IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Employee Attrition Prediction with Neural Networks and SHAP-Based Explainability

[1]Sanjivani Sharma, [2]Sania, [3]Pavan Singh, [4]Amit Kumar

[1]Student, [2] Student, [3] Student, [4] Assistant Professor

Department of Computer Science and Engineering (Artificial Intelligence),

IIMT Collage of Engineering, Greater Noida, India

Abstract: Employee attrition is a big problem for companies around the world, causing issues with how they run and manage their money. Traditional machine learning methods work to some extent, but they usually don't offer enough accuracy or clarity to be useful in making real-world decisions about human resources. This paper introduces a framework that uses a neural network and includes SHAP analysis to predict and explain why employees leave, based on the IBM HR Analytics dataset. The model uses specially designed features, balances the classes with SMOTE, and applies dropout regularization. The system's results show it is correct about 82% of the time, has a ROC-AUC score of 0.73, finds 57% of the actual cases, and has an F1-score of 50%. SHAP analysis shows that working overtime, how much money someone makes each month, and how happy they are with their job are the top factors that predict if someone will leave their job, giving HR managers useful information to take action.

Index Terms — Employee Attrition Prediction, Neural Networks, SHAP Explainability, Human Resource Analytics, Imbalanced Classification

I. INTRODUCTION

Employee leaving a company, either on their own choice or because they are forced to, is one of the biggest and most expensive problems that businesses today must deal with. Research shows that replacing one worker can cost a company anywhere from 50% to 200% of that person's yearly salary. This includes the costs of finding a new person, training them, and the loss of productivity while the new person gets up to speed [1]. Besides the money spent, high employee turnover harms how well a team works together, causes loss of important knowledge, and lowers the overall mood and motivation in the company. As global labour markets become increasingly competitive, the ability to proactively identify employees who are at risk of leaving has become a strategic priority for human resource (HR) departments.

This method serves as an instrument for tackling this issue. The swift incorporation of artificial intelligence (AI) into corporate decision-making has unveiled new opportunities for confronting this challenge [2]. Machine learning and deep learning have shown great ability to find complex, non-linear patterns in large, detailed data sets, which is something traditional statistical methods usually can't do as well. In the field of HR analytics, AI-powered predictive models can help managers make decisions based on data to retain employees, shifting from reactive approaches to proactive actions [5].

Even though many machine learning methods like logistic regression, decision trees, and ensemble techniques work well, there's still a big problem: these models are not easy to understand. When an HR professional sees an attrition risk score, they need more than just a prediction—they also want to understand why it was given. Even when predictions are correct, decision-makers might ignore or use

them wrongly if there are no clear explanations. This lack of understanding has led to more research on explainable AI (XAI) in the field of HR analytics.

This paper tackles both predicting and explaining employee turnover by introducing a feedforward neural network model. The model is trained on data from the IBM HR Analytics Employee Attrition dataset. To improve performance, the data is balanced using SMOTE, and relevant features are carefully selected based on domain knowledge. After the fact, explainability is made possible by using SHAP values, which break down each prediction into contributions from individual features. The main things this work contributes are: (i) creating a special type of neural network that is better at predicting when employees might leave, (ii) making new features that show hidden patterns in how employees behave and how stressed they feel at work, (iii) using SHAP to explain both overall and individual predictions in a way that helps HR make better decisions, and (iv) testing this approach on a well-known dataset to show it works well compared to other methods.

II. Literature Review

Over the last ten years, using machine learning to predict when employees might leave their jobs has become a big topic in both academic research and the business world. Early attempts mainly used classical statistical and simple machine learning methods. Setiawan et al. [9] used logistic regression on HR attrition data and found that factors like job satisfaction, work-life balance, and overtime status were statistically significant predictors. Logistic regression is easy to understand because you can look at the coefficients to see how each feature affects the outcome. However, it assumes that the relationship between the features and the outcome is straight-line, which means it can't handle more complicated, curved relationships between the features very well.

