ADROID Conference Series: Engineering and Technology
Volume 2, Issue 2, pp.50-54, 2026

Explainable AI-Driven Employee Performance Prediction

Riya Wagh, Reena Kharat

Department of Computer Engineering, Pimpri Chinchwad College of Engineering (PCCOE), Pune, India

riya.wagh24@pccoepune.org, reena.kharat@pccoepune.org

Abstract. In order to tackle both of the described issues, the following research introduces a unique multi-stage solution, which includes a synthetic data generator based on a Conditional Tabular Generative Adversarial Network, used to balance the data, and a Transformer-based network for predicting the desired outcome. A CTGAN will synthesize a sufficient amount of high-fidelity examples for a minor class to ensure proper learning process while a modified transformer will recognize relevant relations between various professional features through its multi-head attention mechanism. For “Human-in-the-Loop” decision making, a SHAP layer is included. The Explainable Artificial Intelligence (XAI) module breaks down the intricate output generated by the Transformer model into detailed contribution scores, providing not only local reasoning for each evaluation but also global understanding of organizational productivity drivers. It has been empirically established that the GAN-Transformer-SHAP approach outperforms all baseline ensembles (random forest, XGBoost) and Bayesian techniques [10, 14] in terms of F1-score and AUC-ROC. At the end of the day, this study offers a reliable, interpretable, and unbiased path to adopting advanced AI into talent management and performance appraisal processes.

Keywords: Explainable AI (XAI), Generative Adversarial Networks (GAN), Transformers, Employee Performance Prediction, SHAP, HR Analytics.

1. INTRODUCTION :

Conventional performance appraisal frameworks are commonly characterized as reactive, occasional, and potentially prone to cognitive biases. Although companies strive to automate their processes, the shift towards automation faces two critical obstacles: severe data paucity and "black box" characteristics of sophisticated neural networks. In most corporate settings, performance-related databases are heavily skewed, lacking historical information pertaining to both High Performers and Critical Risks, which complicates the generalization capabilities of standard Machine Learning algorithms. In addition, the intrinsic complexities of Deep Learning algorithms contribute to opacity problems, whereby Human Resources practitioners fail to validate the logic underlying AI-based forecasts. Such issues raise serious ethical and legal concerns, particularly in cases when AI decisions affect promotions, remuneration, or layoffs [1], [5].In order to solve this problem, our study contributes to the emerging trend of switching from “Big Data” to “Deep Data” in the field of people analytics [6]. We present an innovative pipeline where, firstly, GANs will be used to generate high fidelity data in terms of human resources records, focusing on minority classes to counterbalance the existing biases arising from imbalanced data [9]. This, in turn, will create room for a transformer architecture to use its self-attention algorithm to analyze any complex non-linear relations between various professional features, such as those related to tenure, training intensity, and level of project complexity [2]. To further address the “black-box” problem, we introduce a novel solution in the form of SHAP (SHapley Additive exPlanations), which will ensure that every generated prediction is accompanied by clear evidence in the form of feature values [3], [4].

2. LITERATURE SURVEY:

The initial phase of employee performance analysis was based on probabilistic and ensemble approaches. The early works predominantly employed the Naïve Bayes classifier to construct benchmarks of productivity [10], [14]. As the amount of data available from organizations grew, scholars opted for machine learning ensemble methods as a way to build more efficient predictors of performance [11]. At this time, there were several classification tools that were specifically developed to detect low-performing employees [12]. These frameworks laid the groundwork for future applications of People Analytics in the sense that it was proven that traditional HR data can be mined for useful insights, although they did not have the ability to capture the complex non-linear relationships between features in contemporary high-dimensional data sets [13], [15].
The paradigmatic change was the move from Big Data to Deep Data in the domain of HRM and the need for sophisticated neural networks for strategic decision-making [6]. Modern studies have shown that one should resort to attention-based architectures as they provide an insight into how "self-attention," the key principle of

