Grenze International Journal of Engineering and Technology, January Issue

Employee Attrition Rate Prediction - Using Machine Learning

Ramalakshmi K[1], Berin Jeba Jingle I[2], Shirley C P[3], Aishwarya P[4], Joyce Beryl Princess P[5] and Vidhya K[6]
[1]Department of Computer Science and Engineering, Alliance University, Bangalore
Email: ramalakshmi.k@alliance.edu.in
[2-4]Dept. of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: berinjeba@karunya.edu , shirleydavidlivingston@gmail.com , aishwaryap@karunya.edu
[5]Dept. of AI & ML, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: joyceberyl@karunya.edu
[6]Dept. Of DS & CS, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: joyceberyl@karunya.edu

***Abstract*—Today, predicting employee attrition has become a significant issue for businesses. Organizations face a significant problem with employee attrition, particularly when highly skilled, knowledgeable, and important personnel depart in search of better opportunities elsewhere. The cost of replacing a trained employee is incurred as a result. Hence, we analyze data from both current and past employees to understand the prevalent causes of employee turnover. Employing established classification methods like decision trees, logistic regression, SVM, KNN, random forest, and naive Bayes, we leverage human resource data to proactively address employee attrition. Our approach involves implementing feature selection techniques on the data and scrutinizing the outcomes to mitigate the risk of employee loss. Businesses can grow more profitably by reducing the cost of their human resources when they can predict employee turnover.**

***Index Terms*— Employee Attrition; Machine Learning; Intrusion Detection; Anomaly Detection; Network security, Random Forest; SVM; KNN, Data Mining Algorithm.**

I. INTRODUCTION

Attrition among employees is a serious issue that today's HR managers must deal with. The expense of training personnel is increased by attrition in the BPO industry, which also hurts productivity and both the organization's and its employees' "knowledge maturity" levels. Businesses that cannot maintain long-term relationships with their employees will eventually disappear from the market. The stabilization of business operations and the abundance of skilled and experienced labor that may provide a long-term solution to the industry's high turnover rate are key components of the BPO solution.

The call center portion of the BPO industry, in particular, has a high attrition rate due to a lack of job security and advancement opportunities. The quality of the workforce and cost-effectiveness were the two key factors behind the expansion of the BPO sector. There are countless reasons why an employee can decide to leave their company, but these reasons differ depending on the nature of the firm. In the BPO sector, employees take into account a variety of comfort levels when choosing an officer, and the position of the company in the market, over the past few years, industrialists have become aware of the effects of staff attrition pay, room for growth,

Grenze ID: 01.GIJET.11.1.3_1

work environment, and other pertinent aspects can all have a big impact on your choice. stress brought on by the demands of the job and workload, coworkers, the marketability of current roles, and most importantly, future stability with the company.
In the business sector, employee churn is a difficult problem to solve. It significantly affects a company's ability to compete. According to the Ministry of Labor and Employment's Report on Absenteeism, Labour Turnover, Employment, and Labor Cost, the attrition rate in India's automobile industry is 40%. Chronic organizational issues are caused by employee attrition. Industrialists have been aware of the effects of staff attrition during the past few years.

II. Literature Survey

This section contains all of the literature that is required for this paper's purpose. This comprises an explanation of how human resource management can affect organizations, an account of how machine learning is applied in human resources, and an outline of the nature.

A. Establishing a Competitive Advantage is Largely Dependent on Human Resource Management

Being able to outperform competitors is a goal shared by all organizations. HRM, or human resource management, is one crucial area to focus on if you want to acquire a competitive edge [1], there is evidence to suggest that an employee's level of happiness and satisfaction has a direct impact on their engagement and performance, which in turn affects performance within the organization's competitive advantage can be enhanced through effective human resource management (HRM) as it can positively affect employees and enhance their performance [2].

B. Employee Turnover

The primary objective of Human Resource Management (HRM) is to uphold employee satisfaction. Research conducted by Ryan demonstrated a clear association between employee satisfaction and turnover, revealing that higher satisfaction levels are correlated with reduced turnover rates.
Numerous empirical studies, including those by have highlighted the significant impact the effect of staff turnover on the efficacy of an organization, have a direct bearing on competitive advantage [3]
Employee turnover has a variety of effects on how effectively an organization operates. The organization primarily loses both implicit and explicit knowledge when an employee leaves, which lowers organizational knowledge overall and may harm productivity. Secondly, the increased workload on remaining employees can result in diminished efficiency. Thirdly, employee turnover can adversely affect company morale. Lastly, lower turnover rates translate to reduced hiring, thereby lowering costs associated with training and the recruitment process, as noted by[4].
Employee turnover effects encompass both tangible and intangible aspects. A comprehensive understanding of these effects and their impact on competitive advantage necessitates considering both dimensions. Although certain employee turnover can be beneficial, in most cases, turnover proves to be costly and disrupts workflow. For instance, in the U.S. fast food industry, characterized by high turnover, the annual retraining costs for new employees alone amount to a substantial
$4.3 billion [3][5].
To address and mitigate employee turnover, the initial crucial step is understanding its underlying causes. This knowledge empowers organizations to take targeted actions, offering a strategic approach to minimizing the impact of this phenomenon[6].

