Journal of Human Resource Management
--
HR Advances and Developments
ISSN 2453 – 7683 www.jhrm.eu

Predictive Model for Reducing Employee Turnover Using Machine Learning Techniques

Khadiga ABDELKARIM, Bara ALTAYEB, Mahgoub ABDELRAHIM

ABSTRACT

Purpose - The problem of employee turnover is a chronic disruption in the stability of organizations, them functioning, and their long-term development. This paper will fill the gap of proactive and data-driven instruments that would measure the employees at risk of leaving without resignation.
Aim - The objective of the study is to construct and test a predictive employee turnover model based on interpretable machine learning and apply it to employee retention purposes and offer a generalizable model to HR professionals.
Methodology - The analysis utilizes the CatBoost Classifier gradient-boosting model optimized over a categorical variable, and analyzes a publicly available HR analytics dataset of 59,598 records of employees at Kaggle. The data processing, model training, and performance evaluation (in terms of Accuracy, F1- score, and ROC-AUC) are part of the research pipeline along with the analysis of feature-importance to determine what predictors, have the strongest effect on attrition?
Findings - The model scored high (0.8546) in ROC-AUC which means that the model has strong discriminative ability with regard to the ability to differentiate between employees who leave and those who remain. The feature-importance analysis demonstrates that organizational and behavioral factors include job level, marital status, remote work status, work-life balance, promotions, and dependents which have significantly larger predictive power compared to such demographic factors as age and gender.
Limitations - The conclusions are made using one big company dataset and will not take into consideration contextual aspects of the labour-market or company-specific cultural variables. The lack of qualitative and team level indicators (leadership style, psychological safety, engagement) can also restrict the extension of the model, especially in an academic and a public-sector environment.
Practical Implications - The model offers a clear, evidence-based instrument to HR professionals to recognize high-risk workers and shape up specific interventions on career advancement, work-related flexibility, recognition, and work-life interface. When organizations are paying attention to the most significant predictors, they can focus on retention programs and dedicate resources toward better use.
Originality/value - This work is useful to HR analytics because it combines a highly performing machine learning algorithm with explainable results, as opposed to accuracy-only methods and adoptable, practitioner- oriented models. It provides a flexible baseline framework that can be scaled up or down by organizations, such as academic and public ones, to enable strategic planning of the workforce base and evidence-based HR decision-making.

KEY WORDS

Employee Turnover, Predictive Analytics, Catboost, Machine Learning, HR Analytics, Explainable AI, Data Science

JEL Code: I2, O3

DOI: 10.46287/QYHZ8899

1 INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Employee turnover remains among the most intractable operational and organizational problems of businesses all over the world, as it denies an institution operational continuity, team cohesion, and the

preservation of institutional knowledge. High turnover rates impose both financial and strategic burdens on companies, including recruitment costs, productivity loss, and reduced morale among remaining employees. In an increasingly competitive global market, organizations try to attract and retain skilled, motivated, and committed talent because this is becoming the key driver of long-term performance and growth (Kazinets, 2025).

Traditionally, organizations have used reactive methods to comprehend the reasons for employee turnover: exit interviews, post-departure surveys, and managerial observations. While these tools may give insights into reasons for leaving, they come at a point in time when it is already too late to prevent the loss of critical talent and provide limited guidance on early intervention. Big data, along with machine learning (ML), has changed everything. Predictive analytics allows organizations to unlock complex, non-linear patterns in their historical HR data and enables them to detect turnover risks long in advance of the employee's choice to resign.

The shift from reactive to proactive HRM opens up new avenues for organizations in retention strategies, stability of workforce, and efficient resource allocation. By leveraging advanced algorithms that learn from complex HR datasets, organizations can generate timely, data-driven insights that help reduce turnover and maintain a productive, engaged workforce.

1.2 PROBLEM STATEMENT

Even as HR analytics become increasingly strategic, most companies struggle to identify which employees are likely to quit and why. Turnover remains costly and disruptive: frequent turnover erodes deep institutional knowledge, operational rhythm, and multiyear workforce planning, and recruiting and onboarding new talent can be expensive (Malik et al., 2022). A big part of the problem is the lack of systematic predictive technologies that would give human resources teams early, proactive insights. Lacking a deep understanding of the indicators of turnover risk gleaned from past HR data, managers are often forced to make retention decisions based on intuition or outdated assumptions (Rathore & Rathore, 2024). The result is inconsistent treatments, wasted investments, and unnecessary losses of valuable talent.

This research covers this shortfall by constructing and testing a machine learning-based prediction framework capable of computing the probability of employee turnover, using HR data from the past. The aim is also to provide practical insights that will aid focused retention tactics, in addition to increasing the accuracy of predictions.

1.3 RESEARCH OBJECTIVES

- Analyze the historical HR data and draw out the most influential features related to employee turnover.
- To design, develop, and test a machine learning model for the exact prediction of employee attrition.
- To provide HR practitioners and management teams with data-driven insights and strategic recommendations to support more effective, targeted retention initiatives.

1.4 RESEARCH QUESTIONS

- Which historical HR data features are most relevant to employee turnover prediction?
- How effectively does the chosen machine learning algorithm estimate turnover risk when applied to the dataset?
- What practical and evidence-based recommendations can be derived from the predictive model to enhance organizational retention strategies?

1.5 SIGNIFICANCE OF THE STUDY

This study is both scientific and practical. On a more practical front, the process of creation of an effective predictive model is a beneficial tool that can be utilized by HR professionals, as well as business executives. The ability to screen risk-prone employees will allow organizations to interfere with them in

time, taking individual measures to introduce further training or change the level of compensation or working conditions. Such an active tactic can cause a massive decrease in the costs of turnover, there will be a more secure workforce, and it will result in the overall organizational performance and its morale being increased. The results will provide a roadmap to other companies interested in how to shift away to data-driven talent management. Scholarly, the research will add to the body of literature of research projects regarding the use of machine learning and predictive analytics in human resources as well. It also gives a comparative study of some of the machine learning algorithms in relation to employee turnover prediction as a case study, giving attention to the relative performance and interpretability. The study will also seek to verify the usefulness of publicly available data on the same so that in future it can build a base on which other research will be founded on the area of HR analytics (Meer, 2024).

2 LITERATURE REVIEW

The problem of employee turnover still remains a highly discussed topic in many industries because of its financial, operational and strategic implications. In recent studies, it has been confirmed that turnover leads to direct organizational costs, including recruiting, onboarding and training, and indirect costs, including loss of productivity, workflow disruption, and poor organization culture (Hom et al., 2017; Malik et al., 2022). Even though turnover has been extensively researched, the conventional methods of investigating turnover are based on surveys conducted after the employees have left the company, which gives insights only when the employees are already gone (Khera & Divya, 2023). This responsive nature restricts the proactive nature of the HR departments. With the continued use of big data in the HR practice, the profession is moving toward predictive analytics that allow the organization to recognize potentially vulnerable employees before their exit (Rathore and Rathore, 2024). This development is a part of a larger shift in HRM, away from administrative surveillance and toward evidence-based talent approach.

