Grenze International Journal of Engineering and Technology, June Issue

Multi-Model Explainable AI for Employee Attrition Prediction

Kunal Deshmukh[1], Arvind B. Patil[2] and Roshan S. Bhanuse[3]
[1-3]Yeshwantrao Chavan College of Engineering, Nagpur, India
Email: kunaldeshmukh2599@gmail.com, arbhagatpatil@ycce.edu, rsbhanuse@ycce.edu

***Abstract*— In this paper, I have provided a complete and explainable machine learning architecture to predict employee attrition, with a combination of standard classification models, gradient-boosting algorithms, and stacking ensemble. When experiments have been run on a large synthetic-augmented HR dataset, all methods have high success rates in prediction, with the best trade-off between accuracy, F1-macro, and ROC-AUC exhibiting by the stacking ensemble. The ROC and Precision-Recall curves also indicate that the ensemble is reliable in discrimination in the case of class imbalance, which is a requirement in the area of attrition prediction problems.**

In addition to the predictive ability, interpretability is identified as another essential intuition to use in the study in HR decision-making. According to the SHAP-based global elucidations, work stress, job satisfaction, work-life balance, managerial ratings, and variables related to compensation were the most significant turnover factors. To explain the individual level, LIME explanations justify particular predictions and assist the HR teams with the form of understanding the influences involved with high-risk results. Further, DiCE counterfactual analysis offers practical advice through example of how at-risk employees respond to workload, compensation, or managerial condition changes in meaningful ways to reduce the risk of attrition.

In general, the findings indicate that integrating the high-performing ensemble models with the clear interpretability techniques are able to facilitate a stable and pragmatic decision-support system to workforce retention. The suggested framework is proving to be highly predictive with providing the evidence-based insights and practical recommendations to support the proactive efforts of lowering the turnover rates and preserving organizational experience.

***Keywords*— attrition prediction, explainable AI, SHAP, LIME, DiCE counterfactuals, ensemble learning, HR analytics, imbalance correction, stacking, XGBoost, LightGBM.**

I. INTRODUCTION

In the modern competitive business world, companies operating in various industries find it more difficult to retain employees of expertise. With the increasing rate of work force movement, employee turnover has become a significant organizational problem, causing significant financial expenses, loss of productivity and lack of stability in operations continuity. The available literature shows that the turnover of employees has negative effects on employee team performance, knowledge retention, and long-term organizational performance [1]. To address these results, companies have put a stronger focus on the underlying factors behind attrition and formulation of proactive efforts to retain competent talent.

Grenze ID: 01.GIJET.12.2.303_22

Still, it is a complicated task to identify the group of employees who is most apt to leave an organization. HR teams must evaluate voluminous data on demographic, behavioral aspects and organizational factors, but the manual analysis does not necessarily help reveal the complex relations through which employees form their decisions. As Lepri et al. point out in [6], traditional methods of decision making are often less transparent and largely rely on subjective evaluation, which more often than not leads to a high probability of outlived assessment of employee departure risk. Without effective predictive frameworks, organizations experience challenges in adopting timely interventions that eventually increases the risk of turnover and lead to other losses in its operations.

In order to address these problems, machine learning (ML) methods have increasingly been applied in research studies, where large set of data could be analyzed and the accuracy of the attrition prediction increased. Interpretable models were mostly used in early research, including logistic regression and decision trees [12]. As much as these techniques provide predictive capabilities at a foundation, in most cases they do not exhibit the non-linear patterns of workforce. Recent innovations in ensemble learning, such as Random Forest, XGBoost and LightGBM, have improved results by modeling more interactions among features and modelling more complex-decision boundaries [13], [14].

Although these models demonstrate excellent predictive accuracy in HR data, Arrieta et al. in [5] observed that they often work as black boxes such that they do not give the human resource specialists a chance to comprehend the reasoning that underlie certain prediction.

