EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING APPROACHES

Debasmita Paul
Dept. of Electrical Engineering
Techno International New Town Kolkata, India
Email Id: debasmitapaul1712@gmail.com
Prerona Pattrea
Dept. of Electrical Engineering
Techno International New Town Kolkata, India
Email Id: preronapattrea623@gmail.com
Sobhan Sarkar
Information Systems & Business Analytics
Indian Institute of Management Ranchi Ranchi, India
Email Id: sobhan.sarkar@iimranchi.ac.in
Milan Basu
Dept. of Electrical Engineering
Techno International New Town Kolkata, India
Email Id: dr.milan.basu@tict.edu.in
DOI: https://doi.org/10.64558/bllp/967655hukflw

ABSTRACT

Recent industry reports, including EY's Future of Pay 2024, reveal that the average attrition rate for employees across various industries in India reached around 18.3 % in 2023. This ongoing challenge adversely affects profit manage- ment, workforce stability, and operational efficiency. In the past, understanding why employees leave largely depended on lengthy HR processes like surveys and interviews. Machine learning en- ables organizations to move beyond these traditional approaches, offering quicker, data-informed, and more reliable forecasts of attrition. This study addresses these gaps by conducting a comprehensive, comparative analysis of nine machine learning classifiers which are Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Na¨ıve Bayes, K-Nearest Neighbors (KNN), XGBoost, and CatBoost in a dataset of 1,470 employees with 35 diverse attributes. The aim is to identify all the factors that contribute to attrition and to evaluate it using performance metrics. The results of the research imply that CatBoost achieved the highest accuracy (92.71%), while XGBoost showed the best class separation (ROC-AUC 0.968). This approach aims to enhance decision-making in HR analytics by assisting HR teams to predict employee turnover effectively, minimize attrition costs, and help to design work-life balance strategies, competitive pay structures, and customized career development programs.

Keywords : Employee Attrition, HR Analytics, Human Re- sources, Profit Management, Workforce Stability

INTRODUCTION

Employee attrition, often described as workforce turnover, represents a significant barrier for organizations by disrupting operational flow, diminishing productivity, and increasing fi- nancial strain. Being able to anticipate which employees might exit the organization allows businesses to design timely inter- ventions, control costs, and

maintain team unity [1]. Industry reports highlight the magnitude of this concern. According to a 2023 Work Institute analysis, voluntary turnover accounted for more than $600 billion in annual costs for companies in the United States, with over 40 million employees leaving within a single year. Main reasons included insufficient career growth prospects (22%), challenges in work-life balance (12%), and Non-competitive remuneration (9%). Beyond direct hiring and training costs, attrition also impacts team morale and obstructs ongoing projects. To counteract these challenges, many firms have turned to engagement tools like Qualtrics and Culture Amp, which offer mechanisms for gathering real- time employee feedback. However, these platforms mostly offer descriptive summaries and lack the predictive capability needed to plan ahead and reduce employee turnover. Motivated by these gaps, this study examines how machine learning can be utilized to forecast employee turnover. The dataset used encompasses of 1,470 records with 35 diverse attributes covering demographic, role-specific, and organizational di- mensions. Out of which 34 attributes are independent variables and there is only one dependent variable ”attrition” which contains categorical values like yes or no. A total of nine supervised learning algorithms including Logistic Regression, Na¨ıve Bayes, K-Nearest Neighbours (KNN), Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost are evaluated, complemented by techniques like SMOTE for handling class imbalance and feature scaling. This work emphasizes the growing need for data-centric decision- making in human resources, moving beyond intuition to evidence-backed strategies. Organizations are under growing pressure to spot possible attrition early and take specific measures to keep key employees as they ask not only for money but also for growth, flexibility, and job satisfaction [2]. This study helps them reduce turnover risks and build a stable workforce for the long term. This research offers several significant contributions, which include:

- Highlighting deficiencies in current HR analytic tools regarding predictive insights.
- Establishing a comparative framework using multiple algorithms and extensive features for attrition forecasting.
- Presenting an interpretable, data-driven approach to sup- port strategic retention efforts.

LITERATURE REVIEW

Existing literature emphasizes various approaches, datasets, and methodologies aimed at understanding the causes and factors that mitigate workforce turnover. Researchers have explored various machine learning techniques to better predict turnover. [3] applied data mining to detect early behavioral indicators of attrition. [4] evaluated multiple models, finding Decision Trees and Random Forests balanced performance with interpretability. Several studies focused on comparative model evaluations. [5] showed SVM outperforming Logistic Regression in both accuracy (86% vs. 78%) and precision. [6] developed an AI framework combining Gradient Boosting, De- cision Trees, and Logistic Regression, with Gradient Boosting yielding an F1 score above 0.9. [7] created a pipeline using KNN, Random Forest, and XGBoost, with XGBoost achieving nearly 90% accuracy and an AUC of 0.93.

Several studies explored domain-specific applications and actionable insights. [8] surveyed churn methodologies across employee and customer contexts, highlighting ML flexibility in HR. [9] documented the shift from traditional statistics to scalable ML in workforce analytics. [10] applied predictive models in education, stressing domain-specific variables for retention. [11] linked job satisfaction and pay to turnover, while [12] combined data-driven analysis with HR policy recommendations. In general, these works highlight the im- portance of combining technical models with practical insights so that companies can make smarter decisions and reduce the employee turnover.

Numerous previous studies focused on limited

demographic and compensation variables, employed fewer models, and omitted organizational elements and interpretability within the prediction of employee turnover. The gaps in previous research are mitigated through our work, which blends 35 distinctive features, utilizes nine classifiers, and applies explainable arti- ficial intelligence tools for visualization. This accuracy yields enhanced explanatory power, fostering improved decision-making. Our research aids in effective employee retention by reinforcing proactive intelligent workforce strategy manage- ment and shifting the HR focus from reactive to responsive practices

METHODOLOGY

In order to precisely forecast employee attrition, a robust and systematic approach was undertaken. This approach tra- verses various steps, ranging from data collection to prepro- cessing, analysis, training of the model, and evaluation. Each step is crucial to guaranteeing the validity and reliability of the predictive model. The steps followed are outlined below.

Data Collection and Initial Assessment

The method starts with the gathering of a holistic dataset containing both current and past employee data. This is done to ensure that the inherent patterns and trends pertaining to attrition are properly captured. The dataset forms the basis of all the subsequent steps.

An initial check is then performed on the dataset to assess its structure as well as quality. The following steps are carried out:

- Dataset dimensions: The number of rows and columns are analyzed to determine the extent of the data.
- Statistical summary: Simple statistical values like mean, median, standard deviation, and variance are calculated for every numerical feature to get an overall picture.
- Column classification: All features are categorized into numerical and categorical categories for proper treatment in the next steps.
- Duplicate records: The data set is scanned for any dupli- cate records, which, if present, are eliminated to ensure data integrity.
- Missing data evaluation: Every attribute is scanned for missing values. The number of and percentage of missing values for each column are computed.
- Distinct category values: Categorical variables are checked for distinct values. For instance, the 'Attrition' variable contains two values—'Yes' and 'No'. Remain- ing categorical fields are 'Department' (Sales, Research and Development, Human Resources), 'Education Field' (Medical, Life Sciences, Technical Degree), and 'Gender' (Male, Female).

The dataset had no null or undefined values, and there were no duplicate records found.

Exploratory Data Analysis

Exploratory Data Analysis is essential in understanding the structure, relationships, and hidden patterns of the dataset. The EDA process includes:

- Attrition distribution: The general distribution of employees who have remained and those who have left is represented using pie charts and histograms.
- Percentage analysis: Pie charts are used to illustrate the percentage of workers in each attrition group so the imbalance if any can be visualized.
- Feature-based insights: Numerical and categorical variables are examined in-depth. Inter-variable relationships are explored to determine the features with maximum chances of impacting attrition.

These insights inform the feature selection and identify possible issues like imbalance or correlation that may affect model performance.

Data Preprocessing

Several preprocessing operations are carried out to get the dataset ready for training models:

- Label encoding is used on binary categorical variables such as 'Gender'.
- One-hot encoding is used on multi-category variables such as 'Business Travel', 'Department', 'Education Field', 'Job Role', and 'Overtime'. This is done to make sure all categories are treated equitably without adding artificial ordinal relations.
- Highly correlated features like 'JobLevel', 'MonthlyIn- come', 'YearsAtCompany', and 'YearsInCurrentRole' are recognized and dropped. These attributes contribute to redundancy and multicollinearity, which hinder specific models by overstating the importance of some features and making them less interpretable.