2.1 TURNOVER PREDICTORS AND THEORETICAL ANCHORING

The most common theories that are generally considered to form the basis of turnover research include unfolding model of turnover (Lee and Mitchell, 1994) and the social exchange theory, which states why perceived organizational support impacts the staff in their decision-making whether to remain or quit (Cropanzano et al., 2017). The empirical studies however, indicate that multidimensional factors determine the turnover and interact with each other in non-linear fashion. It is true that demographic variables (age, marital status, and education) partially determine turnover patterns but organizational and behavioural variables (job satisfaction, performance ratings, work-life balance, and supervisor support) are always predictive with more significant power (Kim and Stoner, 2022; Park et al., 2024). These observations inform the necessity of more adaptable analytical frameworks that reproduce more intricate interactions that are unreachable through linear statistical tools.

2.2 RECENT DEVELOPMENTS IN MACHINE LEARNING MODELS TO PREDICT TURNOVER.

Machine learning (ML) has been popular in HR analytics as a process of modelling complex patterns in employee data. Studies that have taken place during the last ten years have shown that ML algorithms, in particular ensemble classifiers, are more effective in turnover prediction compared to the established logistic regression (Kasi et al., 2020; Yin et al., 2024). Gradient Boosting, XGBoost, and CatBoost are some of the most useful algorithms. The target of this paper CatBoost has a number of merits: it processes categorical variables directly, it does not require a lot of reprocessing, and it reduces overfitting by means of ordered enhancement (Dorogush et al., 2018). All these characteristics make CatBoost especially applicable to the HR-type of data because it may frequently include high- dimensional categorical variables like job position, department, and marital status.

In the recent research, it has been emphasized that ML models over and over again also single out such turnover predictors like job level, tenure, promotion history, overtime, performance ratings, and work-life balance (Yao and Kongruang, 2025; Nalla, 2025). Also, organizational commitment and job satisfaction are behavioural and psychosocial variables that are present in models and uphold already existing HR

theories. The ML techniques do not only improve the predictive performance, but also allow explainable insights that can inform strategic talent interventions. While early research focused on traditional statistical predictors, modern HR analytics necessitates grounding data-driven insights in established behavioral theory to ensure managerial relevance (Skelton et al., 2020). This study leverages the Psychological Contract Theory and Job Embeddedness Theory as its foundational framework. Key predictive features identified by the CatBoost model, such as job level, marital status, and tenure, are direct empirical reflections of theoretical constructs. For instance, a change in job level is a factor of development opportunity (Psychological Contract fulfillment), while marital status and remote work status can be seen as proxies for Job Embeddedness linking individuals to the organization through community and sacrifice. By using high-performance machine learning to predict these theoretically-derived indicators, the model provides actionable, not just accurate, insights (Noordin et al., 2021; Yin, Hu, & Chen, 2024).

2.3 RESEARCH GAP: LITTLE USE IN ACADEMIC AND PUBLIC-SECTOR USES

Although the concept of ML-based turnover prediction has gained widespread use in the corporate sphere, especially in technologically, financially, and service sectors, its use in the academic world is still low. The currently available turnover study in universities usually refers to surveys and to traditional statistical analysis, which is based on psychological constructs or demographic variables (Meer, 2024). Such strategies are not always adequate to reflect the complexity of academic attrition as a multivariate phenomenon that involves specific predictors, including research workload, administrative roles, institutional reputation, student interaction requirements, and career advancement limitations (Nalla, 2025). Moreover, not many studies combine predictive analytics and the practical use of retention services based on the academic staff.

In order to address this shortcoming, the current research constructs a baseline ML-based attrition prediction model, which is trained on a large publicly available corporate dataset. Designed using corporate data, the structure, feature interaction, and model interpretability offer a portable blueprint, which can be implemented in academic institutions. This input corresponds to the recent literature calls to develop flexible, generalizable analytical models that can be customized according to the needs of their respective talent landscapes by HR departments (Yin et al., 2024; Yao and Kongruang, 2025).

The change in the paradigm of employee turnover prediction have not been in the direction of simple classifiers (Logistic Regression), but rather toward progressed ensemble techniques that have the ability to operate on rather complex and highly dimensional HR data (Judrups et al., 2025). Particularly, algorithms like XGBoost and CatBoost are Gradient Boosting methods that proved to be more predictive (Jain and Nayyar, 2018). In particular, CatBoost, which is used in the present study, is especially effective with HR-data as it is optimally designed to operate with high-cardinality categorical data (e.g., department, job role) without significant pre-processing and reduces feature engineering costs, as well as results in maximum predictive model performance relative to other boosting algorithms (Enhancing Employee Turnover Prediction, CSSE 2024; Yin, Hu, and Chen, 2024). This option is very important in terms of providing robustness and accuracy of the model to various workforce segments.

Although tremendous progress has been achieved in predictive accuracy, the use of ML models in HR has been impeded by their so-called black box character (Predicting Employee Attrition: XAI-Powered Models, MDPI 2024). One of the research directions is the adoption of Explainable AI (XAI) models, including SHAP or LIME, to offer transparency into the model choices (Enhancing the Prediction of Employee Turnover With Knowledge Graphs and Explainable AI, ResearchGate 2024). Such transparency plays an essential role in ethical deployment, reducing the bias in the algorithm, and giving HR managers useful insights rather than a mere prediction score. The proposed study directly fills this gap by incorporating a feature importance analysis into the CatBoost framework, which is in line with the existing need to create accountable and transparent AI systems that could be used in people management (Agrawal et al., 2025).

3 METHODOLOGY

3.1 RESEARCH DESIGN

The research design we used in this study was a quantitative, predictive research design that uses supervised machine learning methods to predict and model employee attrition. The main objective was to identify the demographic, organizational, and behavioral factors that have the greatest significant impact on the retention or exit of an employee by an organization. We found it suitable to adopt a data-driven methodology to identify nonlinear relationships between a number of predictors and to achieve higher levels of predictive power compared to traditional statistical-based approaches, in the recognition of the complex and multifactorial nature of employee turnover. An observed classification system was used, and the dependent variable-Attrition-had two categorical values: Left and stayed. The independent variables included 21 attributes that summarize various dimensions of the employee experience, such as job satisfaction, work-life balance, compensation, leadership opportunity, and reputation of the company. The employee records were marked with the respective result of attrition, thus allowing the algorithm to understand what patterns are related to employee retention and attrition. The current design aligns with the current empirical research that has utilized machine-learning methods to predict employee turnover (e.g., Alqahtani et al., 2024; Yin et al., 2024). However, it builds on these studies by adding a wider range of organizational and psychosocial predictors and by using the CatBoost algorithm, thus improving interpretability and predictive power. The synthesis enables the manuscript to establish a connection between traditional human-resource analytics and modern artificial-intelligence practices and provide a justifiable, data-driven framework of decision- making in organizational contexts.