C. Monitoring of Retention

As a subset of HRM, retention management is concerned with finding ways for companies to hold onto their staff members for longer shifts. As highlighted by], organizations must not only understand but also assess their turnover situation to effectively implement retention management strategies. Simply being aware of turnover is insufficient; organizations need a strategic retention plan that delineates specific actions to retain their workforce before taking any measures[6][7]. Studies indicate that companies with well-executed retention strategies are more likely to retain their current employees, enhancing their attractiveness to potential hires. It is emphasized that organizations should prioritize retaining their top-performing employees, as neglecting to do so may lead to the departure of key talents. This underscores the growing importance of retention management within HRM and its potential effect on advantages in the competition[8].

D. Reasons Why People Continue In Their Current Positions

Moving voluntarily between organizations can have advantages, but it can also present difficulties and obligations for the individual. Over the years, several researchers have examined the reasons behind employees switching from one organization to an additional put it simply, the general answer is yes. Employee turnover is common if they are dissatisfied with their position and show little loyalty to the company[9].

E. The Application of Machine Learning in HRM

It took some time for machine learning to become established in the field of human resources, but]has pointed out that it has recently gained traction. According to a 2018 LinkedIn report, only 22% of businesses have started integrating analytics into their HR departments, and it's still unclear how effective these analytics will be. While big data, such as product sales, is readily accessible for industries like marketing, there are challenges in utilizing machine learning (ML) in HR[10]. Human resources typically have smaller datasets compared to fields like data science, and sometimes no data storage at all Decisions in HR, such as hiring and firing, carry significant consequences Sutherland [9][11].

Several challenges arise in applying machine learning in HR. Firstly, assessing individual performance is complex in roles that heavily depend on teamwork, making it difficult to distinguish individual and group contributions. Additionally, social and psychological ties among coworkers are factors considered in hiring decisions, alongside outward appearance). The last is machine learning algorithms need training, and the training data can influence the algorithm's characteristics. For instance, if an algorithm trained on the existing workforce reflects a majority of white men, it might introduce bias in favor of this group when seeking new applicants. An example of this issue occurred when Amazon faced bias concerns in its hiring model, leading to the explicit removal of gender as a criterion, yet the model continued to exhibit bias toward specific groups.

F. Algorithms Decision-Making

The use of algorithms in decision-making is becoming more widespread in both personal and professional spheres. Frequently, these choices have the potential to significantly impact one's life, HR decisions may involve job performance evaluation, selection of candidates for employment, and screening of applicants for advancement in the process. Let an algorithm make decisions for us, even though algorithmic decision-making is generally advantageous [9][10]. Enumerating thousands of job applications in a matter of seconds allows algorithms to quickly determine which candidates are qualified. When considering the time required for a human to review each application individually, it is evident that utilizing an algorithm would result in significant time savings. Imagine the scenario where the algorithm automatically eliminates the top candidates in the same example. The use of algorithmic decision-making in transformative services, which have the power to significantly impact and change people's lives, has the potential for severe consequences. Decisions that affect people and their families have an effect on organizations as well as on a personal level

Services like healthcare, education, and senior care are examples of transformative services. The literature that is required for this paper's purpose is presented in this section. This comprises an explanation of the influence that human resource management can have on organizations, an account of how machine learning is applied in human resources, and an outline of machine learning's typical applications[11].

III. METHODOLOGY

Features of Dataset :

Ten features are computed from each one of the cells in the Sample which are as follows:

1. Age
2. Business Travel
3. Daily Rate
4. Department
5. Education
6. Education Field
7. Employee Count
8. Employee Number
9. Job Roles
10. Job Satisfaction

A. AdaBoost

AdaBoost, short for Adaptive Boosting, serves as a boosting meta-algorithm in machine learning, offering the theoretical capability to significantly decrease the error of any learning algorithm consistently generating classifiers with suboptimalperformance. It operates as a nominal class classifier, specifically designed for nominal class problems, utilizing the Adaboost ensemble method. While AdaBoost often leads to substantial performance enhancements, there is a risk of occasional overfitting.