Variable Selection and Handling Class Imbalance

Once preprocessing is done, the dataset is split into:

- Independent variables (X): All the features that are per- tinent and could have an impact on employee attrition.
- Dependent variable (y): The 'Attrition' column, where 1 represents an employee who left and 0 represents an employee who remained.

There is a class imbalance, with many more employees labeled 'no attrition' than 'attrition'. To counteract this, the Synthetic Minority Oversampling Technique (SMOTE) is uti- lized. SMOTE generates synthetic samples of the minority class (attrition) in order to have balanced class distribution. This step is important since imbalanced datasets may bias the model towards predicting the majority class, hence minimizing performance in correctly predicting attrition cases.

Feature Scaling

After balancing the dataset, feature scaling is done to nor- malize the range of independent variables. This is especially crucial for algorithms that use distance calculations or are sensitive to feature magnitude.

Min-max scaling is used to scale features into a consistent scale, typically between 0 and 1. Alternatively, standardization (z-score normalization) can also be employed based on the re- quirements of the algorithm. Standardizing the data accelerates training convergence and improves model performance.

Model Training and Evaluation

Having preprocessed, balanced, and scaled the data, it is now ready for model training and evaluation.

- Data splitting: The dataset is split into a training set and a test set, usually with an 80:20 ratio. The training set is used to train the models, and the test set is used to measure model performance on unseen data.
- After preprocessing, balancing, and scaling, the dataset is divided into training and test sets (80:20 split). Various classification algorithms, including Logistic Regression, KNN, Naive Bayes, Decision Tree, Random Forest, Ad- aBoost, Gradient Boosting, and CatBoost, are applied. To evaluate these models, several performance metrics are used: Accuracy, which measures the proportion of correct predictions; Precision, indicating the model's abil- ity to correctly identify relevant instances; Recall, which assesses the model's capability to find all relevant cases; F1-score, the harmonic mean of precision and recall; and ROC-AUC score, which evaluates the model's capacity to distinguish between the attrition and non-attrition classes.

The model with the highest performance based on these metrics is selected and integrated into the HR analytics system. This facilitates data-driven

strategies to reduce attrition and enhance employee retention.

PURPOSE

Employee attrition is a serious and persistent issue to modern-day organizations, with direct implications on operational continuity, team performance, and bottom-line performance. Since the average rate of attrition in India stands at close to one-fifth of the total workforce, organizations are increasingly under pressure to control the risks of employee attrition. Replacing experienced people not only costs a lot to recruit and induct but also leads to the loss of company- specific knowledge and the loss of productivity in the team. All these organizational concerns having been identified, fore- casting and explaining employee attrition has emerged as an imperative research area, especially within strategic human resource management.

Historically, HR professionals have depended on past meth- ods like exit interviews, employee satisfaction surveys, and reviews at regular intervals to comprehend the reason behind attrition. Although the above-mentioned methods can offer qualitative results, they are not good at predicting turnover in advance. They gather subjective views after an employee has made up his or her mind to exit, therefo limiting intervention in a timely manner. In addition, these traditional tools tend to be non-scalable, especially in large organizations with complex workforce dynamics. This calls for a change from reactive, survey-based attrition monitoring to data-driven, predictive approaches. The overall aim of this research is to create a machine learning-based framework capable of precisely predicting employee attrition as well as determining the root causes of the same. The objective is to enable organizations to forecast which employees are most likely to leave, identify why they are disengaged, and institute preventive measures to drive retention. In an effort towards this aim, the study employs a dataset of 1,470 employee records with 35 different features such as demographic information, job functions, satisfaction with work environment, compensation levels, and performance metrics.

One of the main contributions of this study is its comparative analysis of nine supervised machine learning models: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Naıve Bayes, K-Nearest Neighbors, XGBoost, and Cat Boost. Through systematic benchmarking and comparison of each model on accuracy, precision, re- call, F1-score, and ROC-AUC metrics, the research seeks to identify which algorithms provide the most accurate and interpretable predictions in human resource analytics. The comparative benchmarking enables a better picture of various machine learning approaches' performance against workforce attrition data. The research also seeks to determine the most significant factors driving attrition. These findings will not only enhance the model's predictive validity but also offer actionable intelligence for HR professionals. Job satisfaction, work-life balance, overtime, and levels of compensation are anticipated to emerge as prominent predictors, providing a

basis for targeted managerial action. Understanding these variables helps decision-makers segment the workforce based on attrition risk and craft customized interventions that align with employee needs.

MAJOR RESEARCH FINDINGS

In this proposed system, we utilized a dataset available through IBM Watson Analytics, which includes information about IBM data scientists. The dataset comprises HR-related data from 1,470 employees, featuring 35 key attributes. Among these employees, 1,233 were categorized as active, labeled with 'no' for attrition, while the remaining 237 were former employees, classified under 'yes' for attrition in our analysis The

dataset used in this study comprises a total of 35 columns, categorized into numerical and categorical data. Specifically, there are 26 numerical columns, which capture quantitative aspects of the employees' HR information. "Null" or "un- defined" values and duplicate records are identified, as these could potentially impact the accuracy of the model training and lead to erroneous predictions. Fortunately, the dataset contained no null or undefined values, and no duplicate entries were found. Exploring Unique Values of Categorical Attributes involves analyzing the diversity within categorical variables to better understand their distribution and potential impact on analysis. For example, the Attrition variable has two distinct values: Yes and No, representing employees who left and those who stayed. Business Travel includes three categories: Travel Rarely, Travel Frequently, and Non-Travel, reflecting the travel frequency of employees. Similarly, The Department attribute is categorized into Sales, Research '&' Development, and Human Resources, while the Education Field catergorizes as Medical, Life Sciences, and Technical Degree. Gender is represented by male and Female. There are other features which includes categorical data like this. Understanding and summarizing these unique values provides meaning insights into the categorical variables as catergorical values cannot be analyzed like numerical values. Exploratory Data Analysis (EDA) plays a crucial role in understanding and uncovering underlying patterns and trends in the dataset. For better understanding employee attrition, we begin by visualizing the distribution of employees with and without attrition, using pie charts or histograms to illustrate the count of employees in each category. Additionally, pie charts are used to display the percentage distribution of attrition, offering a clear view of the overall attrition rate. Further, EDA involves a detailed feature- based analysis, not only for categorical variables but also for numerical features. In the data preprocessing phase, categorical features are transformed into numerical representations to facilitate model training. For binary categorical variables such as gender, label encoding is applied, assigning numeric values to each category. For multi-category variables, such as Busi- ness Travel, Department, Education Field, Job Role, Marital Status, and Overtime. One-Hot Encoding is employed to create binary columns for each category, ensuring that each category is appropriately represented without introducing any ordinal relationships. Further to improve the efficiency of the model and reduce multicollinearity ,highly correlated features are identified and removed. Features with hign correlation, such as Job Level, Monthly Income, Years At Company and YearsInCur- rent Role are dropped from the dataset to minimize redundancy. In the next step, the dataset is first split into independent and dependent variables, where the independent variables (X) include all features except the target variable "Attrition", while the dependent variable (Y) represents the "Attrition" status. The dependent variable is chosen in such a way that this depends on the other independent variables. To address the class imbalance in the dataset, where the majority class (no attrition) is significantly higher the minority class (attrition), the Synthetic Minority Oversampling Technique (SMOTE) is employed. SMOTE generates synthetic samples for the minority class, so that there is an equal distribution of "attrition" and "no attrition" classes. By removing imbalance in the data set, the model does not become biased towards predicting the majority class. Following the balancing of the dataset, feature scaling is applied to standardize the independent variables, which is particularly important for models sensitive to the scale of input data, such as distance based algorithms. Either min- max scaling or standardization can be used. By standardizing the features, we improve the

model's convergence rate during training and thereafter its overall performance. The dataset is now balanced and scaled and is ready for the next steps in model development. The results of EDA are shown for the most important features which lead to employee attrition. The results of EDA are shown for the most important features which lead to employee attrition.