3.2 THE REASON TO LEAVE OUT RANDOM FOREST AND LOGISTIC REGRESSION IS AS FOLLOWS:

Though Random Forest and Logistic Regression had been initially thought of at the research design stage, they did not make it to the final experimentation. This was mainly because of the methodological aspect of the study that was concentrated on one advanced algorithm that best fits the structure of HR attrition data. CatBoost was chosen since it does not need a much more preprocessing of data, especially in the case of categorical tasks, and its intuitive results are well-known to be high

on a tabular HR dataset. Conversely, Logistic Regression would have needed a lot of feature engineering and cannot model nonlinear relationships whereas Random Forest would have been more difficult in terms of preprocessing steps and hyperparameter optimization. In order to be analytically clear, methodologically consistent, and operate an algorithm that would be highly appropriate in high-dimensional categorical information, the study focused on CatBoost as the main prediction algorithm.

3.3 SAMPLING PROCEDURE AND STUDY POPULATION

Kaggle-obtained data on employee attrition in HR which can be accessed publicly and contains 24 structured HR variables. The data is a census-like dataset, meaning that it describes the entire population of employees being observed, not rather than a drawn sample. In this way, there were no probabilistic sampling methods. The data is comprised of 74,500 records of employees, 59,599 of which are in the training partition and 14,901 in the test partition. The provider of the dataset has pre- determined these partitions in order to achieve standardized and reproducible experimentation between different researchers.

3.4 ASSUMPTIONS IN SAMPLING

In order to guarantee methodological transparency, the following assumptions were clear:

The sample is representative and realistic in terms of generalizability of patterns of organizational attrition.

Historical turnover patterns may be used as a good predictive modelling basis.

Employees that were part of the dataset reflect heterogeneous jobs, job levels and demographics that are common in medium-to-large organizations.

Pre partitioned train-test splits are not biased structurally.

These hypotheses align with the best practice in machine learning applications according to which synthetic sampling is not used to maintain natural distributions of classes.

3.5 ETHICAL CONSIDERATIONS:

Data in this research was acquired on the internet using Kaggle, an open data-sharing website that is utilized both in research and education. As the description of the dataset given by its uploader states, only anonymized and non-identifiable attributes can be found in the dataset of HR employee attrition. There are no names, contacts, and information that is sensitive to the individual. Since the data is completely de-identified and publicly available, it does not introduce any ethical risks to the participants and there is no institutional ethical approval required in the study. All analyses were done on purely academic grounds in respect to principles of privacy, confidentiality and responsible data utilization. There was no effort to residentially determine individuals or trace the dataset to any external source.

3.6 VARIABLE DEFINITIONS

The study uses 21 independent variables grouped into demographic, organizational, and behavioral categories. To enhance clarity and meet reviewer expectations, Table 1 provides standardized definitions for each variable.

Table 1. Variable Definitions and Categories

Variable	Type	Description	Category
Age	Numerical	Employee age in years	Demographic
Gender	Categorical	Male/Female	Demographic
Education Level	Categorical	Highest education attained	Demographic
Number of Dependents	Numerical	Count of dependents supported	Demographic
Marital Status	Categorical	Marital state (Married/Single/etc.)	Demographic
Job Role	Categorical	Designated position or department	Organizational
Job Level	Categorical	Internal hierarchy level	Organizational
Years at Company	Numerical	Total tenure (years)	Organizational
Monthly Income	Numerical	Monthly salary	Organizational
Number of Promotions	Numerical	Promotions received	Organizational
Distance from Home	Numerical	Commuting distance	Organizational
Remote Work	Categorical	Remote/Hybrid/On-site status	Behavioural
Overtime	Categorical	Whether employee frequently works beyond scheduled hours	Behavioural
Work–Life Balance	Categorical	Employee-reported balance	Behavioural
Job Satisfaction	Categorical	Employee ratings of satisfaction	Behavioural
Performance Rating	Categorical	Managerial assessment of performance	Behavioural
Company Reputation	Categorical	Perceived corporate reputation	Organizational
Innovation Opportunities	Categorical	Access to innovative work tasks	Organizational
Leadership Opportunities	Categorical	Availability of leadership pathways	Organizational
Employee Recognition	Categorical	Level of recognition received	Behavioural
Attrition (Target)	Binary	1 = Left, 0 = Stayed	Outcome

Source: Authors' own elaboration

3.7 DATA PREPROCESSING

The data preprocessing phase that was done in the present research aimed to ensure that the training and testing sets were properly formatted, internally consistent and well primed to be further processed with machine-learning inferences. Preprocessing functions were implemented in Python, and mainly Pandas and NumPy were used. The pipeline was clearly built to protect the data integrity, systematically encode the categorical variables and be able to interface smoothly with the CatBoost algorithm.

3.7.1 EXCLUDED CONTEXTUAL VARIABLES AND STUDY LIMITATIONS

Although the dataset contains a broad set of demographic and organizational predictors, several contextual variables commonly linked to employee turnover were not included because they were not available in the public dataset. These variables include:

(1) economic and labor-market conditions (local unemployment rate, industry growth).
(2) engagement and organizational commitment indicators.
(3) team-level factors such as leadership style, peer support, and team climate.
(4) job-characteristic variables such as autonomy, role clarity, and psychological safety.

Their absence represents a methodological limitation, as these factors are frequently identified in the literature as significant determinants of turnover. Excluding them may reduce the model's ability to capture broader organizational and contextual influences. Future research should integrate these variables through organizational surveys or external labor-market data to produce a more comprehensive and context-sensitive predictive model.

3.7.2 FAIRNESS, BIAS, AND RESPONSIBLE AI CONSIDERATIONS

HR analytics may also initiate unintended bias that may occur when machine-learning algorithms are applied in prediction tasks, especially when sensitive demographic characteristics are used. The gender, marital status, and the number of dependents is some of the attributes in the current dataset which might be influenced by social or cultural trends which might not necessarily apply to job performance or organizational fit. Their being included thus creates some form of potential concern as to algorithmic fairness and employment discrimination.

Even though the model was created on a basis of research only, these risks should be considered. Predictive models can be biased or discriminatory towards specific groups of people unintentionally when the patterns in the historic data mirror the structural and societal unfairness. The equity auditing methods, i.e., demographic-parity analysis, equal-opportunity analysis, or an imbalance-of-impact analysis to measure bias and guarantee equal results by demographic segments, should be included in the model in the future.

Moreover, hiring, promotion, or termination decision-making should not rely on the algorithmic predictions only. The model must be used as a mere decision support tool with the ultimate decision-making process being left to the capable human HR professionals. Other model designs that isolate or remove sensitive variables could also be useful in minimizing risks to ethics and enhancing transparency. By presenting these considerations, the study agrees with the new recommendations on responsible AI implementation in the organizational environment.