Decision tree models have also been widely studied and used. Taylor and their team showed that using tree-based methods is better than logistic regression when working with structured human resources data. This is because tree-based approaches are good at dealing with categorical variables and how different features interact with each other. Alduayj and Rajpoot [7] did a comparison of three methods—logistic regression, decision trees, and naive Bayes—on the IBM HR dataset. They found that decision trees gave better results in terms of accuracy, but they also noticed that decision trees had a problem with overfitting when there was an uneven distribution of classes. Usha and Balaji [8] also found that while decision trees gave good accuracy, they were easily affected by changes in the training data and needed a lot of pruning to work well in different situations.

Ensemble methods, particularly Random Forest and Gradient Boosting, have emerged as more robust alternatives. Pratt et al. [12] suggested using a random forest classifier to predict employee attrition and saw better results by reducing variance through the bagging method. Fallucchi and their team [3] did a detailed study comparing several machine learning methods like support vector machines, random forests, and gradient boosting on the IBM HR dataset. They found that using multiple models together, called ensemble methods, worked better than using just one model alone. Gabrani and Kwatra [14] looked into using machine learning to predict employee turnover and stressed how choosing the right features can help make the models work better and easier to understand.

Even with these improvements, a common issue in current research is that many studies don't pay enough attention to explaining how the models work. Most studies share overall performance numbers like accuracy, precision, and recall, but they don't explain why a model gives a specific risk score to a particular employee. Najafi-Zangeneh and their team [11] recognized this issue and suggested better machine learning approaches for predicting employee turnover. They pointed out that models that are too complex and hard to understand are not practical for use in real HR situations, especially where managers need to trust the results and follow legal rules. Yedida and their team [15] looked into different types of neural network setups for predicting employee turnover using the IBM data set, but they didn't go further to explain why their models made certain predictions after they were built. Das [10] emphasized that without ways to explain how the models work to people who aren't experts in technology, the real usefulness of HR analytics tools drops a lot.

The creation of SHAP (SHapley Additive exPlanations) by Lundberg and Lee [18] gave a solid, theory-based method that can be used with any type of machine learning model to explain its predictions.

SHAP is based on cooperative game theory and gives each feature a value that shows how much it contributes. This value follows important rules like being accurate locally, handling missing data, and staying consistent. Even though it's becoming more popular in areas like finance, healthcare, and detecting fraud, there isn't much research on using it for predicting employee turnover, especially when combined with deep learning methods. This paper aims to address this gap directly.

III. RESEARCH METHODOLOGY

A. Dataset

This method serves as the foundation for this study, using the IBM HR Analytics Employee Attrition dataset [16], a widely used benchmark in the HR analytics domain, is the real issue. This dataset includes 1,470 employee records described by 35 features, encompassing demographic attributes (e.g., age, gender, marital status), job-related variables (e.g., job role, department, years at company), compensation metrics (e.g., monthly income, stock option level), and attitudinal measures (e.g., job satisfaction, environment satisfaction, work-life balance). The binary target variable, Attrition, shows if an employee has left the company (positive class) or is still employed (negative class). The dataset has a big problem with class imbalance, where about 16% of the records are in the positive class, which is the attrition class. This makes it hard for regular classification methods to work well.

B. Data Preprocessing

The results from the confusion matrix, explained in words, show that the model correctly identified about 269 true negatives, which are employees who didn't leave and were correctly labelled as such, and 24 true positives, which are employees who did leave and were properly flagged. This method had approximately 18 false negatives were observed—cases where employees who left were incorrectly predicted to stay—alongside approximately 29 false positives, representing employees who were flagged as at-risk but did not ultimately leave. These numbers match the reported precision and recall values and show how hard it is to predict attrition when the dataset is small and not balanced.

C. Feature Engineering

To get a better understanding of ideas that aren't directly shown in the original data, four new features were created from the existing ones.