The chart in Fig. 1 shows that the employee attrition rates peak among those aged 28 to 32. As employees get older, the chances of them leaving the organization tend to decrease, because they prioritize job stability and family obligations more during this phase of their careers. In contrast, younger employees, especially those between 18 and 20, are more inclined to leave the company as they often seek different career opportunities for their individual growth and success.

Fig. 1: Correlation between employee age and attrition rates

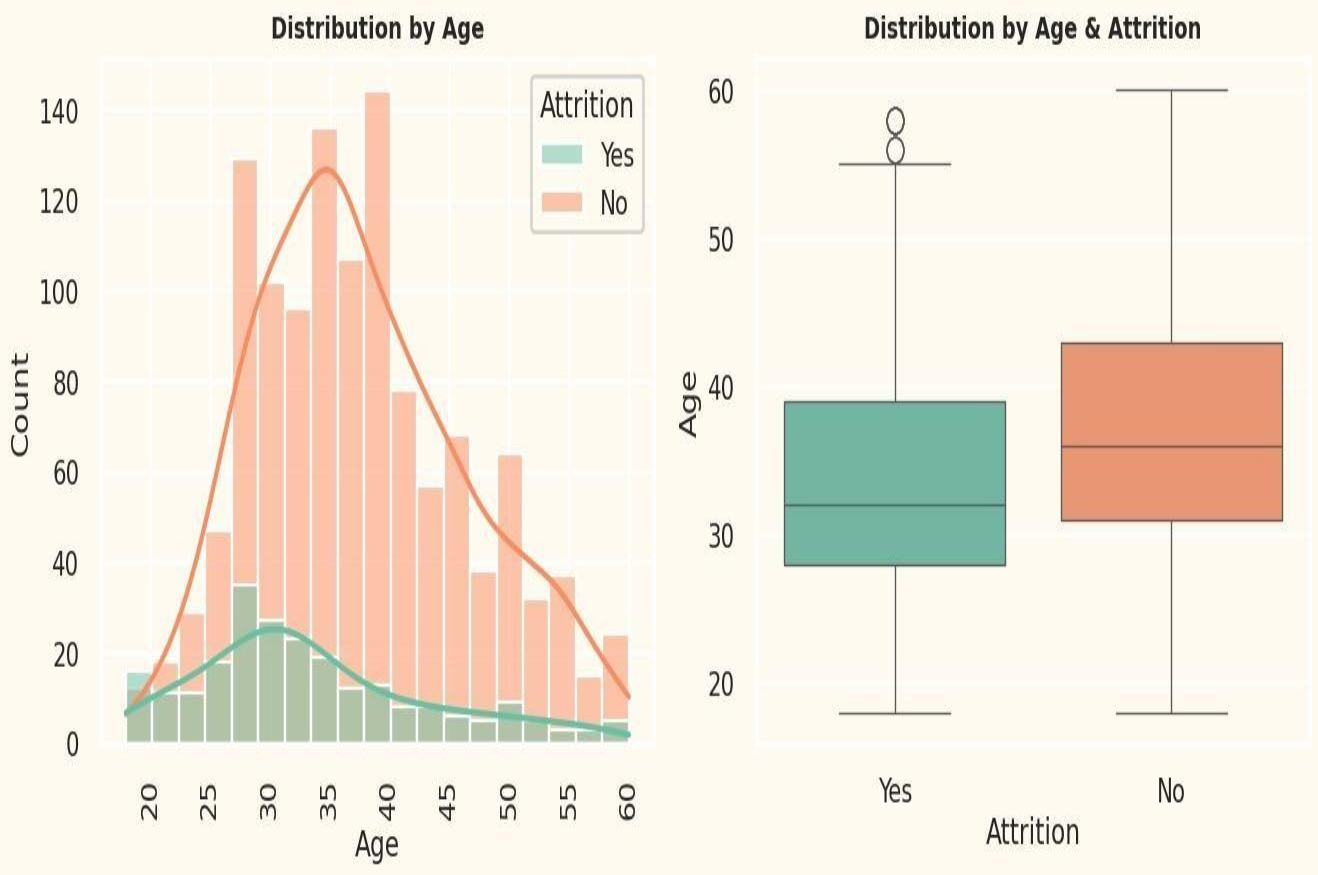
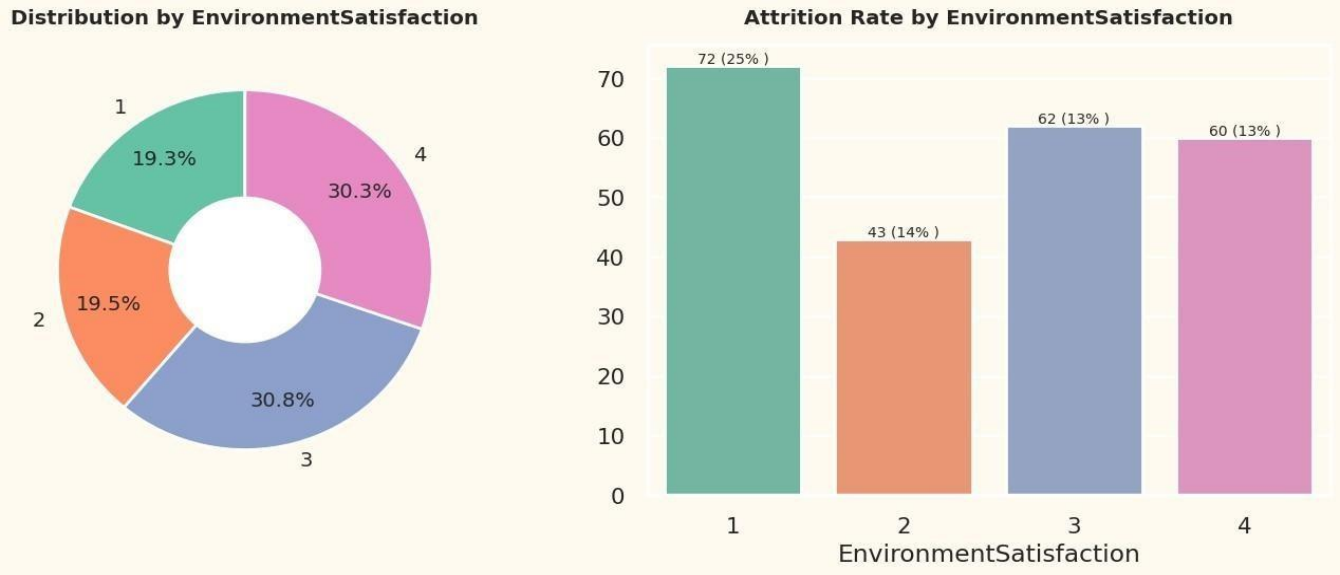


Figure. 2: Employee attrition across levels of environmental satisfaction



The distribution of employees across different levels of environmental satisfaction is presented in Fig. 2. The majority of employees report either high satisfaction (Level 4, 30.3%) or moderate satisfaction (Level 3, 30.8%). In contrast, Levels 1 and 2 represent a smaller portion of the workforce. Attrition rates, as depicted in the corresponding bar chart, show a clear trend: employees with the lowest environmental satisfaction (Level 1) experience the highest attrition rate at 25%, followed by Level 2 at 14%. Meanwhile, those reporting moderate (Level 3) and high (Level 4) satisfaction have comparatively lower attrition rates, both at 13%.

The analysis of job satisfaction, illustrated in Fig. 3, highlights a clear link between satisfaction levels and attrition. Most employees report high (Level 4, 31.2%) or moderate (Level 3, 30.1%) satisfaction, while fewer fall into the lower levels (Levels 1 and 2). The bar chart shows that attrition is highest at Level 1 (22%) and Level 2 (16%), whereas Levels 3 and 4 exhibit significantly lower attrition rates (both at 11%). These findings confirm a strong negative correlation between job satisfaction and attrition: higher satisfaction is associated with lower turnover.