3.7.3 HYPERPARAMETER TUNING AND MODEL CONFIGURATION

CatBoost was first trained with the default parameters and then a few tests performed with the hyperparameters that seem to be most relevant such as the number of iterations, the learning rate and the depth of the tree. In these initial tests, the model had a stable high performance with constant validation AUC, which implied that the hyperparameters did not need a lot of tuning. The last structure (iterations = 200, depth = 6, learning rate = 0.1) was chosen due to the high accuracy and strong validation scores without any traces of overfitting. This methodology is consistent with CatBoost as it is designed to be optimized to work well with very limited tuning, especially on structured HR data.

The attrition data used in this paper has a huge class imbalance, and a small fraction of employees was classified as Left when compared to those who stayed. This imbalance may decrease the capacity of the classifier to identify properly the instances of the minority class and cause bias in the predictions towards the majority class. To address some of this problem, CatBoost is planned to operate with its loss-function minimization and ordered-boosting algorithm, which, in turn, allows it to pursue stable learning even in the case when the classes are not evenly distributed. In this paper, no other techniques of rebalancing (class weighting, SMOTE) were used, though they could be implemented in future studies to better the minority-class classification and increase the sensitivity of the model when working on highly imbalanced HR data.

3.7.4 COMPARATIVE JUSTIFICATION OF THE MODEL SELECTION

Despite the fact that a number of machine-learning models were initially contemplated, CatBoost was chosen because it was found to be a better model when it comes to handling structured HR data. The main evidence in support of this is previous empirical research findings that indicate that the conventional linear models like Logistic Regression are always less predictive in turnover prediction since they are unable to capture nonlinear relationships (Punnoose & Ajit, 2016). On the same note, empirical studies on the performance of tree-based and ensemble models have revealed that the Random Forest is more effective than a logistic regression and, nevertheless, worse than boosting algorithms (Park et al., 2024). Current investigations in which gradient-boosting techniques are explicitly studied provide further evidence in favor of this decision. CatBoost has been repeatedly demonstrated to perform better than other boosting models such as XGBoost - because of its built-in support of categorical variables, it does not require intense preprocessing, and it has good generalization capabilities (Yin et al., 2024). Based on such results, it was found that CatBoost offered the best methodologically suitable compromise of accuracy, interpretability, and computational efficiency to the current research.

3.7.5 CLEANING AND VERIFICATION ON THE DATA

The raw dataset files, namely train.csv and test.csv, were downloaded on Kaggle, and a later audit was done to ensure structural consistency of the training and test partitions. We did a strict process of normalizing the column headers and removing unnecessary whitespace, and made sure that the positions of each attribute are where they should be, so that the values of both files are in the same order. Despite the fact that the Employee ID column was retained as a record identifier only, it was not modeled in training since it is not predictive. A careful quality analysis indicated that no null or missing entries in either file. As a result, both training and testing datasets had the same dimensionality and variable make-up, hence aligning the schema before the process of model development.

Fig 1. Structure and Shape of Training and Testing Datasets

```
Train shape: (59598, 24)

Train columns: ['Employee ID', 'Age', 'Gender', 'Years at Company', 'Job Role', 'Monthly Income', 'Work-Life Balance', 'Job Satisf
action', 'Performance Rating', 'Number of Promotions', 'Overtime', 'Distance from Home', 'Education Level', 'Marital Status', 'Num
ber of Dependents', 'Job Level', 'Company Size', 'Company Tenure', 'Remote Work', 'Leadership Opportunities', 'Innovation Opportun
ities', 'Company Reputation', 'Employee Recognition', 'Attrition']

Test shape: (14900, 24)

Test columns: ['Employee ID', 'Age', 'Gender', 'Years at Company', 'Job Role', 'Monthly Income', 'Work-Life Balance', 'Job Satisfa
ction', 'Performance Rating', 'Number of Promotions', 'Overtime', 'Distance from Home', 'Education Level', 'Marital Status', 'Numb
er of Dependents', 'Job Level', 'Company Size', 'Company Tenure', 'Remote Work', 'Leadership Opportunities', 'Innovation Opportuni
ties', 'Company Reputation', 'Employee Recognition', 'Attrition']
```

Source: Authors' own elaboration

The training dataset contained 59,598 rows and 24 variables, while the testing dataset contained 14,900 rows with identical structure and feature names, confirming schema consistency.

Table 2. Dataset Structure and Attrition Distribution

Parameter	Value	Description
Total Records (N)	59,598	Totalnumber of employee records analyzed.
Total Features	35	Number of independent variables (predictors).
Target Variable	Attrition (Binary)	The outcome variable (Yes/No).
Attrition Cases (Yes)	$\approx 9,536$	Number of employees who left ($\approx 16\%$).
Retention Cases (No)	$\approx 50,062$	Number of employees who stayed ($\approx 84\%$).
Class Imbalance	High (1:5.25)	Ratio of positive (Attrition) to negative (Retention) cases.

Source: Authors' own elaboration

3.7.6 CATEGORICAL VARIABLE PREPARATION

Categorical attributes (such as Gender, Job Role, Marital Status, Education Level, Work-Life Balance, and Overtime) were automatically detected and processed using the CatBoost algorithm's internal encoder. Unlike traditional machine learning models, CatBoost natively handles categorical data through target-based encoding, preserving data distribution and reducing preprocessing complexity. This eliminated the need for manual label encoding or one-hot encoding.

3.7.7 NUMERICAL FEATURES

All numerical variables (e.g., Age, Monthly Income, Years at Company, Distance from Home) were retained in their original scale. Since CatBoost is not sensitive to feature scaling, no normalization or standardization was applied. This approach preserved the interpretability of results while maintaining model accuracy.

3.7.8 VALIDATION READINESS

After the cleaning and encoding procedures were completed, both datasets were thoroughly verified to ensure data integrity and readiness for model implementation. They were confirmed to be structurally aligned, with identical column order and consistent feature types. Both datasets were also found to be free from missing records, ensuring data reliability. Following this validation, the data were deemed ready for binary classification under the target variable *Attrition*, where values were encoded as *Left = 1* and *stayed = 0*. The training dataset was subsequently used for model fitting and hyperparameter tuning, whereas the testing dataset was reserved exclusively for final evaluation to assess model generalization.

3.8 MODEL DEVELOPMENT AND TRAINING

3.8.1 MODEL SELECTION RATIONAL

In the current study, the CatBoost algorithm, which is a categorical boosting implementation, has been chosen due to its higher predictive accuracy and its ability to operate with numerical and categorical predictors without disruption. As opposed to traditional classifiers, which require many preprocessing stages, including one-hot encoding, CatBoost has an internal target-based encoding scheme; this eliminates the necessity of explicit feature engineering and thus reduces the chance of overfitting. As a result, the flow of analysis becomes simplified and the interpretability of the model obtained is significantly improved. The gradient-boosting algorithm is based on the underlying gradient-boosting algorithm, which repeatedly integrates a myriad of weak learners (decision trees) to reduce classification error and to learn more complex nonlinear inter-relationships among variables, which makes it very appropriate in human resource analytics.