Such absence of transparency creates significant hindrance to implementation of ML-based decision systems within the HR settings. Interpretability, fairness and accountability are essential in HR decision-making, especially since the outcomes of predictions affect sensitive processes like how promotions are done, how employees are intervened with, how work is distributed and how much to pay. As highlighted in [7], predictive models need to be explainable so as to be trusted and ethically applied when it comes to human-sensitive areas like hiring and retention. The lack of interpretability might make organizations hesitant to implement solutions based on advanced ML as they are afraid of biased results or black box predictions that can victimize staff or pose a legal liability.

Other than interpretability, employee attrition data is usually very class imbalanced, with the number of remaining employees usually vastly outnumbering the number who have exited. This can favour the majority category of employees on the ML models thus making the models ineffective in identifying truly high-risk employees. The studies of imbalanced learning, such as those conducted by Fernandez et al. [17] and the SMOTE method of Chawla et al. [18], demonstrate the need to use special techniques in order to obtain balanced and fair predictions. Nevertheless, the issue of imbalance treatment is often underdiscussed in most of the studies based on attrition and yields models that are reported to compute high accuracy without hinting at the minority-class cases of most interest to HR practitioners.

The other critical consideration is algorithmic fairness. The direct influence of HR analytics on the employees career development, opportunities, and working experiences is apparent. When used in human-centered contexts, as addressed in fairness-oriented research by Hardt et al. [32] and Bellamy et al. [33], the ML systems employed should be evaluated on the dimension of demographic bias to help in providing ethical and fair judgments. But current attrition prediction studies fail to undertake a systematic test of fairness, and therefore there is too little analysis of the possibility that differences exist.

Explainable AI (XAI) techniques have become the focus of more studies with respect to these issues. SHAP and LIME explanations [8] and the DiCE counterfactual explanations [10] are among methods that offer the interpretable justification of model outputs and so the HR specialists may be able to determine the factors that result in the risk of attrition at both individual and global level. There are previous studies which indicate that XAI increases stakeholder trust and reinforces evidence-based HR decision-making [28]. Although these are the merits, most studies are based on one approach to interpretability or lack integrated explanation and recommendations to be taken, which restricts the application of the study in practical settings of organizations.

All these restrictions show that there is the necessity of a single and readable framework that would provide valid information on prediction of attrition, address the class imbalance, be fair and offer a practical explanation. As the size of HR data increases and complexity of workforce behaviour grows, organisations need the systems that go beyond prediction to allow you to see clear, credible, and actionable insights. To fill in its shortcomings by the previous literature and in tune with the growing popularity of responsible AI in HR, the study will propose a multi-model explainable AI framework that would integrate the classical and ensemble-based ML methods with SHAP, LIME, and DiCE to achieve tiers of interpretability and provide a decision support interface that is ready to be deployed.

II. Related Work

The problem of employee attrition prediction has been progressively studied by researchers as organizations are subject to providing evidence-based solutions to decrease the turnover rate and enhance the stability of workforce. In the early days, people relied on conventional statistical models including Logistic Regression and Decision Trees to help people predict the presence of factors which affect turnover [1]. Although these models were easy to interpret, they were not able to model complex, non-linear relationships in recent HR data. To address these shortcomings, an improved model of ensemble-based machine learning, including Random Forest, Gradient Boosting, XGBoost, and LightGBM, was proposed, which showed significant improvement in its forecasting performance on structured organisational data [12], [13], [14]. These high-performing models, however, as Arrieta et al. point out as noted in [5], can be opaque in the sense that they work as a black-box system and it remains challenging to justify or trust algorithm decisions made by the HR practitioners.

Explainable Artificial Intelligence (XAI) has gained growing popularity due to the absence of transparency in machine learning. The development of such methods as SHAP [8], LIME [9], and DiCE counterfactual explanations [10] help to get an idea of how the individual predictions are made and thus HR stakeholders could learn how the demographic, performance, and workplace variable impacts contribute to attrition results. Previous research has revealed that the introduction of XAI enhances confidence in decision making, responsible AI practices, and helps organizations to develop action plan interventions [28]. Regardless of this development, most of the current research uses a single explanation technique, or does not lead to practical suggestions that HR departments can act upon.