However, every occurrence in the training dataset has a weight. Equation (1) describes the initial weight value that is set.

Here, 'xi' indicates the corresponding training instance, and 'n' indicates the total number of training instances.

$$Weight(xi) = 1/n \qquad \text{-------} \quad (1)$$

B. Random Forest

The popular Random Forest machine learning algorithm, whichcan be used for both regression and classification issues, is a component of the supervised learning methodology. The ensemble learning method, which brings together multiple classifiers to address complex problems and enhance overall model performance, is the foundation of this algorithm. The Random Forest classifier is made up of several decision trees that are built using various subsets of the provided dataset. Thealgorithm attempts to increase the dataset's predictive accuracy by averaging the

of depending on just one decision tree, the Random Forest aggregates predictions from several trees and chooses the best course of action based on the majority of these forecasts.

C. Gradient Boosting

Within sci-kit-learn, the gradient boosting regression class is a widely used technique known as gradient boosting. This approach involves each predictor refining the errors of its predecessor. In contrast to Adaboost, the training instances' weights remain unchanged for each predictor. Instead, the modelutilizes the residual errors from the previous predictor as labels for training. Gradient Boosted Trees is a specific method within this framework, employing CART (Classification and Regression Trees) as its base learner. This technique makes use of a crucial parameter called shrinkage. A tree in an ensemble issaid to have shrunk when its prediction is increased by the learning rate (eta), a variable that ranges from 0 to 1. We call thisphenomenon shrinkage. To attain a specific level of model performance, the ratio of estimators to eta needs to be balanced;a decrease in learning rate necessitates an increase in the number of estimators. Now that every tree has been trained, predictions are feasible. Every tree predicts a label; the final forecast is given by the formula (2).

$$(eta * r1) + (eta * r2) +..+ (eta * rN) = y(pred) \qquad \text{---------} \quad (2)$$

Regressor with Gradient Boosting. Gradient Boosting Classifier is a similar classification algorithm that is similar.

D. Decision Tree

Supervised learning techniques, such as decision trees, are commonly employed to address classification and regression problems, for which they are particularly well-suited. A decision tree is a type of tree structure classifier in which the decision-making process is represented by branches, internal nodes represent features in thedataset, and leaf nodes indicate the classification outcome.

Decision trees are composed of two node types: decision nodes and leaf nodes. Decision nodes are used for choices and have several branches, while leaf nodes, which represent the final decisions made, have only one branch. The characteristics included in the provided dataset are the basis for the decision-making process or tests.

This graphical representation explores all possible solutions or choices within predetermined parameters, commencing with theroot node and expanding through subsequent branches in a tree-like structure, resembling an actual tree. Hence, it is aptly nameda decision tree. The construction of the tree is facilitated by the CART algorithm, denoting the Classification and Regression Tree algorithm.

E. XG Boost

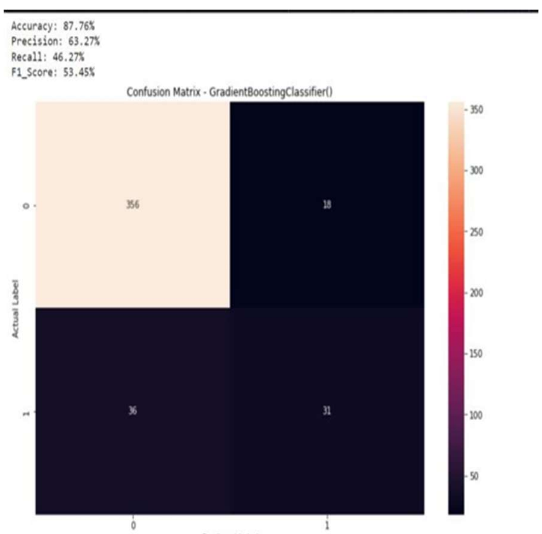
XG Boost is well-known for its key components, which include Newton Boosting, proportionate leaf node shrinking, astute tree penalization, and an additional randomization parameter. It can operate on a single machine and distributed processing frameworks like Dask, Flink, Hadoop, and Spark. It is optimized for distributed computing. Gradient boosting implementations other than XG Boost have been surpassed in speed and performance, making it the preferred choice in applied machine learning and Kaggle competitions for structured or tabular data.

