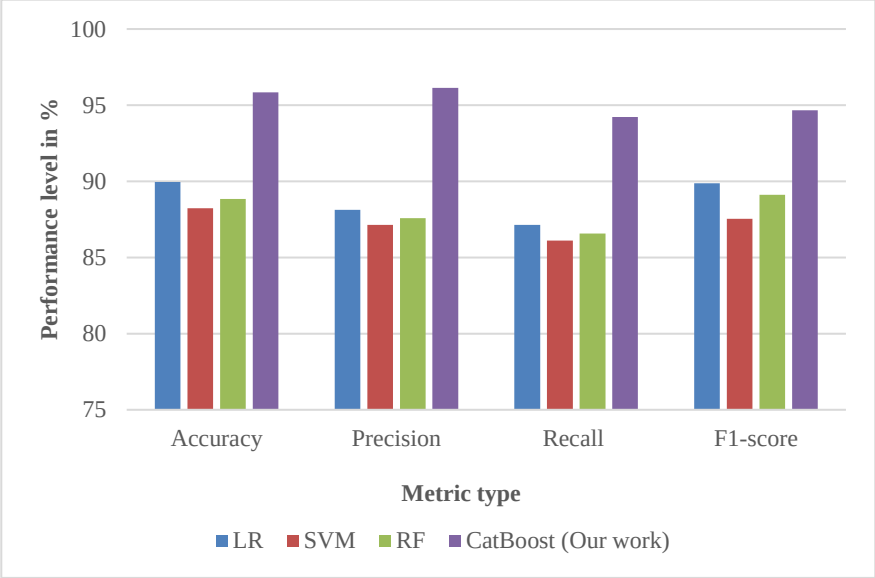

Fig 5: Comparison Performance metrics of four ML models

Table 2: Computational time in sec

Type of Model	Base Model	Hyper Parameter Tuning
LR	23.3	2.98
SVM	39.95	4.75
RF	35.54	3.71
CatBoost (Our work)	19.5	2.15

From Figure 6, it is clear that Our CatBoost takes less time as compared to all the other three ML approaches.

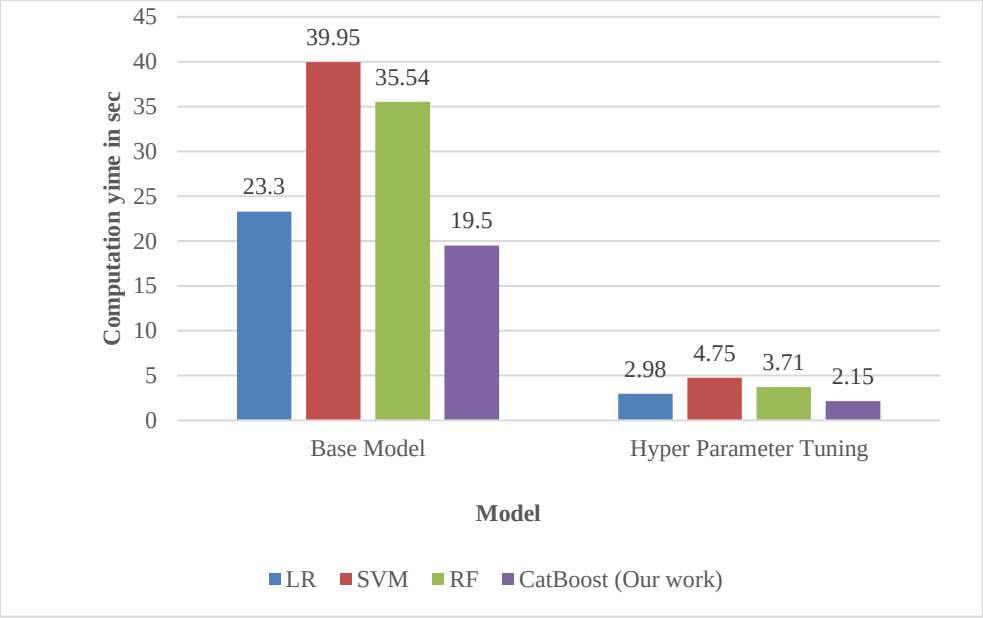

Fig 6: Comparison of computation time of four ML models

This section clarifies the exact model settings and crucial features during the data training. When determining the degree to which various machine learning models can make accurate predictions, the major areas in which we concentrate our focus are the metrics like ROC, AUC, and accuracy.

Table 3: Comparative result of ML algorithms

Model	**LR**	**SVM**	**RF**
Accuracy	89.96	88.24	88.85
AUC	90.14	96.22	98.52

Table 4: Results of ML algorithms of the Boosting Type ensemble learning

Model	**CatBoost**	CatBoost (Our work)
Accuracy	93.35	95.84
AUC	98.59	98.80

The dataset is applied to all the ML models to evaluate their overall performance and determine which parameter tweaking methods get the best results. The findings, including the accuracy and the AUC, are presented in Tables 3 and 4. Popular machine learning algorithms that only use boosting do relatively well in accuracy, except LR. In particular, SVM attains the lowest degree of accuracy (88.24%) possible. Random forest obtains a great level of accuracy, above 98%, when employed as a bagging algorithm in conjunction with other types of algorithms that boost. This successful model performance is achieved by incorporating ensemble learning on employ attrition datasets compared to other related works.