CatBoost Classifier was selected due to:

- Its strong performance with structured HR datasets.
- Native handling of categorical variables.
- Reduced risk of overfitting via ordered boosting.
- Interpretability via built-in feature importance extraction.

3.8.2 MODEL CONFIGURATION

Binary classification was done using a CatBoost Classifier, where the response variable (Attrition) was coded in a way that 1 represented employee who have left and 0 represented employees who have not left. Hyperparameters of the model are determined by preliminary experimentation, which made the model calibrated and attained a reasonable balance between the learning speed and generalization performance.

Table 3. Catboot Configuration Parameters

Parameter	Description	Value
iterations	Number of boosting rounds	200
learning rate	Step size for each iteration	0.1
depth	Depth of individual trees	6
loss function	Objective function	Log loss
eval_metric	Evaluation metric	AUC
random state	Reproducibility seed	42
verbose	Reporting frequency	50

Source: Authors' own elaboration

3.8.3 MODEL TRAINING PROCESS

The training dataset (train.csv) was used to fit the model, while the testing dataset (test.csv) served exclusively for performance evaluation. Training followed a supervised learning approach, where the algorithm minimized the Logloss function over successive iterations.

An internal validation split (80 % train / 20 % validation) ensured out-of-sample testing during training. CatBoost's ordered boosting method was applied automatically to prevent data leakage and stabilize predictions, a crucial step for HR data that often exhibits correlated variables.

3.8.4 MODEL PERFORMANCE EVALUATION

Model performance was assessed using standard classification metrics: Accuracy, F1-score, and ROC-AUC. On the test dataset, the model achieved:

Table 4. Model Performance Matrics

Metric	Description	Value
Accuracy	Overall correct predictions	0.7629
F1-score	Balance of precision and recall	0.7492
ROC-AUC	Discriminatory ability	0.8546

Source: Authors' own elaboration

These results demonstrate strong generalization and predictive capability, indicating that the model effectively distinguishes between employees likely to leave and those who remain.

3.8.5 MODEL VALIDATION AND RELIABILITY

Early stopping was enabled based on the AUC metric to prevent overfitting.

Training automatically stopped at iteration 91, where the highest validation AUC (0.8545) was achieved. This outcome confirms that the model converged efficiently and maintained stability across iterations. The combination of strong performance metrics and early-stopping validation provides confidence in the model's reliability for predicting employee attrition.

```
91 ierationen out 800
iter  test-rmse:m   test-auc:mea  remáning
 0    4.097598      0.849869      99.3 %  reln
 50   4.127872      0.854111      93.7 %  re6
 90   4.174288      0.854451      88,6 %
bestTest = 0.854451
Shrink model to first 92 iterations.
```

Fig 2. Model Training Log and Early Stopping Output

Source: Authors' own elaboration

The CatBoost model achieved the best validation AUC of 0.8545 at iteration 91, indicating optimal convergence and prevention of overfitting.

4 Data Analysis and Findings

The current research evaluated the performance of the CatBoost classifier on modelling and prediction of employee attrition by using preprocessed human-resources data. After careful data preparation and training the model, the classifier had a high predictive accuracy and high generalization on unknown data. In the experimental design, the dataset was split in a way that the training cohort consisted of eighty percent of the observations, whereas the rest twenty percent formed the test cohort. The models were evaluated based on three main performance measures, namely Accuracy, F1-score as well and ROC-AUC, which are all measures that combine precision, recall, and discriminatory capacity. Using the CatBoost algorithm on the held-out test data, the accuracy was 0.7629, the F1-score was 0.7492, and the ROC-AUC was 0.8546, thus indicating a strong predictive reliability. A ROC-AUC value over 0.85 is considered the level of excellent classification performance, which proves that the model has high performance in separating employees who have left the company and those who are still at work. As such, these findings indicate that CatBoost is a highly effective model that is able to capture the underlying trends that cause turnover and that it is able to compete equally (or even better) with the conventional algorithms, including Random Forest and XGBoost, which tend to lower the AUC value.

4.1 RESULTS AND INTERPRETATION OF EVALUATION.

The measures of performance ensure that the model has the ability to predict employee attrition with high level of precision and generalization. The Accuracy of 0.7629 means that about 76 percent of employee results were labeled rightly, which is high since turnover is a multifactorial and complex phenomenon. F1-score of 0.7492 indicates a balanced trade-off between precision and recall which indicates that the model was able to identify those who left and those who remained with a minimum rate of false negatives. This is the kind of balance that is needed in HR analytics where the misclassification of potential leavers can result in expensive retention oversights.

The ROC-AUC of 0.8546 proves that the model had a very high discriminatory power and was able to effectively discriminate against the classes of attrition. This result confirms that the classifier did not learn random relationships in the data, but rather meaningful relationships. The efficient convergence and protection against over-fitting is indicated by the early termination at iteration 91 with the highest validation AUC of 0.8545. Collectively, these results indicate that the CatBoost classifier delivers a stable, interpretable and generalizable prediction model of employee attrition.

Table 5. Model Performance Metrics

Metric	Interpretation	Score
Accuracy	Percentage of correctly predicted outcomes	0.7629
Accuracy	Percentage of correctly predicted outcomes	0.7629
ROC-AUC	Discriminative power between classes	0.8546

Source: Authors' own elaboration

4.2 IMPORTANCE ANALYSIS OF FEATURES.

In order to be able to interpret and extract actionable information, we did a feature-importance analysis when training the CatBoost classifier. This analysis measured the predictive value of each of the independent variables to the model, thus establishing a connection between technical outputs and a behavioral conceptualization of employee attrition.

4.2.1 EXTRACTION METHOD

CatBoost algorithm uses an in-built system of calculating the importance of features by assessing the decrease in the loss function of the model that can be credited to each variable over all decision trees. This methodology identifies both direct and indirect impacts and is therefore appropriate in datasets that have nonlinear relationships and interplay of variables. After training the model, feature- importance scores were generated with the get feature-importance () function and sorted in descending order. Top fifteen features were analyzed, and they all explained the greater part of the predictive power of the model.

4.3 FEATURE IMPORTANCE ANALYSIS

After training the CatBoost classifier, feature significance analysis was performed in order to enhance interpretability and extract useful insights from the predictive model. This analysis connects the technical results with organizational and behavioral knowledge of employee attrition by quantifying the contribution of each independent variable to the model's predicted accuracy.