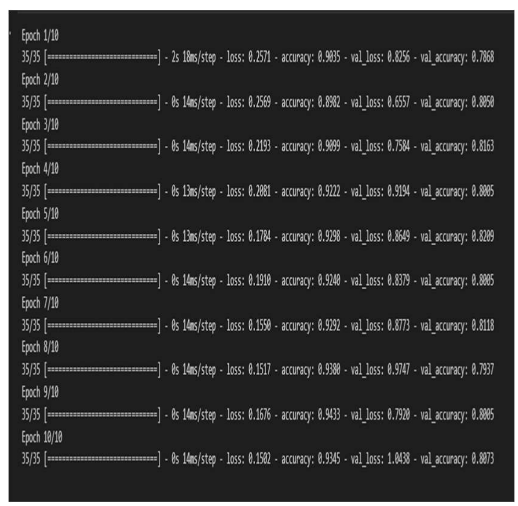
F. CNN Algorithm

Among the deep learning systems, Convolutional Neural Networks (CNNs) are particularly good at processing and identifying images. Convolutional, pooling, and fully connected layers are among the layers in this architecture. The convolutional layers of a CNN are essential because they use filters to extract features from the input image, such as edges, textures, and shapes. Following the convolutional layers, the information output is subjected to pooling layers, serving to downsample feature maps. This process reduces spatial dimensions while maintaining important information. After that, the output of the pooled layer is subjected to one or more fully connected layers to aid in the classification or prediction of images. CNNs are trained using extensive datasets of labeled images to recognize patterns and features linked to particular objects or classes. Beyond image classification, CNNs can also be utilized for tasks like object detection and image segmentation by extracting relevant features.

IV. RESULT AND DISCUSSION

By employing RandomizedSearchCV, we achieved a notable improvement in the accuracy score of the Gradient Boosting Classifier, reaching 87.76%, which stood out as the highest among all models. However, the initial Ada Boost Classifier model maintained the top recall score at 62.69%, accurately predicting a substantial number of employees likely to leave while maintaining a commendable accuracy score of 85.53%. Despite efforts to fine-tune the Ada Boost model, resulting in an 83.58% recall score, there was a noticeable trade-off with accuracy. This adjustment led to a significant increase in false positives, suggesting a potential bias in the algorithm towards consistently identifying employees as likely to leave Fig 1 and 2.





```
Epoch 1/10
35/35 [==============================] - 2s 18ms/step - loss: 0.2571 - accuracy: 0.9035 - val_loss: 0.8256 - val_accuracy: 0.7868
Epoch 2/10
35/35 [==============================] - 0s 14ms/step - loss: 0.2569 - accuracy: 0.8982 - val_loss: 0.6557 - val_accuracy: 0.8050
Epoch 3/10
35/35 [==============================] - 0s 14ms/step - loss: 0.2193 - accuracy: 0.9099 - val_loss: 0.7584 - val_accuracy: 0.8163
Epoch 4/10
35/35 [==============================] - 0s 13ms/step - loss: 0.2081 - accuracy: 0.9222 - val_loss: 0.9194 - val_accuracy: 0.8005
Epoch 5/10
35/35 [==============================] - 0s 13ms/step - loss: 0.1784 - accuracy: 0.9298 - val_loss: 0.8649 - val_accuracy: 0.8209
Epoch 6/10
35/35 [==============================] - 0s 14ms/step - loss: 0.1910 - accuracy: 0.9240 - val_loss: 0.8379 - val_accuracy: 0.8005
Epoch 7/10
35/35 [==============================] - 0s 14ms/step - loss: 0.1550 - accuracy: 0.9292 - val_loss: 0.8773 - val_accuracy: 0.8118
Epoch 8/10
35/35 [==============================] - 0s 14ms/step - loss: 0.1517 - accuracy: 0.9380 - val_loss: 0.9747 - val_accuracy: 0.7937
Epoch 9/10
35/35 [==============================] - 0s 14ms/step - loss: 0.1676 - accuracy: 0.9433 - val_loss: 0.7920 - val_accuracy: 0.8005
Epoch 10/10
35/35 [==============================] - 0s 14ms/step - loss: 0.1502 - accuracy: 0.9345 - val_loss: 1.0438 - val_accuracy: 0.8073
```

Fig 1:Data Loss

V. CONCLUSION

It is feasible to anticipate if an employee will depart from the company due to attrition concerns. The organization can use this information to determine which employees are most likely to leave and provide them with restricted incentives. Furthermore, there can be "false positives," which happen when HR predicts that an employee will leave the company shortly but the employee decides to stay. Even though these mistakes are expensive and inconvenient for human resources as well as employees, relationships are better off as a result. However, it may also be a false negative if a human resources department fails to promote or elevate staff members and they ultimately leave the organization.