Class imbalance is another significant issue raised in literature and in which there are far more retained employees than there are employees who leave. Algorithms (SMOTE [18]) and class weighting are often proposed to help to detect minority attrition cases. Nevertheless, not all studies put into account the aspect of fairness, despite the fact that the overall predictions of the Ml might unknowingly disadvantage certain groups of people. The evaluation frameworks of fairness such as Hardt, et al. [32] and Bellamy, et al. [34] emphasize that when training models, it is important to ensure fair model performance in terms of gender, age, job level among other attributes which are protected.

Despite the current methods having helped us to achieve a much better accuracy of attrition prediction, we are still lacking the opportunity to incorporate the high-performing ensemble models alongside with the ability to view models in multi-layers and explain their contribution to the results and its fairness in one whole model. The proposed work will address this gap in research through the integration of a multi-model learning pipeline and SHAP, LIME, and DiCE explanations, with fairness evaluation and deployment-ready reporting to support them.

III. Proposed Methodology

This section presents the dataset adopted in this study, the preprocessing procedures implemented, and the explainable machine learning approaches employed to deliver transparent and actionable attrition predictions. The proposed framework combines multi-model learning, fairness evaluation, and interpretability components to generate reliable predictions that are appropriate for real-world HR deployment

A. Dataset Description

This part outlines the dataset used in this paper, data processing tasks that have been applied and the interpretable machine learning methodologies that will be used to provide clear and practical predictions of attrition. This framework is a combination of multi-model learning, fairness assessment and interpretability elements to produce viable predictions which can be applied in the real HR world.

The dataset includes several behavioral, feedback, and sentiment-driven features to reflect experience in the workplace better. They are individual feedback score of peers, employee satisfaction score, survey sentiment score, work stress index, job satisfaction, work-life balance, training time, KPI performance, awards, and the frequency of remote-work. Also, workload stress variables, including status of overtime, overtime work hours, overtime ratio, and work hours per week, will be incorporated to identify the trend of burnout. These new feedback-based and sentiment-based indicators are integrated to fit into a more realistic picture of the role recognition and the workplace condition have on attrition.

This data has no value gaps to preprocess, thus making it easy to do so. Numerical attributes are based on realistic HR oriented distributions, and categorical variables maintain realistic organizational proportions. Even though data is synthetic, the dataset will have significant statistical relationships which are in line with the trends found in the Kaggle IBM dataset and other HR resources. As a result, it can be used to measure predictive models and predictive interpretability methods like SHAP, LIME, and DiCE.

B. Data Preprocessing

Since the dataset involves numerical, categorical, behavioral, and feedback-based factors, a systematic preprocessing pipeline was constructed to provide consistency and prepare modeling. Firstly, the feature names were standardized throughout to a standard lowercase set of naming with underscores so as to be generally readable during the analysis process.

Handling Missing Values

Although missing values are not present in the original IBM data set, they could also be introduced in the model by the use of the synthetic records. To ensure reliability, numbers variables are imputed by using the median of the variables of each variable and the categorical variables are imputed by the most common class.

Encoding of Categorical Attributes

The nominal variables that are to be converted into numerical values are Department, Business Travel, Education Field, Job Role, Marital Status, Gender and Over Time. As such, One-Hot Encoding use is implemented by a column-wise transformer, where each category is binary-indicated without ordinal assumptions.

Feature Scaling

To mitigate the effects of scale that are especially sensitive to models like the Logistic Regression, XG-Boost, and Light-GBM, numerical variables are normalized with Z-Score to prevent this issue. In this process, the numerical attributes are contributed proportional when optimizing the model.