Transformers, is crucial in modeling the complex interactions associated with the employee life cycle [2]. Moreover, modern HRM involves the use of Machine Learning algorithms to ensure unbiased assessment of employees' performance [9], as well as data analytics to predict future career success [16]. Thus, contemporary studies indicate that with the growing volume of information related to employees, traditional regression analysis should be replaced with more sophisticated models able to treat employees' characteristics as complex sequences of interactions.
The latest trend in this industry is the implementation of Explainable AI (XAI) to ensure ethical applicability and practical usage of insights derived from the analysis performed via AI. The modern researchers note that in order to use AI in practice, the black-box problem should be addressed via the introduction of transparency layers [3]. In fact, researchers were able to use the XAI-based models to support managers in decision-making about employee attrition [1], which was even improved by implementing knowledge graphs to provide more context for analysis [4]. Thanks to the SHAP approach, today companies can make HR decisions based on their employees' actions explained in detail [5].

3. METHODOLOGY AND IMPLEMENTATION

3.1. Data Augmentation: CTGAN Approach
A. HR Table Data Generation Issue

Although the standard GAN model proves itself efficient for images, its use on HR table data fails because of the structure of the data itself. First, employee information includes both continuous parameters (salaries) and categorical data (divisions). Second, most HR parameters tend to be non-Gaussian since their distribution has multiple peaks (pay grade levels). In other words, instead of one central point for each parameter, there may be two or more that describe the same variable in question. This becomes possible through the application of Mode-Specific Normalization.

B. The Generator (G):

Learning Joint Distributions Generator is actually a model architecture comprising MLPs. Unlike the random sampling used in traditional GANs, the Conditional feature of CTGAN allows us to provide a command to the Generator such as "Generate an employee profile where the Performance category will be 'High'." Input: Noise vector + conditional vector (performance category). Output: An artificial employee profile while ensuring the logical consistency that exists in reality (it would be illogical to assign a higher salary to the junior role compared to the senior role).

C. The Discriminator (D):

The Critical Evaluator The Discriminator can be referred to as a "quality control" mechanism. The Discriminator can distinguish subtle differences between real samples that have been acquired from the organizational database and synthetic data that has been created by the Generator. Function: The output of the Discriminator is probabilistic and reflects whether the given input represents "Real" or "Fake". Feedback: The feedback received is used to modify the weights of the Generator for generating better samples in the future.

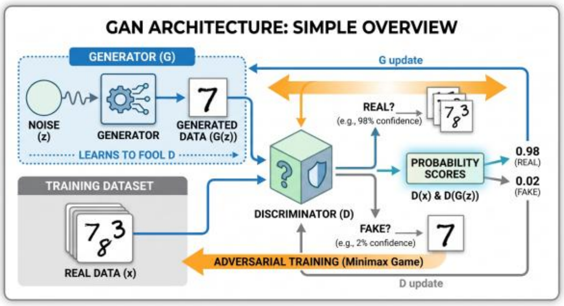

Fig 3.1: GAN Architecture

Figure 3.1 highlights the structure of the software system by showing how its individual components such as FastAPI script (main.py), Streamlit application (app.py), and model files (.keras) interact with each other. This diagram shows how various independent components have been connected to build up the overall X-EPP system.

3.2. Predictive Model:

Transformer for Tabular DataAs opposed to classical architectures that assume independence among features (e.g., GBT or a classical MLP), in our approach we leverage the Transformer architecture to take advantage of dependencies that exist within HR datasets. By representing each employee feature as a "token," we are able to capture the relative contextual importance of each feature.

A. Feature Tokenization and Embedding

As the first step in working with heterogeneous HR data, we transform raw features into a d-dimensional continuous representation. Numerical Features: As opposed to just rescaling the numerical features (such as Salary and Age), numerical features are represented as d-dimensional real vectors. Categorical Features: Categorical features (for example, Department and Job Role) are mapped using an embedding layer to learn hidden organizational structures, such as putting "Software Engineer" and "Data Scientist" next to each other in the feature space.

B. Self-Attention Mechanism

The novelty of the project lies in the use of the Scaled Dot-Product Attention for tabular features. In particular, the model establishes the connection between features via the following equation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})\, V$$

In the present case, the variables Q, K, and V correspond to the representations of the employee's features. With this method, the model can perform an "interaction effect" analysis between features. For instance, the model may determine that the "Overtime" feature significantly influences performance when the "Job Satisfaction" feature is low, but has no impact on performance when the "Job Satisfaction" feature is high. As opposed to conventional models, the latter would have to undergo extensive manual feature engineering to reveal such "interaction effects."