PromotionDelay is the gap between how long someone has been working at the company and how long it has been since they last got promoted, and it is calculated as:

$$\text{PromotionDelay} = \text{YearsAtCompany} - \text{YearsSinceLastPromotion}$$

This feature shows how much an employee's time with the company has gone beyond the recognition and promotion they've received, which is a predictor based on theory that can lead to disengagement and leaving on their own [4].

StressScore is made by combining different factors that show how hard someone is working and how unhappy they are with their environment:

$$\text{StressScore} = \frac{\text{OverTime\_encoded} + (5 - \text{WorkLifeBalance}) + (5 - \text{EnvironmentSatisfaction})}{3}$$

EngagementScore aggregates intrinsic motivational indicators**:**

$$\text{EngagementScore} = \frac{\text{JobInvolvement} + \text{JobSatisfaction} + \text{RelationshipSatisfaction}}{3}$$

StabilityIndex captures organizational embeddedness through tenure-related variables:

$$\text{StabilityIndex} = \text{YearsAtCompany} + \text{YearsWithCurrManager} + \text{YearsInCurrentRole}$$

These engineered features are added to the preprocessed feature matrix before model training, resulting in a final feature dimensionality of 54 input variables.

D. Class Imbalance Handling

To reduce the impact of having too many samples from the majority class, the Synthetic Minority Oversampling Technique (SMOTE) [17] is used only on the training data after splitting the dataset into training and testing parts. SMOTE creates new examples for the smaller group by mixing similar

examples from that group in the data space, which helps make the training data more balanced between the groups without using any information from the test data. This method is better than randomly removing some data, which would lose a lot of information from the larger group, and it's also better than just copying data, which doesn't add variety in the features.

E. Model Architecture

The suggested model is a type of neural network that has four layers, each connected completely to the next. The input layer takes in a vector with 54 different features. The first hidden layer has 128 neurons using ReLU activation, then it goes through Batch Normalization and a Dropout layer with a dropout rate of 0.4. The second hidden layer has 64 neurons that use ReLU activation, and then it goes through Batch Normalization and a Dropout layer with a dropout rate of 0.3. The third hidden layer has 32 neurons and uses ReLU activation. The output layer has one neuron that uses a sigmoid activation function, which gives a single probability value for the positive attrition class.

The sigmoid activation function at the output layer is defined as:

$$\sigma(z) = 1 + e - z1$$

where z represents the weighted sum of the activations from the last layer before applying the activation function. The goal of the training is to reduce the binary cross-entropy loss function, which is written formally as:

$$L = -N1\sum i = 1N\left[yi\log\left(y^{i}\right) + (1 - yi)\log\left(1 - y^{i}\right)\right]$$

where N denotes the total number of training samples, $y_i \in \{0,1\}$ is the true label for sample i, and $\hat{y}_i \in (0,1)$ is the predicted probability output by the sigmoid activation function. Batch normalization helps speed up the training process by adjusting the input to each layer based on the data in small groups, which decreases the problem of internal covariate shift. Dropout regularization helps prevent overfitting by randomly turning off some neurons during training, which pushes the network to spread out the learning across more neurons and avoid relying too much on any single one.

F. Training Process

The dataset, which has been preprocessed and enhanced with SMOTE, is divided into training and test groups using a stratified split of 80% for training and 20% for testing. This method keeps the same class distribution as the original data in the test set, ensuring the evaluation is fair and representative. The model uses the Adam optimizer, starts with a learning rate of 0.001, and processes data in batches of 32. The process uses early stopping with a patience of 15 epochs, which keeps track of the validation loss to stop training when performance starts to decline, helping to avoid overfitting and find the best time to stop. Model checkpointing saves the weights that correspond to the lowest validation loss, and these saved weights are later used to run predictions or make inferences. The training process usually finishes between 60 and 90 cycles, and the loss on the validation data becomes steady, matching the early stopping rule.