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

Stratified Splitting and Cross-Validation

Since the target of attrition is skewed, stratified splitting is used to keep the proportions of classes constant in the training and validation sets. The stability of the model is then measured by Stratified K-Fold Cross-Validation and the total score calculated. This is a process that ensures less evaluation bias and enhances the reliability of estimated performance.

$$\text{CVScore} = \frac{1}{K} \sum_{k=1}^{K} \text{Score}_k$$

C. Machine Learning Models

The research uses multiple machine learning functions to try to uptake varied and non-linear trends linked to staff conduct and turnover. The reason behind using Logistic Regression is to create interpretable baseline in order to clearly understand the linear relationships among the dataset. The use of Random Forest is to represent more complex dependencies through a combination of decision trees. Also, XGBoost and LightGBM are high-performance gradient boosting algorithms that are used because of their efficiency in structured HR data. Both models are optimized systematically with RandomizedSearchCV in order to increase performance and minimize overfitting. To additionally increase predictive robustness especially on detecting cases of minority attrition; a Stacking Ensemble of all base models result in predictions is pooled, where the Logistic Regression assumes the role of meta-learner. Such a combined approach enhances trustworthiness and contributes to identification more accurate staff members at risk of leaving the company.

D. Interpretable Machine Learning Methods

Predicting attrition is often based on advanced machine learning to simulate a black-box system. Such models can be highly predictive, but since they are not as transparent as desired, it can only be used in a few situations in HR, where the reasons behind prediction have to be communicated. The proposed framework would combine three complementary features of explanation, such as LIME, SHAP, and DiCE, which will help shed light on the model decision-making process, as well as explain it on an individual and an international level.

Local Interpretable Model-Agnostic Explanations

LIME elucidates the prediction of a given employee by forming a straightforward, understandable surrogate model of the given instance. Instead of estimating the entire black-box model, LIME concentrates on one local region of perturbed examples of the samples achieved through minor modifications in the value of the employees features. The objective maximized by LIME is the following: here f is the original complex model, g is the

interpretable surrogate model, the weighting of perturbed samples is denoted by px, and the value Ω (g) is used to ensure that the surrogate model is not overly complex.

$$\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Shapley Additive Explanations (SHAP)

SHAP can be used to offer local and global interpretability through Shapley values, which are assigned to feature j, meaning the influence of feature j on the model prediction. Basing their Shapley values on cooperative game theory, Shapley values quantify the contribution of any given feature to the prediction given a set of other features. The SHAP value of feature j is calculated as; F the set of all features, S is a set of features and fs(x) is the prediction value obtained using set S.

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x) - f_S(x) \right]$$

DiCE Counterfactual Explanations

Whereas LIME and SHAP clarify about the causes behind an attrition prediction, DiCE (Diverse Counterfactual Explanations) concentrates on the way in which the prediction may be modified. DiCE can produce actionable what-if scenarios, i.e. by finding minimal and realistic modifications to features that can decrease predicted attrition risk. It is an optimization algorithm to get counterfactual instances x, where d(x',x) is the difference between the counterfactual and the actual employee profile, II(f(x)!= ytarget) is used to reinforce prediction of the desired outcome, and Div(x',C) is used to reinforce diversity and non-redundancy in the recommendations. Such counterfactual explanations allow HR managers to translate the model outputs into the form of feasible retention interventions.

$$\min_{x'} d(x', x) + \lambda_1 \cdot \mathbb{I}(f(x') \neq y_{\text{target}}) + \lambda_2 \cdot \text{Div}(x', C)$$

IV. Experimental Results

This section presents the experimental results of the suggested employee attrition prediction framework, that is the multi-model ensemble with an interpretability pipeline. The appraisal includes an analysis of classes distribution, validated performance metrics, end results of the models and outputs explainability created using SHAP, LIME, and DiCE.