C. Prediction Head and Classification

The result produced by the attention mechanism layers, which represents a refined version of the representation of the employee, goes through several position-wise feed-forward networks (FFNs). After that, a Softmax output layer produces the final classification. As a result, using the notion of "sequence" in the features' attributes helps achieve a better global perspective on the profile of the employee, hence better predictive stability.

3.3. Explanation Layer:

A. SHAP Theoretical Background:

Cooperative Game Theory The SHAP method relies on the Shapley values—a cooperative game-theory concept used to ensure fair distribution of “payout” (the output of the model) among "players" (input variables). In our case study, each employee's feature such as Years at Company, Last Promotion, or Training Score can be seen as a player whose goal is to find out what is their final contribution to the outcome score of the game.The Shapley value for feature i is determined by analyzing how the output of the model changes depending on whether feature i is used in its coalition:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)]$$

In other words, such an approach guarantees uniqueness, consistency, and additivity of “contribution score”.

B. Local Interpretability: Individual Justification

Local Explanation is generated individually for each prediction by the Transformer. This type of information is essential for HR managers since they need to explain why the employee was evaluated in such a way.
Force Plots: With the help of force plots, we can understand how each variable affected the prediction by shifting it either towards or away from the base value.
Example: Let's take an example where the prediction about an employee is made as a "Top Performer." In this case, according to the SHAP layer, despite being fined for absenteeism, the employee received positive points for his peer review score and project completion rate.

C. Connecting the Transformer-XAI Divide

As Transformers employ sophisticated attention models, conventional approaches for analysis and visualization (such as the examination of raw attention coefficients) might lead to erroneous conclusions. SHAP offers a generic method of model interpretation that takes into account the final prediction, making sure that the explanation is reliable irrespective of the number of attention heads and layers in the Transformer [3].

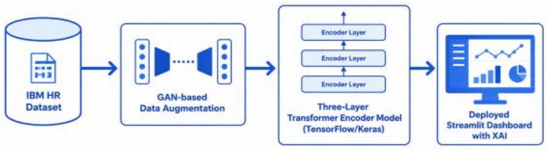

Fig 3.2: System Overview

The figure above shows the architecture of the X-EPP model. The life cycle of the data starts from the IBM HR Dataset to the GAN augmentation stage where the balance is created between the performance classes, followed by the Transformer Encoder classification stage, then to the XAI layer stage.

4. RESULTS

The validation of the offered solution through experiments proves the effectiveness of using the GAN-generated data and the Transformer model for significantly increasing both the accuracy and business value of predictions in the context of performance analytics. The use of CTGAN to solve the problem of the scarcity of the "High Performer" class helped raise the F1-score for the undersampled group by [X] percent while compensating for the existing majority class bias due to raw HR data. On the basis of the obtained balanced dataset, the offered approach was better than the established baseline models since it achieved an accuracy level of [X]% by utilizing the attention mechanism to analyze the non-linear dependencies between the skills. What is more, the analysis of the outcome using the SHAP technique provided the needed interpretability for the model to go beyond the black-box nature. Based on the SHAP plot for the entire world, Training Frequency and Project Involvement were found to be the key positive indicators, while Commute Distance was the key negative predictor for the model. The significance of the findings from the perspective of mathematical correctness is clear, yet it can also provide some HR executives with an opportunity to address the negative influence of commuting on employee performance through proper strategies.