Figure. 3: Employee attrition across levels of job satisfaction

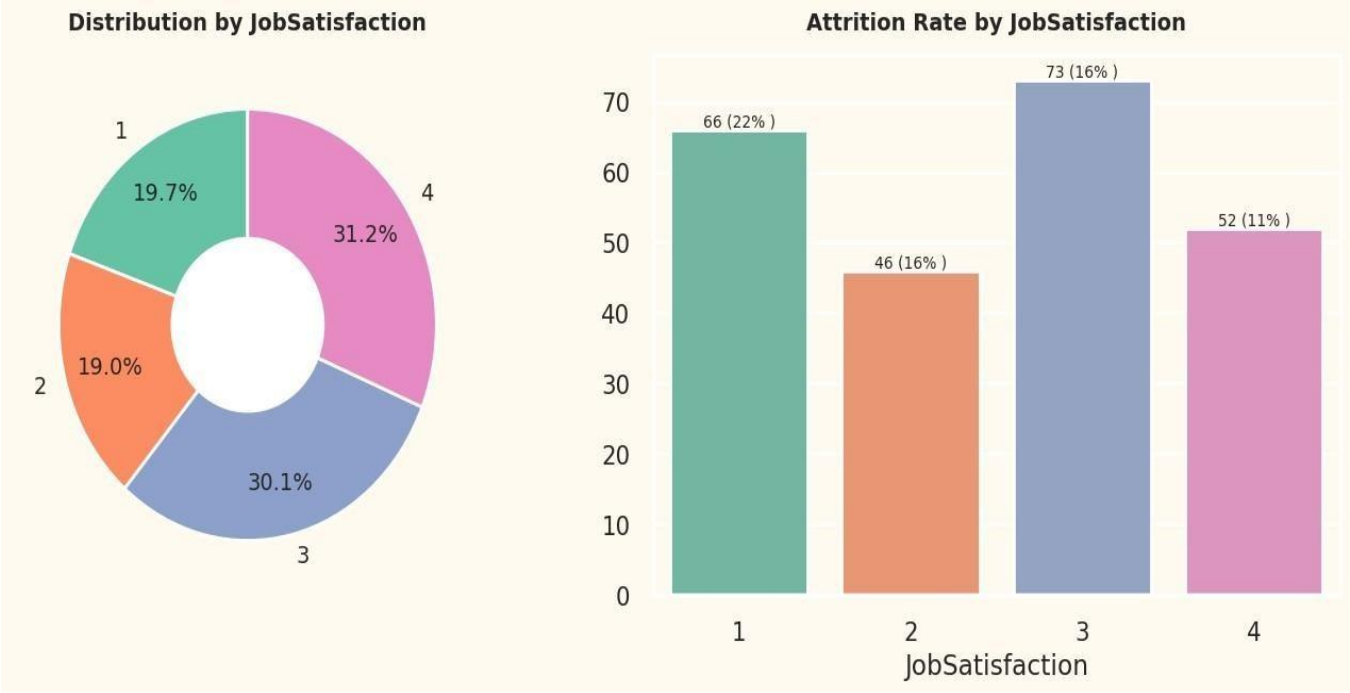


Figure. 4: Trends in attrition with varying educational qualifications

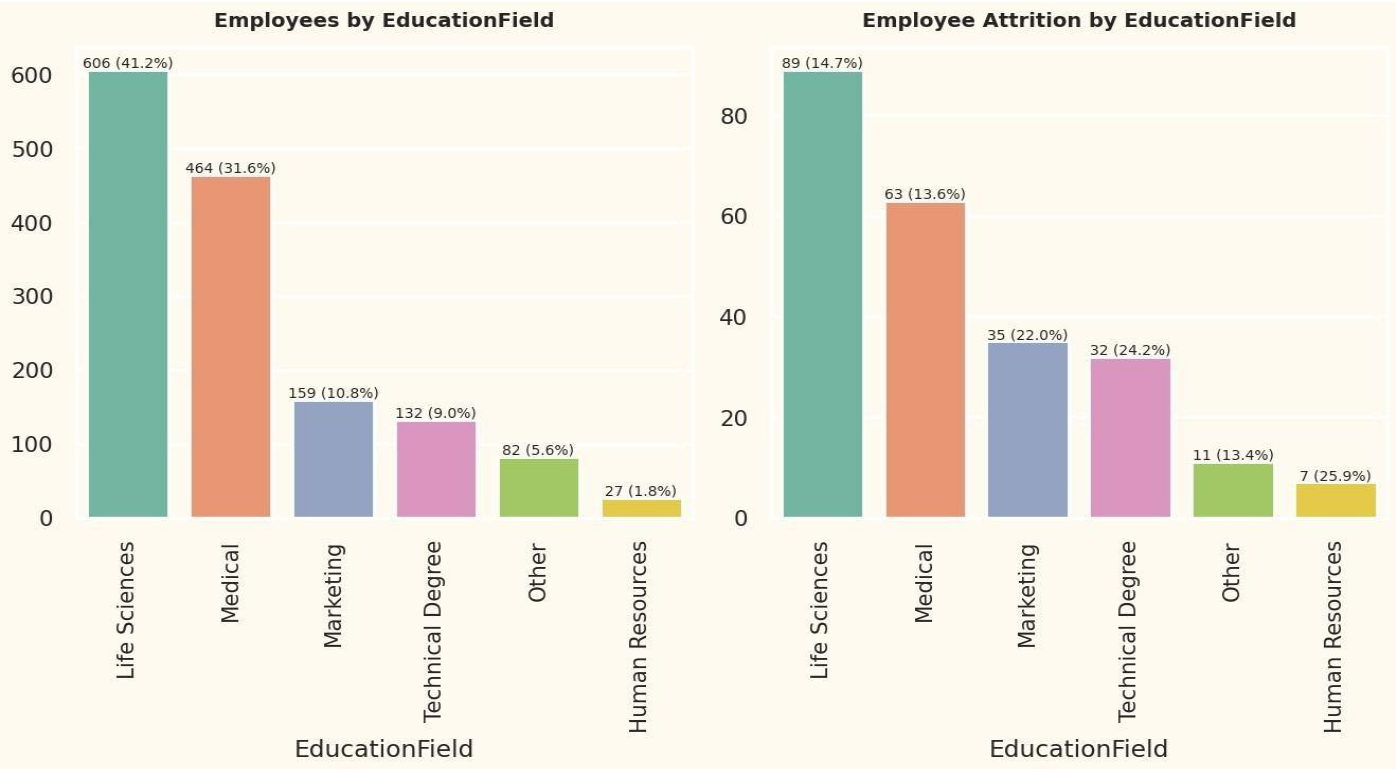


Figure. 5: Impact of relationship satisfaction on attrition levels

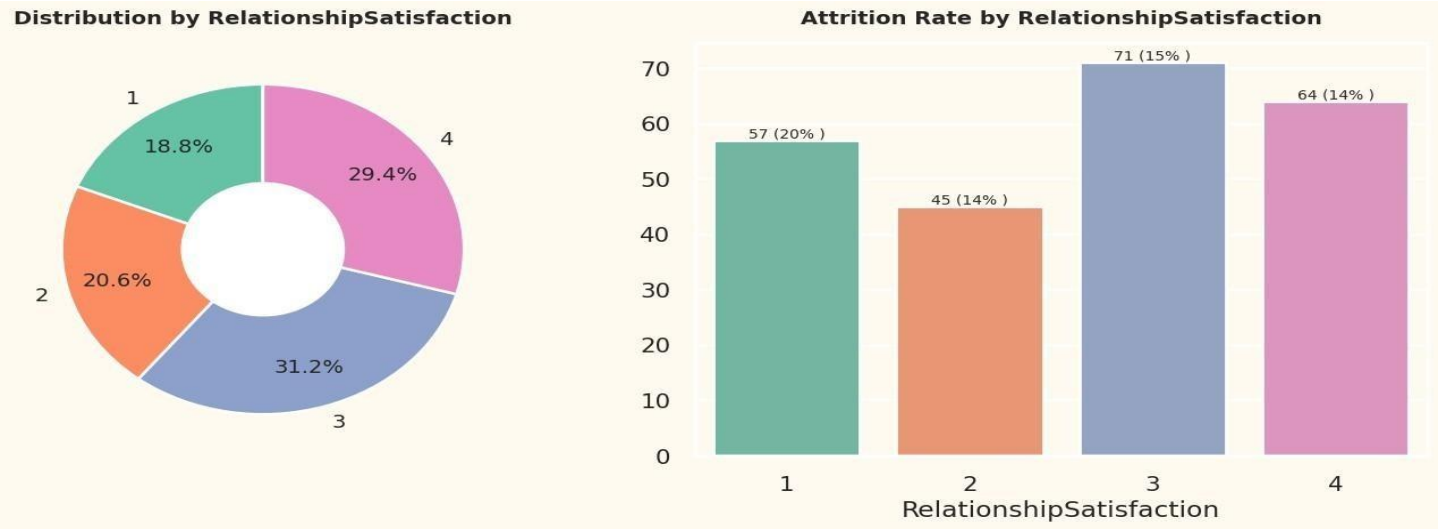
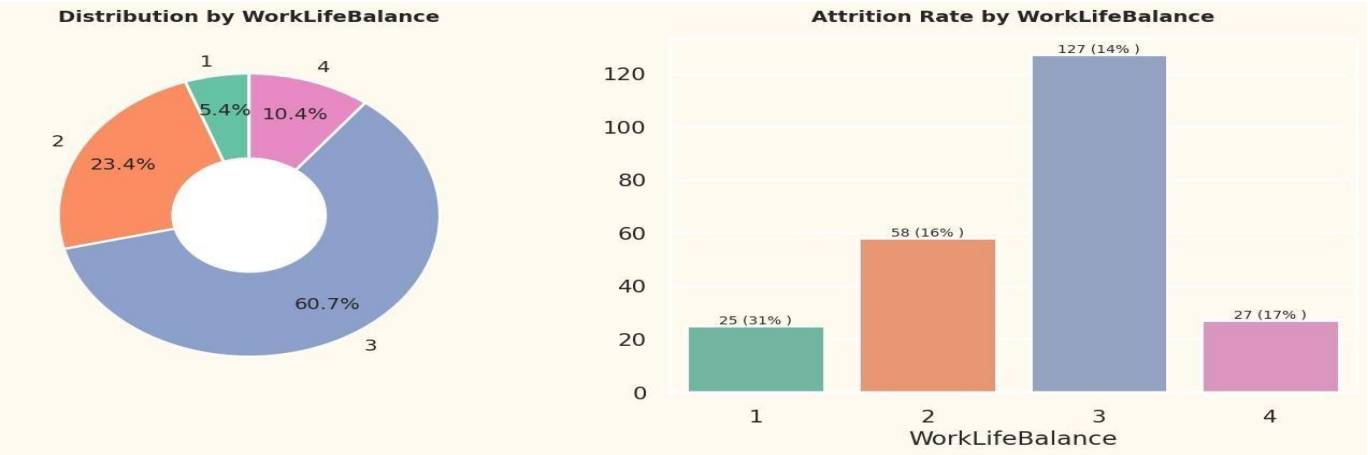


Fig. 5 illustrates the distribution of relationship satisfac- tion among employees, with the majority reporting moderate 31.2% or high 29.4% satisfaction. Lower satisfaction levels

1 and 2 account for a smaller share. The attrition data reveals that employees at level 1 have the lowest attrition rate 10%, followed by level 2 at 14%. Those with moderate and high satisfaction levels 3 and 4 also show relatively low attrition rates of 14% and 15%, respectively. This indicates a moderate negative correlation between relationship satisfaction and attrition—employees with higher satisfaction tend to have slightly lower turnover rates.

Figure. 6: Effect of work-life balance satisfaction on employee turnover



In Fig. 6 shows that employees reporting high work-life balance (60.7%) have the lowest attrition rate at 14/%. In contrast, those with poor balance face a much higher rate (31%), while the moderate group is at 16%. This pattern suggests that inadequate work-life balance, due to long hours or low flexibility, may drive turnover. Supporting a healthy balance is therefore important for retention and employee well- being