A. Class Distribution

The visualization of the classes distribution (Fig. 1) demonstrates that there is a heavy imbalance between the active (label 0) and the resigned staff (label 1). In the entire synthetic-augmented dataset, the proportion of the instances of the class 0 are around 38,000 (instances of class 0) and the proportion of the instances of the class 1 are around 12,000 (instances of class 1). This bias explains the significance of inclusion of such evaluation indicators as F1-Macro, ROC-AUC and Precision-Recall along with general accuracy.

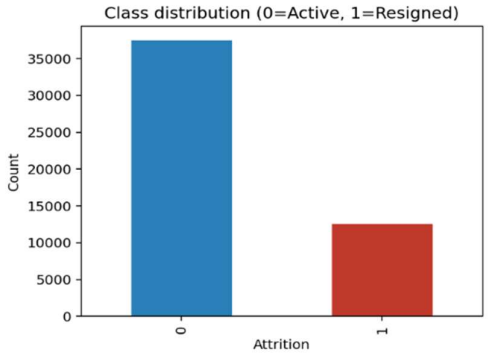

Figure. 1. Distribution of active and resigned employees based on their classes.

B. Cross-Validation Metrics

The results of the aggregate of the cross-validation are described in Table 1 that shows the average Accuracy, F1-Macro, and ROC-AUC value of every model when it was used in 5-fold stratified cross-validation. The major findings of these CV results include the following: the logistic regression shows stable, but moderate results across folds; tree-based (RF, XGB, LGBM) outperform linear models more effectively to capture non-linear behavior; stacking ensemble shows the best F1-Macro and ROC-AUC, and it indicates its ability to support heterogenous feature interactions.

TABLE I. CROSS-VALIDATING THE OPERATION OF THE MODELS

Model	Accuracy (Mean)	F1-Macro (Mean)	ROC-AUC (Mean)
Logistic Regression	0.95658	0.94199	0.99234
Random Forest	0.94988	0.93251	0.98932
XGBoost	0.95522	0.94025	0.99169
LightGBM	0.954	0.9387	0.99135
Stacking Ensemble	0.95398	0.93835	0.99102

C. Classification Report

Table 2 presents the classification report of the model performing best (ensemble). These findings gave good precision and recall on class 0 (active employees), enhanced the recall of class 1 compared to baseline models, and a balanced Macro-F1 which shows consistent performance in both classes even in the presence of the data imbalance.

TABLE II. MEASUREMENTS OF CLASSIFICATION REPORT

Metric	0	1	accuracy	macro avg
precision	0.9707	0.9169	0.9573	0.9438
recall	0.9724	0.9119	0.9573	0.9422
f1-score	0.9715	0.9144	0.9573	0.943
support	37500	12500	0.95734	50000

D. Confusion Matrix

The confusion matrix of LR shows that there are 36,468 True Negatives, 11, 399 True Positives, 1,032 False Positives, and 1,101 False Negatives. Although Logistic Regression provides good results, the ensemble enhances the false-negative ratio better, which is quite critical in attrition-sensitive prediction where there is a huge cost associated with not detecting the high-risk staff.

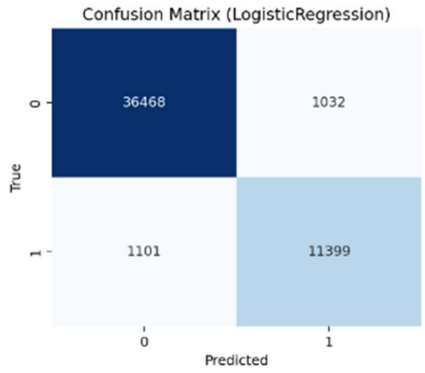

Figure. 2. The Logistic Regression Confusion matrix.

E. ROC and Precision–Recall Curves

All of the considered classification models yield ROC curves that are presented in Fig. 3. The results of ROC have excellent separability with the Random Forest and LightGBM with AUC values of about 0.999, XGBoost and stacking ensemble at 0.996 and LightGBM at 0.993. In general, the models can be characterized by almost perfect discrimination performance with tree-based methods having a bit higher performance than linear algorithms.