The training process can be explained like this: first, raw data is taken in and prepared through preprocessing and creating features. Then, the features are divided into two parts: one for training and one for testing. Next, SMOTE is used on the training part to balance the data. The neural network is then set up and trained step by step using mini-batch gradient descent, and the training stops early if there's no improvement. The best version of the model is kept for further testing and use.

G. Inference Pipeline

During inference, new employee records go through the same steps for cleaning and preparing data that were used when training the model, making sure the features are represented in the same way every time. The trained neural network creates a probability score $\hat{y}_i$ for each record. Instead of using the usual decision threshold of 0.5, a threshold of 0.40 is chosen after looking at the precision-recall curve from the validation set. This helps improve recall, which means it's better at identifying employees who might leave but are not wrongly labelled as staying. Employees who have a predicted probability higher than this threshold are marked for HR to act. After identifying the records that need attention, SHAP values are calculated to provide clear, detailed explanations for each individual case regarding the risk of attrition.

IV. RESULTS AND DISCUSSION

A. Performance Metrics

The trained neural network is tested on a separate set of 294 records to see how well it performs. This method achieves an overall accuracy of 82%, reflecting its ability to correctly classify the majority of test instances. The model has a problem with cement and gravel. However, because there's a natural imbalance in the classes, relying only on accuracy doesn't give a full picture of how well the model is performing. The accuracy for the attrition-positive class is 45%, which means out of all the cases predicted to be attrition-positive, about 45% are actually correct cases of attrition. The recall for the positive class of employees who leave is 57%, which means the model correctly finds 57% of all the actual cases where employees leave. The F1-score, which balances precision and recall in a balanced way, is 50%. This method has an area under the receiver operating characteristic curve (ROC-AUC) of 0.73, indicating a moderate to good discriminative capacity between attrition and non-attrition classes that substantially exceeds random classification (AUC = 0.50). The cement and gravel are the real issue, although the sand is in sufficient quantity.

B. Precision-Recall Tradeoff and Threshold Optimization

The precision-recall setup shows a purposeful design decision based on the uneven costs involved in predicting attrition. In this area, false negatives—missing an employee who later leaves—cost the organization more than false positives—incorrectly identifying a retained employee as someone needing action. So, the decision threshold is set to 0.40, which means it's more focused on recalling as many positive cases as possible rather than making sure each positive case is correct. This threshold selection is done by looking at the validation precision-recall curve, which shows how precision and recall change as the threshold value changes for all possible settings. The curve shows how the two measures relate in the opposite way: when the threshold gets lower, recall goes up but precision goes down. The chosen threshold of 0.40 is the point where the improvement in recall gives the most overall benefit compared to the cost of having more false positives in the organization.

C. ROC Curve Analysis

The ROC curve shows how the True Positive Rate, also called Recall, changes compared to the False Positive Rate as we adjust the threshold for classification. It shows that as the threshold gets lower, the model becomes more sensitive, but this also leads to more false positives. At the default threshold of 0.50, the model operates at a comparatively conservative operating point. Lowering the threshold to 0.40 shifts the operating point towards higher recall in the upper-left area of the ROC space. The ROC-AUC score of 0.73 means that if you pick two employees at random, one who left the company and one who stayed, the model will correctly predict that the one who left has a higher chance of leaving in 73% of cases. This performance is similar to what was found in other studies that used neural networks on the IBM HR dataset, and it shows a real improvement compared to a simple classifier that just predicts the most common class, which would have an AUC of 0.50.

D. Confusion Matrix Analysis

This method's test set reveals that among the approximately 47 true attrition-positive instances in the test partition (reflecting the 16% base rate), the model correctly identifies approximately 27 as attrition-positive (true positives) and misclassifies approximately 20 as non-attrition (false negatives). The confusion matrix shows that the cement and gravel are the real problem—there's enough sand, but the cement and gravel are the real issue. This method serves as a negative instance of attrition, the model correctly classifies around 214 as non-attrition (true negatives), while marking about 33 as false positives. Fortunately, the sand is in sufficient quantity. However, the real issue lies with the cement and gravel; there's a significant shortage there. This distribution shows that the model is good at correctly identifying the majority class and also picks up a useful share of the minority class that is leaving, with the false negative rate showing how hard it is to spot attrition when there's a big imbalance between the classes.