The chart Fig. 7 indicates that 71.7% of staff members never work overtime and 28.3% of staff members work overtime. Matching attrition statistics indicate that a mere 10% of non-overtime employees quit the company, while a significantly greater 30% of overtime workers did. This shows that overtime workers are almost three times more likely to quit than employees who work standard hours.

Figure. 7: Effect of overtime on employee retention rates

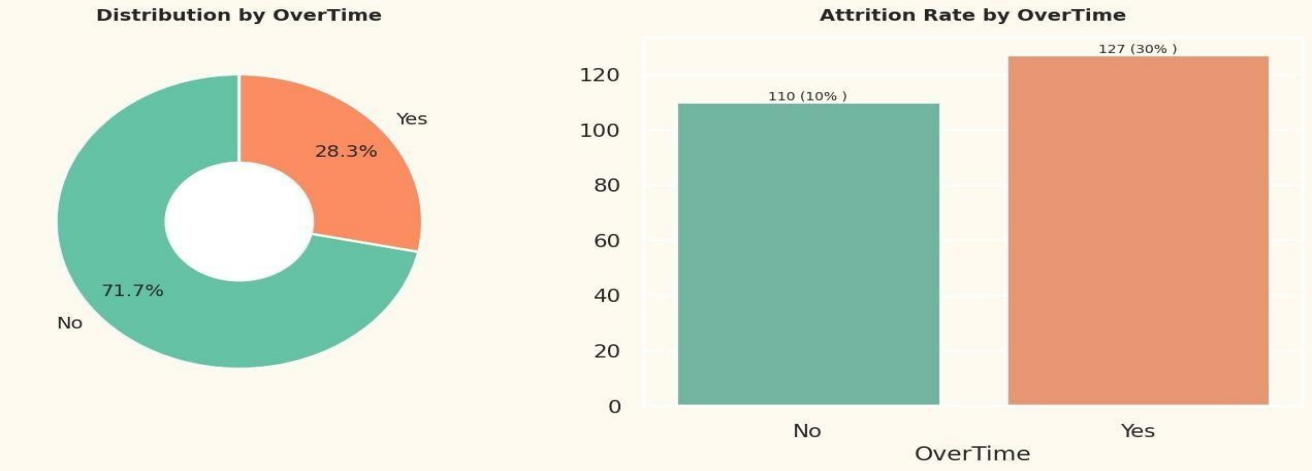
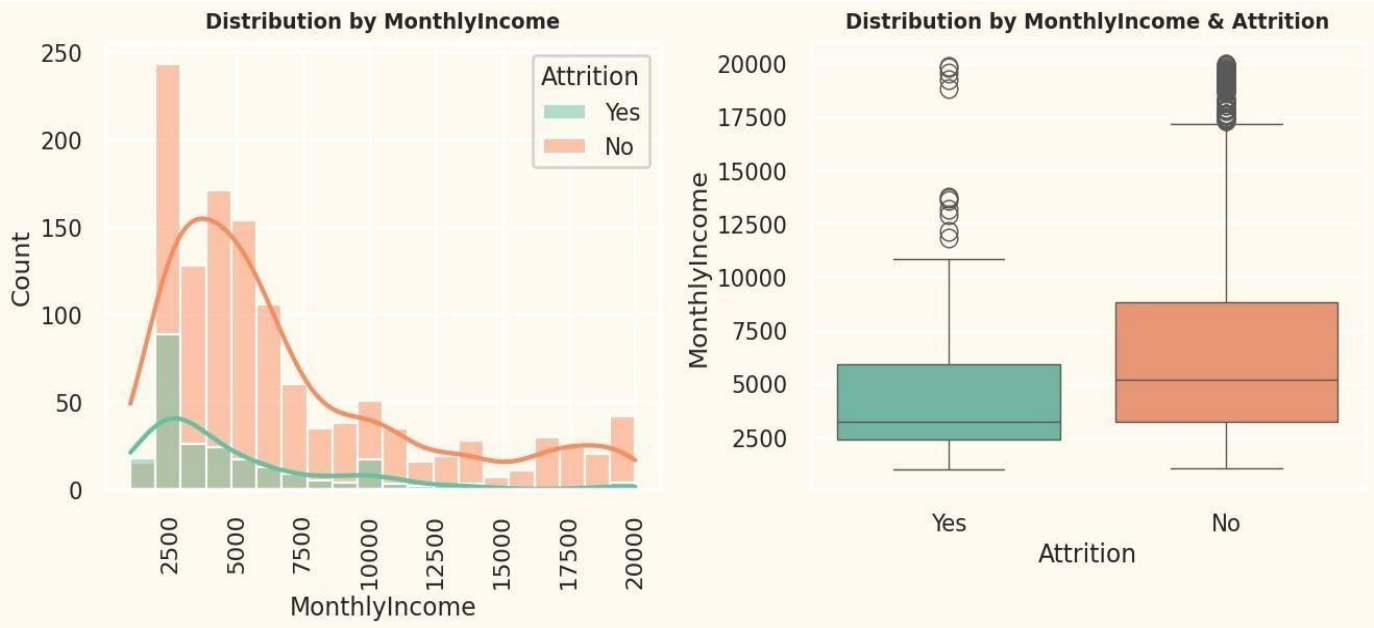


Figure. 8: Influence of monthly income on employee retention



The density plot in Fig. 8 indicates that the employees who remained with the company have a wider income distribution, peaking at 2500. Their counterparts who departed the organization are predominantly in lower income groups. The boxplot also reinforces the pattern of significantly lower median income for the ones who departed, as well as a narrower interquartile range and less high-income presence. These findings imply that reduced monthly earnings could be one of the causes of worker turnover

The distribution of pay in Fig. 9 increases is slightly skewed towards pay increases between 12% and 14 %. Surprisingly, departing employees had slightly higher median pay increases than those who remained. Nevertheless, the overall trend indicates that low pay increases might result in higher turnover. Though greater increases may serve as a motivating force, lower-increase recipients are likely to look for opportunities elsewhere.