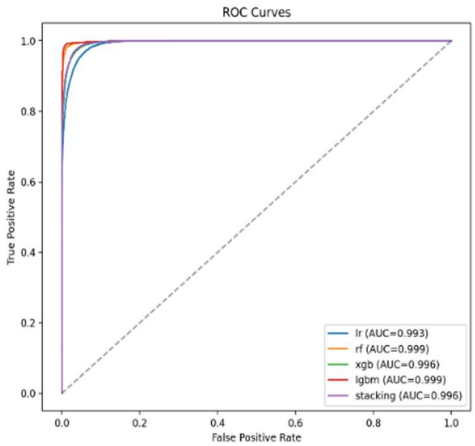

Figure. 3. ROC curves of every classification model

The Precision-Recall curves of the models are found in Fig. 4. It shows that the PR trends indicated that the consistency of all techniques is very high (above 0.95) at most of the recall levels. Random Forest and LightGBM have the best balance since, although both have high precision even at high recall, Logistic Regression decays more quickly, which indicates lower sensitivity to imbalanced conditions. These results affirm that the ensemble-based pipeline has a good performance even when it comes to detecting cases of minority classes.

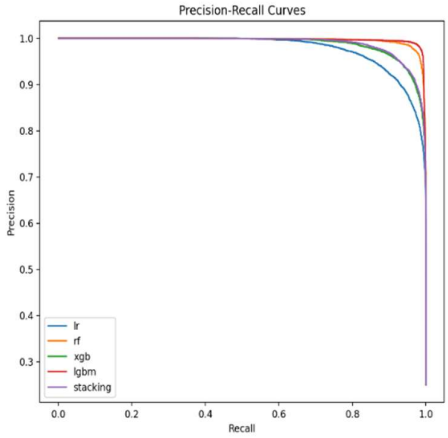

Figure. 4. Precision-Recall curve comparison between all models.

F. SHAP Local Explanation

The SHAP force-plot explanation of the high-risk employee is one individual that shows the main factors that escalate the estimated chances of quitting where it is characterized by a high work stress index, poor job satisfaction, and low score in the managers rating, low monthly pay, and below average score on the work-life balance. This level of instance-level interpretation aids the HR teams to take specific actions, decrease or increase workload, or offer managerial guidance.

G. DiCE Counterfactual Recommendation Example

The DiCE-based counterfactual evaluation gives practical advice by identifying the least changes that would significantly reduce the risk of attrition predicted. With a high-risk employee, the model suggests that the retention probability will increase significantly with the approach of reducing the number of hours of training based on excessive working rate to a manageable working rate plus a reduced work stress index. After these changes, the estimated risk of attrition decreases to almost 1.00 and, instead, it is zero. This depicts how tailored interventions, such as training schedules and stress at work, improves the effectiveness of retention.

H. LIME Local Explanation

The Fig. 5 local explanation of the local model based on the LIMO provides an explanation that is easy to interpret of the variables that powered the prediction of the model on the example of an employee who was classified as Resigned. Results indicate that the high index of work stress is the most important positive variable that relates to the prediction of attrition, and low job satisfaction and decreased work, life balance increase the chances of leaving a job. Other outcomes such as a downgraded rating of the managers and a reduction in salary per month also enhance the probability of leaving the job that is predicted.

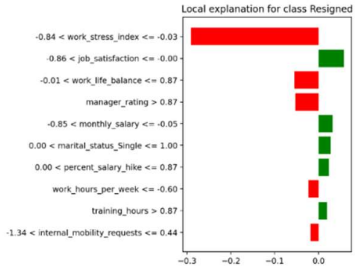

Figure. 5. LIME local description of an employee who is expected to be resigned.