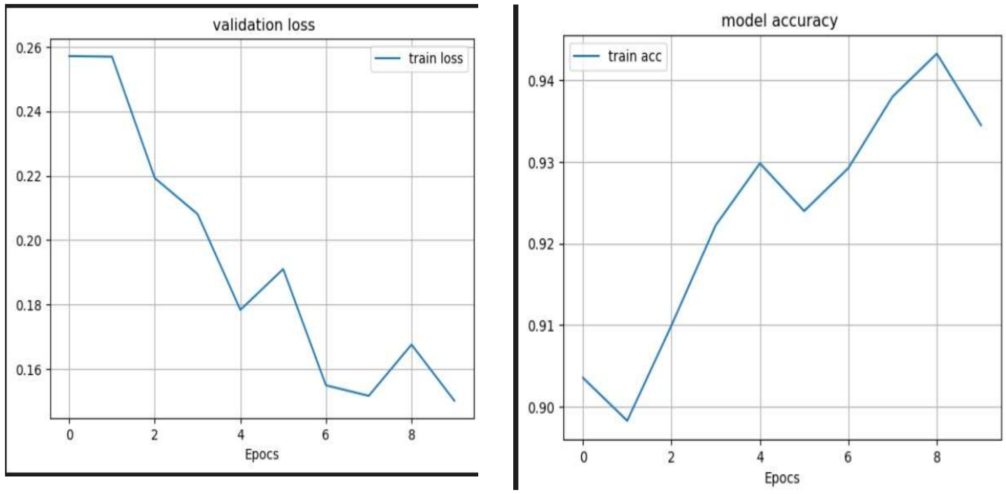


Fig 2: Graph of Data Loss

Due to this human resources issue, the company not only runs the risk of losing a worker but also pays more for finding, developing, and employing a replacement. The kind of treatment can be divided into different categories according to the employee income range, depending on the circumstances. If an employee gets paid more than they should, that could be considered unfair treatment. It is imperative to take the treatment plan's cost into account. Human judgment is a factor in employee attrition prediction, and this study's human resources dataset has been subjected to several machine-learning approaches. The results unequivocally show that Random Forest (RF) performs better than other techniques.

REFERENCES

[1] Holtom, Brooks & Mitchell, Terence & Lee, Thomas & Inderrieden, Edward. (2005). Shocks as causes of turnover: What they are and how organizations can manage them. Human Resource Management. 44. 337 - 352. 10.1002/hrm.20074.

[2] Kinnie, Nicholas & Hutchinson, Sue & Purcell, John & Rayton, Bruce & Swart, Juani. (2006). Satisfaction with HR Practices and Commitment to the Organisation: Why One Size Does Not Fit All. Human Resource Management Journal. 15. 9 - 29. 10.1111/j.1748-8583.2005.tb00293.x.

[3] Raza A, Munir K, Almutairi M, Younas F, Fareed MMS. Predicting Employee Attrition Using Machine Learning Approaches. Applied Sciences. 2022; 12(13):6424. https://doi.org/10.3390/app12136424

[4] Fallucchi F, Coladangelo M, Giuliano R, William De Luca E. Predicting Employee Attrition Using Machine Learning Techniques. Computers. 2020; 9(4):86. https://doi.org/10.3390/computers9040086

[5] Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting.

[6] Guerranti, F., & Dimitri, G. M. (2022). A Comparison of Machine Learning Approaches for Predicting Employee Attrition. Applied Sciences, 13(1), 267.

[7] Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2022). A review of machine learning applications in human resource management. International Journal of Productivity and Performance Management, 71(5), 1590-1610.

[8] Rustam, F., Ashraf, I., Shafique, R., Mehmood, A., Ullah, S., & Sang Choi, G. (2021). Review prognosis system to predict employees' job satisfaction using deep neural network. Computational Intelligence, 37(2), 924-950.

[9] Elanwer, R. (2021). HR analytics in practice: Using machine learning to predict employee absenteeism.

[10] Lata, K., & Garg, N. (2023). Predicting non-violent work behavior among employees using machine learning techniques. International Journal of Conflict Management, 34(5), 931-944.

[11] Chourey, A., Phulre, S., & Mishra, S. (2019). Employee attrition prediction using various machine learning techniques. The International Journal of Analytical and Experimental Modal Analysis, 11(11), 2718-2724.