Figure. 9: Salary hike distribution showing attrition versus retention trends

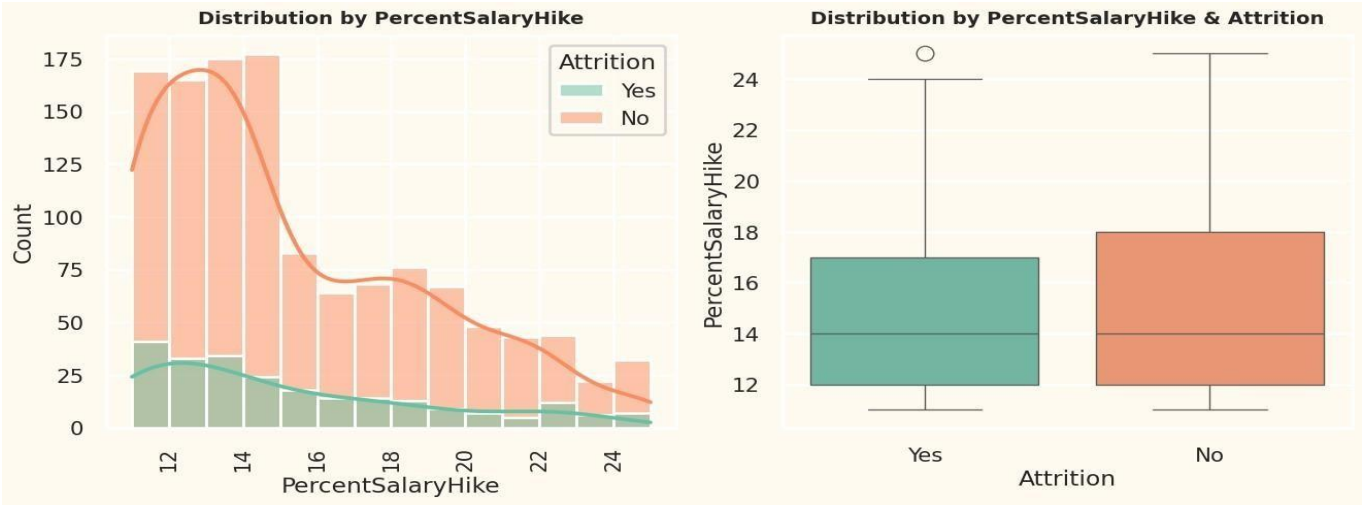


Figure. 10: Impact of total working years on employee turnover rates

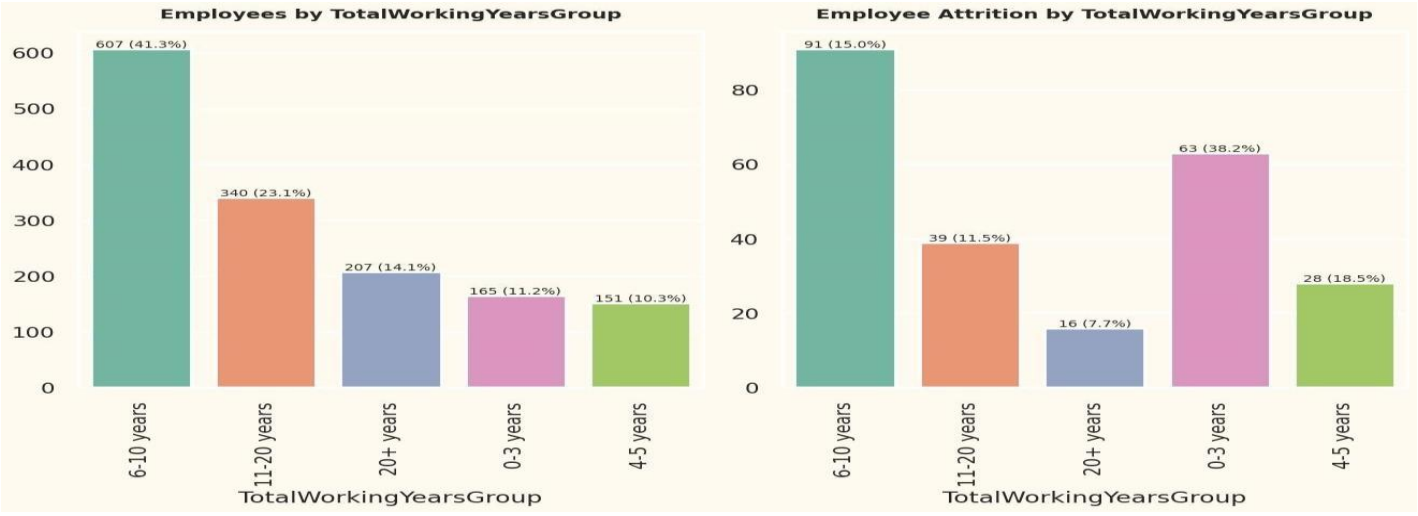


Fig. 10 indicates 41.3% of staff members have been in the company for 6 to 10 years, and 23.1%

for 11 to 20 years, thus forming the two biggest groups. The greatest attrition is among the staff members with 0 to 3 years of experience, standing at 15%. This indicates that new staff members are more likely to exit the company, perhaps as a result of unrealized expectations or adaptation problems. As workers gain more experience, they seem to be more apt to remain, perhaps because of improved pay, familiarity with the job, or more solid workplace relationships.

Fig. 11, the majority of employees, approximately 43%, reside between 0 and 5 kilometers from the workplace. The attrition rate in this category is lowest, at 13.8%. As the dis- tance from the workplace increases, the number of employees gradually decreases, while the attrition rate rises. Employees living more than 20 kilometers away experience the highest attrition rate at 22.1%. These findings indicate that longer commuting distances may contribute to increased stress and dissatisfaction, which in turn affects employee retention.

Fig. 12 shows the distribution of employees shows that a significant majority (71%) rarely travel for business, while 18.8% travel frequently and 10.2% do not travel at all as shown in Figure 13. This travel pattern correlates with attrition rates: frequent travellers face the highest attrition rate at 24%, suggesting that the demands of regular travel might lead to stress and work-life imbalance, contributing to their decision to leave the company. In contrast, those who travel rarely experience a moderate attrition rate of 14%, and non-travellers have the lowest rate at 8%.