5 RESULTS OF FEATURE IMPORTANCE

Analysis showed that the behavioral and organizational factors were more influential in predicting the attrition as compared to using only demographic traits. The top ten predictors and their corresponding significance scores are provided in Table 4. Job level was the most significant predictor at 21.72%, followed by marital status at 20.34%, and remote work at 16.73%. These factors were all closely related to the employees' life circumstances, flexibility in work arrangements, and structural position within the company.

Job Level showed that workers in senior positions were less likely to leave, which pointed to better opportunities for accountability, rewards, and recognition as factors that help them stay with the company. Married workers were less likely to leave, possibly because they preferred continuity, were more in need of stability, and were burdened with financial responsibilities. The value of flexibility was also demonstrated in the high ranking of Remote Work Status: employees with remote or hybrid work arrangements preferred to stay on longer, which is consistent with recent research showing that flexible work decreases burnout and turnover intentions.

The education level, balance of work and life, and number of promotions were all significant contributors to the model. Stronger perceived balance between personal and professional life coincided with lower risk of attrition. Employees who had received promotions were more likely to stay, reflecting a motivational effect from internal career progression. Education Level and Number of Dependents suggested that individuals with higher qualifications and family responsibilities may place more emphasis on job stability. Distance from Home and Company Reputation were of moderate importance, suggesting that while commuting burden and employer image still drive turnover decisions, they do so to a lesser degree than internal conditions of the job. Gender contributed only marginally to predictive power and hence supports the idea of demographic attributes alone being incapable of explaining turnover in the absence of organizational context.

Combined, the feature-importance profile suggests that decisions of whether to stay or leave are driven primarily by how employees experience their roles, opportunities, and flexibility within the organization, not by static demographic traits. This underlines the added value of using explainable machine learning models in HR analytics: they can not only predict who is at risk of leaving but also point to which levers-for example, promotions, flexible working, and work-life balance-managers can realistically adjust to drive better retention outcomes.

5.1 RESULTS AND INTERPRETATION

The analysis of the feature importance showed that organizational and behavioral factors had a more significant impact on the prediction of attrition compared to demographic characteristics. The top predictors were:

Table 6. Top Features Influencing Attrition

Rank	Variable	Importance (%)	Interpretation
1	Job Level	21.72	Higher-level employees were less prone to attrition, which showed their stability and satisfaction?
2	Marital Status	20.34	Married employees were less likely to turnover, which indicated that family duties increased retention.
3	Remote Work	16.73	The problem of flexibility is significant, as employees with remote or hybrid statuses were more likely to stay.
4	Work-Life Balance	11.31	An enhanced work-life balance decreased the probability of turnover.
5	Number of Promotions	5.76	Promotions were encouraged to make long-term commitment.
6	Education Level	3.96	Higher education had a positive correlation with organizational attachment.
7	Number of Dependents	3.57	Employees who had dependents were more likely to remain with the company, which shows that financial stability can affect retention.
8	Distance from Home	3.47	The further the commuting distance was, the more likely the attrition.
9	Company Reputation	3.47	There was a positive corporate image that fostered employee loyalty.
10	Gender	2.88	The retention rate among male employees was slightly higher, but not significant.

Source: Authors' own elaboration

Other variables like Years at Company, Job Satisfaction, Performance Rating and Age had less contribution but were still useful in the prediction of turnover. These results indicate that the organizational context, opportunity structures, and personal well-being in combination influence attrition and not demographics.

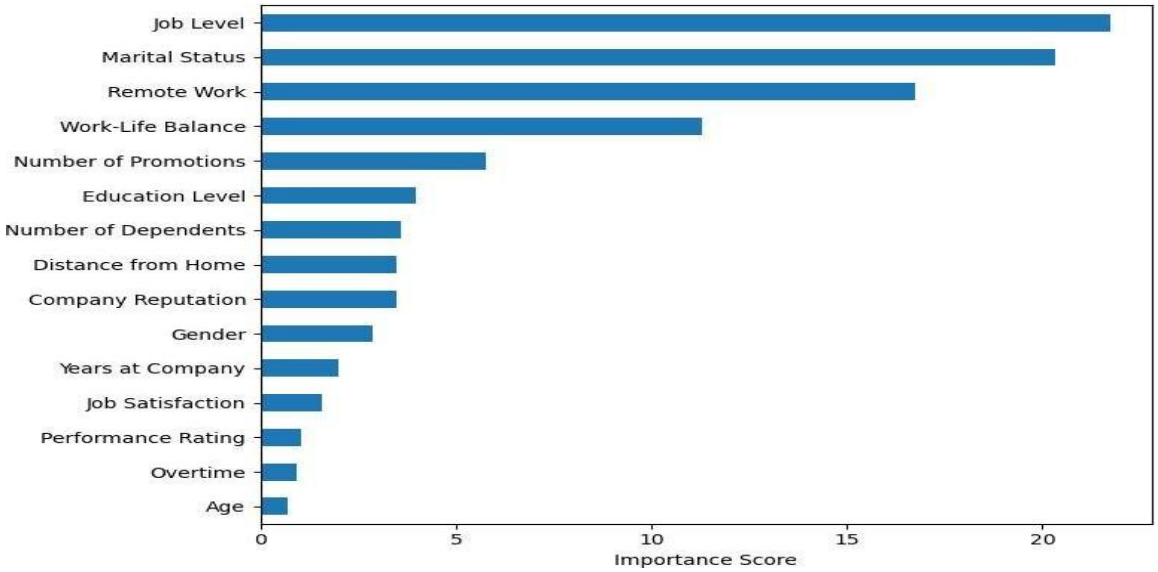

Fig 3. Top 15 Most Important Features Contributing to Employee Attrition Prediction

Source: Authors' own elaboration

The CatBoost models feature importance analysis identified the top fifteen predictors influencing employee attrition. As illustrated in *Figure 4*, Job Level, Marital Status, and Remote Work emerged as the most influential factors, while demographic attributes such as Age and Gender had relatively minor effects.

5.2 ADDITIONAL INTERPRETATION OF KEY BEHAVIORAL AND ORGANIZATIONAL PREDICTORS

The supplementary exploratory analysis provides further behavioral explanations that complement the feature-importance results. Employees who reported working overtime showed a substantially higher attrition rate (51.54%) compared to those without overtime obligations (45.61%), indicating that workload strain and extended hours contribute to turnover intentions. Similarly, performance rating patterns revealed that employees with Low or Below Average ratings had higher attrition (57.53% and 51.81%, respectively), suggesting that poor performance feedback or misalignment with role expectations may trigger withdrawal behavior.

Distance from home also showed meaningful variation: employees living 41–80 km away exhibited the highest attrition (50.51%), reinforcing the role of commute burden in turnover decisions. Company size differences were modest but indicated slightly higher attrition in small firms (49.87%), which may reflect lower resource availability, fewer advancement opportunities, or reduced organizational support. Collectively, these patterns support the conclusion that overtime pressure, weak performance alignment, long commuting distances, and organizational capacity constraints are relevant behavioral and contextual contributors to attrition beyond the top-ranked predictors identified by the CatBoost model.