E. SHAP-Based Explainability Analysis

SHAP, which stands for SHapley Additive exPlanations, is used on the trained neural network through the KernelSHAP method. This approach treats the model like a black box and calculates Shapley values

by looking at different combinations of features. It gives a fair way to explain the model's predictions based on cooperative game theory principles.

This method serves as a feature importance tool, with the SHAP value ranking showing that OverTime is the single most influential predictor, displaying significant positive SHAP contributions for overtime workers. MonthlyIncome is the second most important factor, and higher income is linked to more negative SHAP values, showing that better pay helps protect against the outcome. Job satisfaction and environment satisfaction also have negative SHAP values when they are high. The specially designed features StressScore, PromotionDelay, and EngagementScore all show up as major factors, which supports the feature engineering approach based on domain knowledge. YearsAtCompany gives negative SHAP values to employees who have been with the company for a longer time, which matches what organizational embeddedness theory [4] suggests.

Local Explanation: For a typical high-risk case, having OverTime set to Yes adds a small amount of risk (+0.18), along with low MonthlyIncome (+0.12), high StressScore (+0.09), and low JobSatisfaction (+0.08). These factors together make the prediction higher than the average. However, having a moderately high number of years at the company slightly reduces the risk (−0.06).

The top HR actions that should be focused on, according to the analysis, are reducing overtime, comparing pay fairly with industry standards, and offering specific career growth opportunities. These results match what has been found in previous studies [3], [7], [12], but SHAP offers more accurate insights at the individual case level compared to the standard feature importance methods used in tree-based models.

F. Comparison with Previous Studies

This method's performance profile is contextually competitive with prior work. Fallucchi and their team [3] found that traditional machine learning models had accuracy ranging from 78% to 86% on similar datasets, but they didn't often mention the recall rate for the less common class. Alduayj and Rajpoot [7] got 88% accuracy using Random Forest, but they found that the recall for the attrition-positive cases was much lower. This shows that the high accuracy was mainly because the model correctly identified the majority of non-attrition cases. Najafi-Zangeneh and their team [11] found better results when using an ensemble approach, but they didn't include methods to explain how the results were reached. This study stands out not just because of its strong prediction results, but because it carefully combines SHAP-based explainability in a structured way. This approach tackles a key problem that has been pointed out in previous research, making the model better suited for clear and open HR decisions.

V. CONCLUSION

This paper introduced a full framework for predicting employee turnover that combines a regularized feedforward neural network with techniques to handle imbalanced data using SMOTE, features created based on domain knowledge, optimization of decision thresholds, and explanations of the model's predictions using SHAP. When tested on the IBM HR Analytics dataset, the proposed model shows an accuracy of 82%, precision of 45%, recall of 57%, an F1-score of 50%, and an ROC-AUC of 0.73. The balance between precision and recall shows how hard it is to predict employee turnover when the data is not balanced, but using a strategy to adjust the threshold helps focus more on recalling potential leavers. This reduces the costly mistakes of missing out on someone who might leave, which weakens the usefulness of these prediction systems in real-world situations.

The main part of this work is combining SHAP explainability with the neural network's prediction process, which changes the model's results from unclear probability scores into clear and useful explanations. A worldwide study using SHAP analysis shows that factors like OverTime, MonthlyIncome, JobSatisfaction, and StressScore are the main reasons employees leave. Additionally, local explanations help explain why each employee might be at risk, allowing HR professionals to communicate more effectively with individuals. This two-part system brings together accurate predictions and clear explanations after the fact, directly tackling the lack of transparency that has been a big issue in previous studies. It also helps make AI-based systems for predicting employee turnover more useful in real-world situations.