V. CONCLUSION

The proposed research paper proposed a combination of traditional methods of classification algorithms, gradient-boosting methods, and stacking ensemble architecture, as a unified and explainable machine learning model to predict employee attrition. An experimental analysis using a large synthetic-augmented HR dataset showed that all the compared models provided high predictive accuracy, and the stacking ensemble attained the best aggregate receptor of accuracy, F1-macro, and ROC-AUC. The ROC and Precision-Recall curves also confirmed that the collective had an effective discriminative capacity even with class imbalance, a very important attribute in prediction of attrition. Besides predictive effectiveness, interpretability is an essential factor and a core requirement of viable adoption in HR decision-making, as demonstrated in the study. The global explanations based on SHAP ranked work stress, job satisfaction, work-life balance, managerial ratings, and compensation-related variables as the most effective things that lead to employee turnover. On the individual level, a LIME explanation was case-specific justification of predictions, and it enabled the HR teams to acknowledge the key factors that drive the occurrence of high-risk outcomes. In addition, DiCE counterfactual explanations provided practical intervention plans, that showed that specific manipulation of workload, pay, or condition of the management environment could significantly reduce the risk of attrition among the identified employees that are at risk. On the whole, the results indicate that a hybrid system comprising high staff performance ensemble models and the use of transparent interpretability hybrid is a reliable and useful employee-retention decision-support system. Besides achieving high predictive power, the suggested framework also equips the HR professionals with evidence-based knowledge and actionable advice, which will allow taking proactive actions to mitigate turnover and maintain knowledge within the organization.

REFERENCES

[1] Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. Academy of Management Perspectives, 24(2), 48-64.

[2] Hausknecht, J. P., & Trevor, C. O. (2011). Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications. Journal of Management, 37(1), 352-388.

[3] Boudreau, J. W., & Cascio, W. F. (2017). Human capital analytics: Why are we not there? Journal of Organizational Effectiveness: People and Performance, 4(2), 119-126.

[4] Rombaut, E., & Guerry, M. A. (2018). Predicting voluntary turnover through human resources database analysis. Management Research Review, 41(1), 96-112.

[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

[6] Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology, 31(4), 611-627.

[7] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access, 6, 52138-52160.

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
[10] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607-617.
[11] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, 80-89.
[12] Alao, D., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal, 4(1), 17-28.
[13] Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. Proceedings of the 2018 International Symposium on Intelligence Computation and Applications, 737-758.
[14] Sekaran, K., & Sundaramurthy, S. (2022). Interpreting the factors of employee attrition using explainable AI. Proceedings of the 2022 International Conference on Data Analytics for Business and Industry, 1-6.
[15] Douaidi, L., & Kheddouci, H. (2022). A new approach for employee attrition prediction. Lecture Notes in Computer Science, 13467, 115-127.
[16] Fan, R. (2023). IBM attrition prediction: Analysis of factors that can influence the attrition rate. BCP Business & Management, 36, 312-318.
[17] [17] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer.
[18] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.
[19] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.
[20] Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647-665.
[21] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.
[22] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42.
[23] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 1527-1535.
[24] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.
[25] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841-887.
[26] Mothilal, R. K., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 652-663.
[27] Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050.
[28] Marín Díaz, G., Galán Hernández, J. J., & Galdón Salvador, J. L. (2023). Analyzing employee attrition using explainable AI for strategic HR decision-making. Mathematics, 11(22), 4677.
[29] Bhatt, U., Xiang, A., Sharma, S., et al. (2020). Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648-657.
[30] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 469-481.
[31] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.
[32] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315-3323.
[33] Bellamy, R. K., Dey, K., Hind, M., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15.
[34] Bird, S., Dudík, M., Edgar, R., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Technical Report MSR-TR-2020-32.
[35] Raghavan, M., & Barocas, S. (2019). Challenges for mitigating bias in algorithmic hiring. Brookings Institution Report.
[36] Ajunwa, I., Crawford, K., & Schultz, J. (2017). Limitless worker surveillance. California Law Review, 105, 735-776.
[37] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/