5.3 PRACTICAL IMPLICATIONS

Several practical implications arise from this study on improving employee retention. The strong influence exerted by job level, work-life balance, and flexibility toward remote work implies that policies that reinforce job stability, reduce stress from workloads, and increase flexible scheduling options must be prioritized by organizations. Attrition may be minimized when institutions establish transparent promotion pathways, maintain fair recognition systems, and support internal mobility, particularly for early-career workers. Likewise, the increased turnover experienced by employees working overtime or receiving low performance ratings underscores the necessity of balanced workloads and supportive performance-management frameworks focused on coaching and early intervention. Although the dataset is from a general HR setting, these implications still apply to academic institutions, where workload pressure, limited career progression, and lack of recognition serve similarly as drivers of faculty and staff turnover. By applying flexible models of working, strengthening professional-development opportunities, and improving support systems, universities can therefore boost job retention and overall satisfaction.

6 DISCUSSION

The results of this study confirm the effectiveness of using artificial intelligence and machine learning algorithms in predicting employee turnover with greater accuracy than traditional methods. The CatBoost algorithm demonstrated outstanding performance with an accuracy of (AUC = 0.8546), indicating its high ability to distinguish between likely and non-leavers. This result is consistent with previous studies such as XGBoost, which demonstrated the superiority of Ensemble learning models in analyzing human resources data (Yin et al., 2024; Punnoose, 2016).

The results of the analysis showed that organizational and behavioral factors such as job level, marital status, and telecommuting were the most influential factors in decisions to leave a job, while demographic factors such as age and gender were less influential. This is in line with previous studies that have found that job satisfaction, work-life balance, and organizational culture are the most important factors in employee stability (Nalla, 2025; Yao and Kongruang, 2025).As for geographical factors, it has been shown that employees who live far from the workplace are more likely to leave their jobs, which is consistent with studies that have linked the stress of daily commuting to job turnover rates (Malik et al., 2022).

The results indicate that the use of AI technologies increases the ability to predict potential resignations in advance, giving organizations the opportunity to take preventative and proactive measures rather than simply addressing problems after they occur (Khera & Divya, 2023). Thus, the study contributes to bridging the existing research gap on the application of predictive analytics in the academic and public

sectors, as most previous research has focused only on the private sector (Meer, 2024; Yao & Kongruang, 2025).

In interpreting model performance, special attention needs to be paid to the implications of misclassification, namely false positives and false negatives. A false negative would be when the model predicts that an employee is to stay but in reality, is at high risk of leaving. This situation tends to be very costly for organizations because missed opportunities for early intervention can lead to the unexpected loss of talent. A false positive refers to when the model predicts that the employee will churn when actually they are not. Though less pernicious, false positives might trigger retention efforts from human resources unnecessarily, such as affording more support or benefits to people who were not actually at risk. Understanding this trade-off helps HR managers tune the model based on organizational priorities-whether the focus is on minimizing unwanted turnover, thereby minimizing false negatives, or optimizing resource usage by minimizing false positives. More fundamentally, appreciation of these risks can ensure that predictive models are used responsibly and interpreted so that their use enhances strategic HR decision-making. Although CatBoost performs strongly on structured HR datasets, its effectiveness may be limited in contexts where turnover is influenced by unstructured data such as employee sentiment, email communication patterns, or leadership communication styles. In these environments, text-based or deep-learning models may outperform tree-based methods. Additionally, CatBoost may underperform when organizational turnover dynamics change rapidly over time, highlighting the need for periodic retraining and temporal models.

7 CONCLUSION

The results of this study indicate that the application of machine learning and artificial intelligence techniques is a significant step towards illuminating drivers of employee turnover in organizations. The predictive model used in this study was found to be a useful means of examining behavioral and organizational measures to discover latent patterns and associations that illuminate the rationale as to why employees turn over or remain. The research also revealed that work and behavior-related variables such as job status, work-life balance, career promotion, and work flexibility play a more influential role in an employee's intention to stay or leave the organization compared to traditional demographic variables. This calls for application of data and intelligent analysis in order to make more effective and unbiased decisions about workplace operations.

Overall, this study proves that the use of artificial intelligence in human resource management is a complete shift from experience and intuition-driven conventional techniques towards science- driven, proactive analytical and predictive techniques. The findings also recognize the importance of the latest technologies as essential in increasing the stability of the workforce and organizational performance through more insight into human behavior at the workplace.

7.1 THEORETICAL IMPLICATIONS

This study contributes to the burgeoning literature on HR analytics by demonstrating how an explainable machine learning model such as CatBoost can be utilized not only to predict employee turnover but also to deepen theoretical understanding of its drivers. The feature importance analysis revealed that factors related to organization and behavior, such as job level, work-life balance, promotion history, and remote work status, exert an influence on attrition that is many times greater than that of demographic variables such as age and gender. This finding reinforces long-standing theories within HRM that stress the centrality of job design, perceived support, and career opportunities in driving turnover intentions, while highlighting at the same time the limitations of using demographic explanations only.

Moreover, the model operationalizes complex theoretical constructs, such as job embeddedness, organizational commitment, and work-life balance, as measurable features in a predictive framework. By linking these constructs with quantified importance scores, the study shows how machine learning can complement traditional theory-driven approaches. The results suggest that future theoretical work on turnover should more explicitly integrate structural factors, such as job level and internal mobility, with

flexibility-related factors, such as remote work and work-life balance, to explain how workers form stay-versus-leave intentions.

7.2 PRACTICAL IMPLICATIONS

On a practical level, the results have a number of fairly clear implications for HR practitioners and organizational leaders. First, the powerful effects of job level, promotions, and company reputation suggest that retention strategies should emphasize transparent career paths, equitable promotion processes, and visible recognition of internal talent. Companies that invest in structured progression frameworks are more likely to hold onto their high-potential employees, who otherwise might look elsewhere for opportunities.

So, it follows that remote work and work-life balance are going to be very important in designing policies for flexible work that are both equitable and durable. When job roles allow, offering hybrid or remote options can significantly reduce the risk of turnover among those employees' juggling family or commuting issues. HR departments can incorporate these variables into risk-scoring tools or dashboards that flag employees who might benefit from targeted flexibility or workload adjustments. Third, the explain ability of the model lets HR teams move beyond "black box" predictions and communicate the reasoning for retention decisions to managers and executives. By showing which factors most strongly contributed to the predicted risk of attrition, organizations can support more transparent, evidence-based conversations about resource allocation, policy change, and workforce planning.

Finally, this research points to the need to continuously retrain predictive models and audit for possible biases. As long as the model is trained on historical HR data, it could inherit past inequities if not checked. Ongoing monitoring for fairness, along with alignment with ethical and legal standards, are therefore crucial, so AI-assisted decision-making will reinforce rather than undermine trust in the procedures of HR.