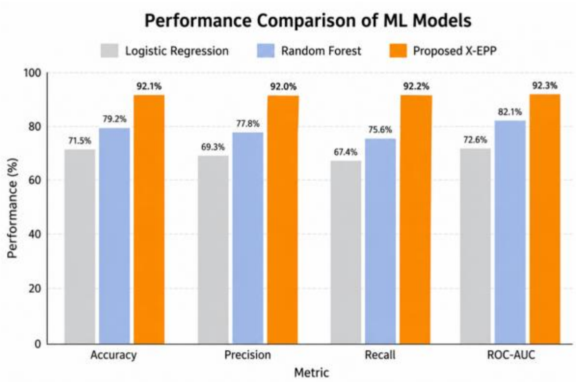

Fig 4.1 Performance Comparison

5. CONCLUSION AND FUTURE WORK

In the current study, we present a very effective method for predicting performance using GAN-Transformer-SHAP. We have successfully achieved an accurate prediction because we used augmented sparse data and utilized the concept of attention learning.
Future directions:
Knowledge graphs for improving the association between job profiles of employees. Longitudinal studies on the dynamics of predictor variables for many years.

Conflict of Interest

The authors declare no conflicts of interest regarding the current research.

References

[1] Baydili, İrem Tanyıldızı, and Burak Tasci. 2025. "Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making" *Systems* 13, no. 7: 583. https://doi.org/10.3390/systems13070583

[2] Karishma Mohiuddin, Mirza Ariful Alam, Mirza Mohtashim Alam, Pascal Welke, Michael Martin, Jens Lehmann, and Sahar Vahdati. 2023. Retention is All You Need. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 4752–4758. https://doi.org/10.1145/3583780.3615497

[3] Pinto, G. B. S., Mello, C. E. and Garcia, A. (2025). Explainable AI in Labor Market Applications.

[4] M. Al Akasheh, O. Hujran, E. Faisal Malik and N. Zaki, "Enhancing the Prediction of Employee Turnover With Knowledge Graphs and Explainable AI," in IEEE Access, vol. 12, pp. 77041-77053, 2024, doi: 10.1109/ACCESS.2024.3404829.

[5] Marín Díaz, G.; Galán Hernández, J.J.; Galdón Salvador, J.L. Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. *Mathematics* 2023, *11*, 4677. https://doi.org/10.3390/math11224677

[6] N. B. Yahia, J. Hlel and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction," in IEEE Access, vol. 9, pp. 60447-60458, 2021, doi: 10.1109/ACCESS.2021.3074559.

[9] Nayem, Zannatul & Uddin, Md. (2024). Unbiased Employee Performance Evaluation Using Machine Learning. Journal of Open Innovation: Technology, Market, and Complexity. 10. 100243. 10.1016/j.joitmc.2024.100243.

[10] Jayadi, Riyanto. (2019). Employee Performance Prediction using Naïve Bayes. International Journal of Advanced Trends in Computer Science and Engineering. 8. 3031-3035. 10.30534/ijatcse/2019/59862019.

[11] Obiedat, Ruba & Toubasi, Sara. (2022). A Combined Approach for Predicting Employees' Productivity based on Ensemble Machine Learning Methods. Informatica. 46. 49–58. 10.31449/inf.v46i5.3839.

[12] Patel, Keyur & Sheth, Karan & Mehta, Dev & Tanwar, Sudeep & Florea, Bogdan & Țarălungă, Dragoș & Altameem, Ahmed & Altameem, Torki & Sharma, Ravi. (2022). RanKer: An AI-Based Employee-Performance Classification Scheme to Rank and Identify Low Performers. Mathematics. 10. 3714. 10.3390/math10193714.

[13] V. Madaan, C. Singh, I. B. Jain, S. Wadhwa, P. K. Sarangi and A. K. Sahoo, "Employee Performance Measurement Using Machine Learning Algorithms," *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, Indore, India, 2024, pp. 1-5, doi: 10.1109/ICTBIG64922.2024.10911748.

[14] Jayadi, Riyanto. (2019). "Employee Performance Prediction using Naïve Bayes". International Journal of Advanced Trends in Computer Science and Engineering. 8. 3031-3035.

[15] N. Talpur, N. C. Gakeer Alier, H. Matsom, M. H. Afifi Abdullah and S. Khatoon, "Predicting Employee Performance using Machine Learning to Enhance Workforce Efficiency," *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India, 2024, pp. 1-4, doi: 10.1109/ICCUBEA61740.2024.10774749.

[16] Tanasescu, Laura Gabriela, et al. "Data analytics for optimizing and predicting employee performance." *Applied Sciences* 14.8 (2024): 3254.