VI. RESULTS AND DISCUSSION

Several directions merit future investigation. The framework needs to be tested on bigger and more varied sets of data from different organizations to see if it works well beyond the IBM HR example. Second, connecting with real-time HR systems would allow for ongoing tracking of employee turnover and timely alerts about potential risks as employee data changes. Third, using advanced deep learning methods like Transformer models or graph neural networks, which can understand relationships in employee data, might lead to better predictions. This method serves as a tool to compare predicted models against actual attrition outcomes over time, offering rigorous validation of operational utility. The longitudinal study design tracks these predictions against real attrition outcomes over time, which would provide rigorous validation of operational utility. This method serves as an alternative explainability approach compared to SHAP, LIME and Integrated Gradients are still being evaluated. It would contribute to broader XAI literature in HR analytics contexts.

VII. ACKNOWLEDGMENT

The authors would like to express their heartfelt gratitude to the Department of Computer Science and Engineering at IIMT College of Engineering, Greater Noida, India, for offering the necessary computational resources and institutional support that helped in conducting this research. Special thanks are extended to my mentor, Mr. Amit Kumar, for his valuable guidance, helpful feedback, and consistent support throughout this work. The authors also appreciate the broader research community for making the IBM HR Analytics dataset available freely, which allows researchers to conduct studies that can be replicated and analysed within the field of HR analytics.

REFERENCES

[1] **[1]** M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," Business Horizons, vol. 61, no. 4, pp. 577–586, 2018

[2] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data—Evolution, challenges and research agenda," International Journal of Information Management, vol. 48, pp. 63–71, 2019.

[3] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020.

[4] J. M. Zelenski, S. A. Murphy, and D. A. Jenkins, "The happy-productive worker thesis revisited," Journal of Happiness Studies, vol. 9, no. 4, pp. 521–537, 2008.

[5] H. Varian, "Artificial Intelligence, Economics, and Industrial Organization," in The Economics of Artificial Intelligence: An Agenda, Chicago, IL, USA: University of Chicago Press, 2019, pp. 399–422.

[6] P. Vardarlier and C. Zafer, "Use of artificial intelligence as business strategy in recruitment process and social perspective," in Digital Business Strategies in Blockchain Ecosystems, Cham, Switzerland: Springer, 2020, pp. 355–373.

[7] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in Proc. Int. Conf. Innovations in Information Technology (IIT), Al Ain, UAE, 2018, pp. 93–98.

[8] P. M. Usha and N. Balaji, "Analysing employee attrition using machine learning," Karpagam Journal of Computer Science, vol. 13, pp. 277–282, 2019.

I. Setiawan, S. Suprihanto, A. Nugraha, and J. Hutahaean, "HR analytics: Employee attrition analysis using logistic regression," in IOP Conf. Series: Materials Science and Engineering, vol. 830, 2020, p. 032001.

[9] R. C. Das, "Conceptualizing the importance of HR analytics in attrition reduction," International Research Journal of Advanced Science and Hub, vol. 2, pp. 40–48, 2020.

[10] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, "An improved machine learning-based employees attrition prediction framework with emphasis on feature selection," Mathematics, vol. 9, no. 11, p. 1226, 2021.

[11] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," Baltic Journal of Modern Computing, vol. 9, no. 1, pp. 49–66, 2021.

[12] S. Taylor, N. El-Rayes, and M. Smith, "An explicative and predictive study of employee attrition using tree-based models," in Proc. 53rd Hawaii Int. Conf. System Sciences (HICSS), Hawaii, USA, 2020.

[13] G. Gabrani and A. Kwatra, "Machine learning based predictive model for risk assessment of employee attrition," in Int. Conf. Computational Science and Its Applications, Melbourne, Australia, 2018, pp. 189–201.

[14] R. Yedida et al., "Employee attrition prediction," arXiv preprint arXiv:1806.10480, 2018.

[15] IBM, "IBM HR Analytics Employee Attrition Dataset." [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/hremployee-attrition/

[16] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN), Hong Kong, 2008, pp. 1322–1328.

[17] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.