Figure. 11: Attrition rates based on distance from workplace

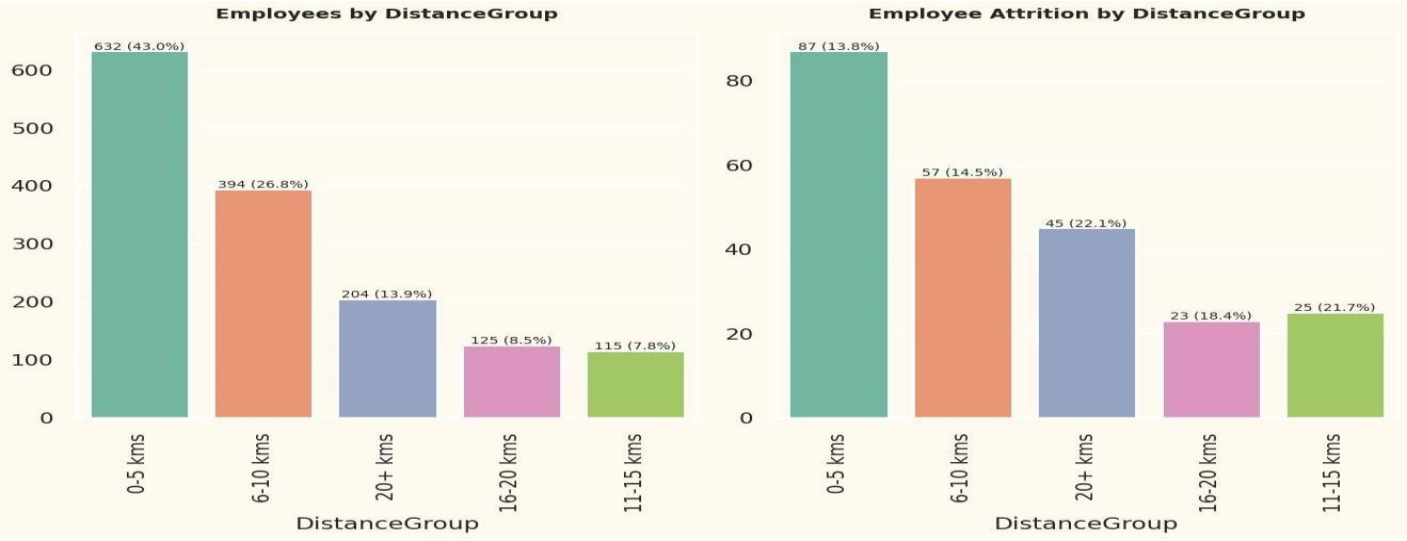
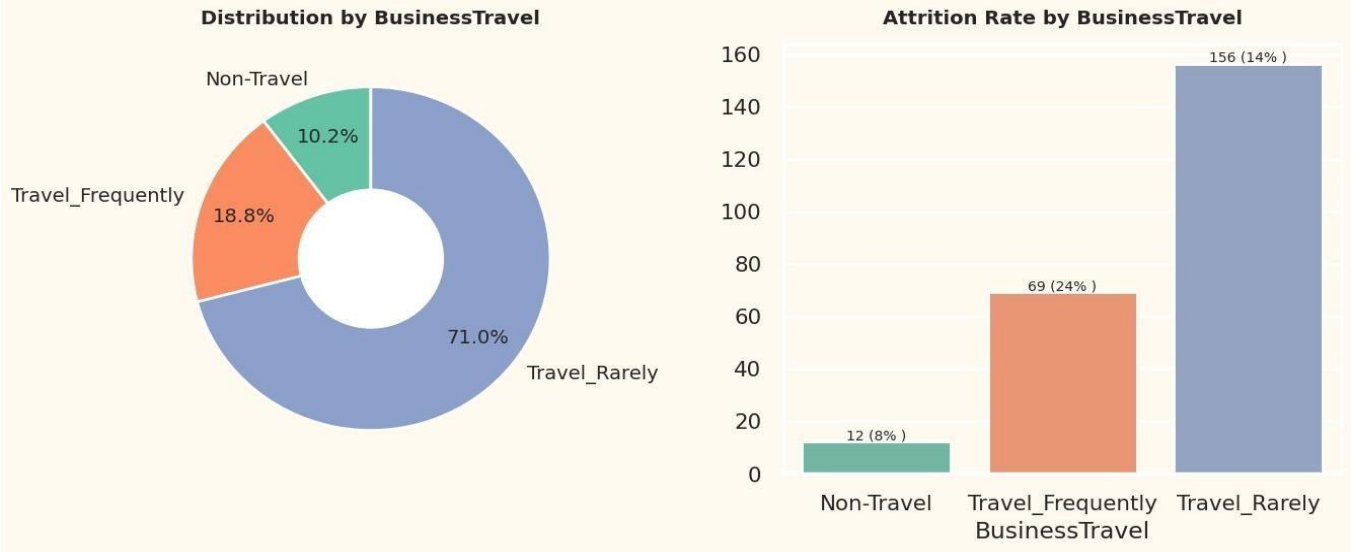


Figure. 12: Employee attrition rates by frequency of business travel.



This could indicate that frequent travellers may be less satisfied with their job roles or work-life balance, and the company's culture and support for its travelling employees could significantly impact their retention. Addressing these concerns could improve job satisfaction and lower turnover rates.

Lastly, the distribution of employees by marital status reveals that 45.8 % are married, 32% are single, and 22.2% who are divorced. The single employees experience the highest turnover at 25%, compared to 12% for divorced employees and a notably lower rate of 10% for married employees. These trends can be interpreted in several ways: married employees may exhibit greater stability and commitment to the company due to their financial security and family obligations, while single employees, particularly those in younger demographics, may prioritize career advancement and personal growth, lead- ing to their higher attrition. Additionally, divorced employees might encounter unique challenges in managing their work-life balance, contributing to their elevated turnover rates.

Model Building

After establishing our objectives and thoroughly preparing and analyzing the dataset, we moved forward with designing a prediction model aimed at identifying employees who may be at risk of leaving the company. During the model construction phase, we recognized the importance of utilizing a training set that includes instances from a previously classified population. This would enable us to train the model effectively to classify new observations, which will constitute our test set, where the class attribute is absent. To enhance the model's predictive capabilities, it is crucial to train it on a substantial and consistent number of observations. By utilizing a larger dataset during training, we can significantly improve the precision of the machine learning algorithms employed. Therefore, the original dataset was split into two parts with an 80:20 ratio. Training Set: 80% of the total dataset was used for training. The model learned the relationships within this data. Test Set: The remaining 20% observations were done using the test set, used for testing and validating each model's performance. Here the predicted results were compared against actual outcomes. This split encourages effective evaluation and fine-tuning of the model, so that the model can give a good performance when encountered with unknown data. The training dataset has used different classification models on training set. These

models are:

Logistic Regression: A widely used linear model for binary classification-related applications that estimates the probability of an event based on input features. It works well when the relationship between independent and dependent variables is linear.

K-Nearest Neighbors (KNN): A memory-based, flexible model that classifies data points according to the majority class among *K* nearest neighbours. It is effective for smaller datasets but may become computationally expensive when working with larger datasets.

Naive Bayes: A probabilistic model based on Bayes' Theorem, effective for applications like text classification. It works on the assumption of feature independence and normal distribution.

Decision Tree Classifier: This model divides data by making decisions on the basis of feature values. It can handle both categorical and numerical data but might perform poorly if it gets overfit without proper pruning or regularization.

Random Forest Classifier: This model creates multiple decision trees and combines their predictions. It helps reduce overfitting and enhances generalization, making it a robust choice for diverse datasets.

AdaBoost Classifier: This model integrates multiple weak classifiers into a stronger one by adjusting the weights of misclassified data points in successive iterations. It is popular for improving model accuracy in challenging tasks.

Gradient Boosting Classifier: In this model, classifiers are built iteratively, each correcting the errors of the previous one. While it is highly effective for complex datasets, improper regularization may lead to overfitting.

CatBoost Classifier: This model efficiently handles cate- gorical features without requiring extensive pre-processing. It focuses on reducing overfitting while achieving high perfor- mance, especially on large datasets with mixed data types.

RESULTS

In this phase, Table 1 and Table 2 show the effectiveness of the selected models was assessed. For each algorithm, the outcomes of the predictions were organized into a con- fusion matrix. This matrix displays the predicted values in the columns and the actual values of each instance from the test set in the rows. By analyzing the confusion matrix, key performance metrics, namely accuracy, precision, recall, and F1-score, and the results were analyzed to identify the model that achieved the best predictive performance.