5. Conclusion

This paper identifies and analyzes the prediction shift issue in all existing gradient boosting implementations. We suggest an inclusive solution to the problem: ordered boosting using orders. Both parameter and machine learning techniques are used in this work to generate an appropriate prediction model for employee attrition. According to the results of our research, females had a higher attrition rate than males by 0.659 times, while married and divorced people had an attrition rate of 0.427% and 0.30%, greater than those seen in single people, respectively. In addition, the attrition rate among frequent travelers was two and a half times greater than among occasional travelers. Those who only travel occasionally had a lower rate of attrition. Using the dataset to make predictions about employee turnover, we trained with 70%, tested with 30%, and recorded the accuracy of the test set. This allowed us to evaluate the effectiveness of the model we used to predict employee attrition. The proposed CatBoost outperformed with an accuracy of 95.84%, Recall of 94.66%, and consumes a time of 2.15 sec related to SVM, RF, and LR melds. This demonstrates that CatBoost provided a better fit for our dataset and performed better when making predictions.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Alduayj, Sarah & Rajpoot, Kashif. (2018). Predicting Employee Attrition using Machine Learning. 93-98.

10.1109/INNOVATIONS.2018.8605976.

[2] Sexton, Randall & McMurtrey, Shannon & Michalopoulos, Joanna & Smith, Angela. (2005). Employee turnover: A neural network solution. Computers & Operations Research. 32. 2635-2651. 10.1016/j.cor.2004.06.022.

[3] Ashworth, Michael. (2006). Preserving knowledge legacies: Workforce aging, turnover and human resource issues in the US electric power industry. International Journal of Human Resource Management. 17. 1659-1688.

10.1080/09585190600878600.

[4] Droege, S. B. & Hoobler, M. J. 2003. Employee Turnover And Tacit Knowledge Diffusion: A Network Perspective. Journal of Managerial Issues 15. 50-64. https://www.jstor.org/stable/40604414

[5] J. Rohan, A. Shahid, S. Saud, and J. Ramirez, "IBM HR analytics employee attrition & performance," January 2018 [Online]http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html#business_problem

[6] Raja D V A J and Kumar R A R 2016, A Study to Reduce Employee Attrition in IT Industries. International Journal of Marketing and Human Resource Management (IJMHRM) 7(1) 1-14

[7] Bindra, Harlieen & Sehgal, Krishna & Jain, Rachna. (2019). Optimisation of C5.0 Using Association Rules and Prediction of Employee Attrition: Proceedings of ICICC 2018, Volume 2. 10.1007/978-981-13-2354-6_3.

[8] Van Reenen, J. Human resource management and productivity. In Handbook of Labor Economics; Elsevier: Amsterdam, The Netherlands, 2011.

[9] Deepak, K.D.; Guthrie, J.; Wright, P. Human Resource Management and Labor Productivity: Does Industry Matter? Acad. Manag. J. 2005, 48, 135–145.

[10] Keramati, A.; Jafari-Marandi, R.; Aliannejadi, M.; Ahmadian, I.; Mozaffari, M.; Abbasi, U. Improved churn prediction in telecommunication industry using data mining techniques. Appl. Soft Comput. 2014, 24, 994–1012

[11] Fallucchi, Francesca & Coladangelo, Marco & Giuliano, Romeo & De Luca, Ernesto. (2020). Predicting Employee Attrition Using Machine Learning Techniques. Computers. 9. 86. 10.3390/computers9040086.

[12] Srivastava, Dr. Praveen & Eachempati, Prajwal. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi- Criteria Decision-Making Approach. Journal of Global Information Management. 29. 1-29.

[13] 10.4018/JGIM.20211101.oa23.

[14] Khera, Shikha & Divya,. (2019). Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. Vision: The Journal of Business Perspective. 23. 12-21. 10.1177/0972262918821221.

[15] Mansor, Norsuhada & S Sani, Nor & Aliff, Mohd. (2021). Machine Learning for Predicting Employee Attrition. International Journal of Advanced Computer Science and Applications. 12. 435-445.

[16] 10.14569/IJACSA.2021.0121149.

[17] El-Rayes, Nesreen & Fang, Ming & Smith, Michael & Taylor, Stephen. (2020). Predicting employee attrition using tree-based models. International Journal of Organizational Analysis. ahead-of-print. 10.1108/IJOA-10-2019-1903.

[18] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. 2017a.

[20] Li Wei, Machine Learning in Fraudulent E-commerce Review Detection , Machine Learning Applications Conference Proceedings, Vol 2 2022.

[21] Harris, K., Green, L., Perez, A., Fernández, C., & Pérez, C. Exploring Reinforcement Learning for Optimal Resource Allocation. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/155

[22] Rajesh, N. (2022). Stock price prediction using hybrid deep learning technique for accurate performance. Paper presented at the IEEE International Conference on Knowledge Engineering and Communication Systems, ICKES 2022, doi:10.1109/ICKECS56523.2022.10060833 Retrieved from www.scopus.com