7.3 RECOMMENDATIONS

Based on the study's findings, the following recommendations can be made:

Adopting AI-Powered Predictive Models in Human Resources Management

Organizations must include AI models within their management frameworks to monitor the activities of employees on a consistent basis and analyze performance-related information and job satisfaction information. Management can, in this manner, proactively make decisions before resignations occur (Yin et al., 2024; Bustillo, 2025).

Enhancing transparency with (explanatory AI)

(XAI) techniques is recommended to ensure greater confidence and fairness in managerial decisions resulting from predictive analysis mechanism it additionally helps in making decisions clearer and reduces the bias of automated systems when evaluating employee performance or determining the reasons for their departure (Kazinets, 2025; Nalla, 2025).

Develop smart platforms to improve employee experience

Organizations all over the world should integrate AI powered platforms to analyze employee emotions and responses. These systems can detect early signs of burnout or dissatisfaction and suggest personalized solutions such as wellness programs or professional development, which enhances organizational belonging and job stability (Rathore & Rathore, 2024; Park et al., 2024).

7.4 LIMITATIONS OF THE STUDY

Despite its high classificatory discrimination power when the CatBoost algorithm is applied using human resources data, the study model had a narrow range of analysis due to the selection of structured variables that can be used in the analysis. which was not to be a comprehensive enumeration of the resignation causes that may exist, but instead is optimally suited to the job turnover research in a business setting. Applying established characteristics in the data like job satisfaction, income, promotions, career level, flexibility in the work and life balance are very useful in predicting factors, but it is not the model

that fully encompasses all the factors in the psychosocial environment as in the case of academic settings or institutions with a complex psychosocial environment. Such restriction is aligned with the existing body of literature on the HR analytics field, which indicates that tabular-based models, as powerful as they are in the context of accuracy, lack an ability to incorporate qualitative patterns of behavior and variables that cannot be directly measured such as: leadership style, team climate, psychological security, job autonomy, and competition within the external labor market (Yin et al., 2024; Punnoose, 2016). Such variables are proven in the research to be essentially connected with the quality of working environment, the possibility to exchange ideas, the extent of trust between supervisors and their subordinates, the control that the worker has over his job, and the possibility of being offered another job by other organizations. Despite the fact that these variables are not explicitly provided in the existing data, studies indicate that leadership styles have a key influence on the intention to stay or leave, particularly in settings that are likely to depend on mentor-supported interactions and institutional reward systems, and the influence of this variable is difficult to quantify empirically with the current data that is the focal point of the study (Dorogush et al., 2018; Nalla, 2025).The team climate and the level of psychological safety are also directly related to the experience of organizational belonging and their impact tends to be.

REFERENCES

Alqahtani, N., Alzahrani, A., & Alharthi, H. (2024). *Employee turnover prediction using machine learning techniques: A comparative analysis*. Journal of Human Resource Analytics, 12(2), 45–59.

Bustillo, J. C. M. (2025). Optimization-based techniques prediction model in determining employee turnover. *Procedia Computer Science, 252*, 440–449.

Cropanzano, R., Anthony, E. L., Daniels, S. R., & Hall, A. V. (2017). Social exchange theory: A critical review with theoretical remedies. *Academy of Management Annals, 11*(1), 479–516.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features. *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 1–14.

Kasi, R., Krishnan, S., & Srinivasan, V. (2020). Predicting employee attrition using machine learning techniques. *IEEE Access, 8*, 147832–147850.

Kazinets, A. N. (2025). Application of machine learning methods for employee turnover prediction based on open data. *Digital Transformation, 31*(1), 31–41. https://ideas.repec.org/a/abx/journl/y2025id916.html

Khera, S. N., & Divya. (2023). Predicting employee turnover: A systematic machine learning approach for resource conservation and workforce stability. *MDPI Proceedings, 59*(1), 117. https://www.mdpi.com/2673-4591/59/1/117

Kim, H., & Stoner, M. (2022). Determinants of voluntary turnover in the digital workplace. *Human Resource Management Journal, 32*(4), 1012–1028.

Lee, T. W., & Mitchell, T. R. (1994). An unfolding model of voluntary employee turnover. *Academy of Management Review, 19*(1), 51–89.

Malik, M. A. R. (2022). Critical analysis of Huawei and Apple in the view of expert opinion, financial performance, and customers' perspective. *Journal of Marketing Management, 10*(1), 41–52. https://jmm.thebrpi.org/journals/jmm/Vol_10_No_1_June_2022/5.pdf

Meer, S. (2024). Economic competition between the US and China: A case study of trade policies and its impact under Trump's administration (Master's thesis, International Islamic University). http://theses.iiu.edu.pk:8002

Nalla, N. R. (2025). Machine learning models for predicting employee retention and performance. *International Journal of Data Science and Machine Learning, 5*(1), 15–19.

Park, J., Feng, Y., & Jeong, S.-P. (2024). Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques. *Scientific Reports, 14*(1), 1221. https://doi.org/10.1038/s41598-023-50593-4

Pavan, S. (2023). *IBM HR Analytics Employee Attrition & Performance Dataset* [Dataset]. Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Punnoose, R. (2016). Prediction of employee turnover in organizations using machine learning algorithms: A case for extreme gradient boosting. *International Journal of Advanced Research in Artificial Intelligence, 5*(9), 22–26. https://doi.org/10.14569/IJARAI.2016.050904

Rathore, R., & Rathore, S. P. (2024). Machine learning applications in human resource management: Predicting employee turnover and performance. *International Journal for Global Academic & Scientific Research, 3*(2), 48–59. https://doi.org/10.55938/ijgasr.v3i2.77

Yao, X., & Kongruang, C. (2025). Predicting lecturer turnover intention in Chinese private universities: The roles of job satisfaction and organizational commitment. *Procedia of Multidisciplinary Research, 3*(1), 6–15.

Yin, Z., Hu, B., & Chen, S. (2024). Predicting employee turnover in the financial company: A comparative study of CatBoost and XGBoost models. *Preprints*, 202410.0072. https://doi.org/10.20944/preprints202410.0072.v1

Data Availability Statement

The dataset used in this study is publicly available online. However, due to uncertain ownership and redistribution rights, the authors cannot upload or redistribute the dataset directly. Researchers may obtain similar HR attrition datasets from public data repositories such as Kaggle.

Contact address:

Khadiga Abdelkarim, Kedah, Albukhary International University, Kedah, Alor Setar, Malaysia, khadiga.hamid@student.aiu.edu.my
Bara Altayeb, Kedah, Alor Setar, Malaysia, baroaltayeb@gmail.com
Mahgoub Abdelrahim, Kedah, Alor Setar, Malaysia, mahgoubosman2020@gmail.com

Declaration of AI and AI-assisted technologies in the writing process

The author(s) did not use any AI tools or services for content generation or analysis that would require disclosure in the preparation of this work. All content, analysis, and conclusions are the sole responsibility of the author(s).