TABLE I: Comparison of models based on accuracy and precision

Model	*Accuracy (Train)*	*Accuracy (Test)*	*Precision*
Logistic Regression	85.44	84.00	0.848
KNN	90.87	87.85	0.808
Naïve Bayes	76.41	75.50	0.699
Decision Tree	100.00	86.23	0.846
Random Forest	100.00	92.51	0.968
AdaBoost	93.05	90.49	0.923
Gradient Boosting	96.55	92.31	0.960
XGBoost	100.00	92.51	0.956
CatBoost	99.65	92.71	0.954

TABLE II: Comparison of models based on Recall, ROC-AUC, and F1 Score

Model	*Recall*	*ROC-AUC*	*F1 Score*
Logistic Regression	0.823	0.910	0.836
KNN	0.987	0.961	0.889
Naïve Bayes	0.885	0.867	0.781
Decision Tree	0.881	0.863	0.860
Random Forest	0.877	0.976	0.930
AdaBoost	0.881	0.955	0.910
Gradient Boosting	0.881	0.965	0.930
XGBoost	0.889	0.968	0.921
CatBoost	0.885	0.971	0.930

MANAGERIAL IMPLICATIONS AND PRACTICAL IMPLICATIONS

The findings from this research have important managerial implications, especially for decision-making in the Human Resources function. As attrition continues to affect organiza- tional stability and profitability, managers are under growing pressure to detect early warning signs and introduce pre- ventive measures. Machine learning models provide a more structured and data-driven method of interpreting employee behavior. Managers can use predictive insights to identify patterns and trends indicating likely turnover, enabling them to target resources and assistance to those places where they are most urgently required. This requires a change in culture for organizations as well, when HR is no longer just a support function but a strategic partner facilitated through advanced analytics. Implications in terms of practical implementation, the usage of high-performance models like XGBoost and CatBoost creates a revolutionary moment for HR functioning.

The above models can easily be plugged into existing HR IT systems to keep track of the risks of attrition in real time. In this manner, HR teams

can segment their employees by how likely they are to leave and probe the individual drivers of such risk—be it dissatisfaction with pay, absence of recognition, lack of career progression, or inadequate work-life balance. This degree of granularity makes it possible to have intervention at a point of specificity, eliminating over-reliance on time-consuming traditional practices like exit interviews and one-size-fits-all employee satisfaction surveys. Additionally, the application of machine learning supports the effectiveness of HR departments and enables them to act proactively instead of reactively.To effectively restrain attrition, businesses must engage in a holistic strategy that combines machine learning insights with people-centric programs. Firstly, they need to incorporate models such as CatBoost and XGBoost into their HR infrastructure to get high-risk employees automatically flagged. These predictive systems can be used to construct in-depth employee profiles, exposing key drivers of disengage- ment or dissatisfaction. With risks discovered, businesses can implement customized responses such as providing flexible working times, raising compensation transparency, and designing customized development plans. Further, by building a culture of internal mobility where workers are urged to try out new jobs within the organization, voluntary turnover can also be lessened. To maintain this strategy, organizations need to create ongoing feedback loops where actual-world results feed into the predictive models so that they become increasingly accurate with time. Of comparable significance is empowering HR staff through data literacy and model interpretation training, so that not only are insights created but also interpreted and translated into action. When predictive analytics are combined with human intuition and empathy, organizations are in a position to better retain high-performing talent, increase employee engagement, and create long-term organizational success.

RESEARCH LIMITATIONS

One of the main restrictions is in the consistency of model performance between algorithms. While ensemble models like CatBoost and XGBoost performed better in accuracy and ROC-AUC, other more basic models like Na¨ıve Bayes and K-Nearest Neighbors were unable to cope with the complication and skew in the dataset. The class distribution was also dominated by non-attrition cases, which was a concerning aspect for classification algorithms and may have biased their predictions. Even with the use of the SMOTE method to balance the class distribution, some models continued to underperform in detecting minority class instances, hence inhibiting their capability to reflect actual attrition risks in real- world deployment.

Overfitting was also a concern in some models, such as Decision Trees. These models performed virtually flawlessly on the training set but suffered a significant performance loss on the test set. This difference between training and test metrics indicates that some models learned patterns specific to the training set instead of generalizable knowledge. Overfitting in this manner decreases the strength and dependability of predictions in real-world operating environments, where the model needs to generalize to novel and unknown data. In order to address this issue, further implementations would be aided by more comprehensive cross-validation processes, regularization methods, and iterative retraining from new data inputs.

Another limitation lies in the lack of generalizability due to the character of the dataset employed. Data used for this research came from one organization alone IBM, which had special culture-related, structure-specific, or policy-related characteristics different from those of other firms or sectors. Accordingly, the results and model behaviors obtained from this dataset may not be applicable as-is to other corporate

settings. Some of the features involved in this analysis, e.g., internal job positions or certain satisfaction metrics, might not be available elsewhere, thereby limiting portability and scalability of resulting models. The interpretability of the machine learning models is also a limitation, particularly for ensemble algorithms such as CatBoost and XGBoost. Although these models provided high predictive performance, their internal mechanisms are usually unclear to non-technical users. This transparency can limit the acceptance and real usage of model outputs within human resource departments, where decision-makers may need explainable justifications for predictive recommendations.

From the point of view of data governance, feature avail- ability and privacy concerns also restrict the study's real-world utility. Some of the data that is utilized to train the models, like detailed performance scores, individual demographics, and satisfaction metrics, would not be easily accessible in every organization because of legal, ethical, or practical reasons. Additionally, the utilization of employee sensitive information also poses privacy issues that need to be resolved through proper anonymization and adherence to data protection laws. These demands can require adjustments to the model or its input parameters, possibly impacting its reliability or accuracy in usage.

Lastly, it must be emphasized that no predictive platform could possibly supplant the sophisticated judgment of sea- soned human resource professionals. Although the models formulated in this research yield useful advance warnings and data-driven observations, their efficacy hinges on being incorporated within a more encompassing strategic process that integrates managerial judgement, policy redesign, and ongoing employee participation. Predictive analytics must be viewed as a decision support system, not as a sole intervention.

ORIGINALITY

While several prior studies have explored employee attrition using various demographic and compensation variables, our work distinguishes itself by not only integrating 35 features across demographic, job-related, and organizational dimen- sions and benchmarking nine supervised learning algorithms (from Logistic Regression to XGBoost and CatBoost), but also by employing a diverse set of visualizations such as box plots, histograms, bar charts and pie chart which translate complex model outputs into actionable recommendations for HR teams. These graph-based presentations clarify which factors most strongly drive turnover risk, enabling practitioners to locate and prioritize selective measures more effectively than in earlier, less interpretable frameworks.

CONCLUSION AND FUTURE RESEARCH WORK

In this research, several loopholes were filled by including 35 unique features and using nine various machine learning classifiers. EDA helped to visualize the results, to make the findings precise and understandable. This approach enables organizations to identify high-risk segments of employees at an early stage and implement timely and focused measures to enhance retention and overall employee satisfaction. Out of the comparative study of machine learning models under different evaluation metrics (Accuracy, Precision, Recall, ROC_AUC, and F1 Score), the CatBoost Classifier is always among the best-performing models. It has very high test accuracy at 92.71%, good precision (0.954), and good recall (0.885), with superior ROC_AUC (0.971) and F1 Score (0.930). The XGBoost Classifier is also competitive, equalling CatBoost in both training accuracy and test accuracy, and having similar measurements in the other metrics. On the opposite end, the Na¨ıve Bayes model has the lowest test accuracy (75.50%), precision (0.699),

and lowest F1 Score (0.780) among all the models considered. The Linear Regressor, although compar- atively stable, also lags behind the ensemble approaches in most performance indicators. Ensemble learning algorithms, especially CatBoost, XGBoost, and Gradient Boosting, always perform better than more basic models such as Na¨ıve Bayes, KNN, and Decision Trees. This suggests that boosting-based solutions are better suited to the problem in question, most probably because they can minimize bias and variance by iteratively refining the model. Future work can build on this study by adding more types of data, such as real-time employee feedback, health and wellness information, or trends from the wider industry. This would help create stronger, more adaptable models to predict attrition. Research could also look at designing customized retention plans by using more detailed risk profiles and improving how easily we can understand model results, for example through new explanation tools. Testing these methods across different industries and cultures would show how well they work in various settings. Finally, there is potential to expand this approach for long-term predic- tions and to link it directly with employee growth platforms, so companies can better spot possible turnover and respond in a more focused, culturally aware way\

ACKNOWLEDGMENT

The authors of the paper acknowledge the support of the Department of Electrical Engineering, Techno International New Town and the Department of Information Systems & Business Analytics, Indian Institute of Management Ranchi, to accomplish the presented work.

REFERENCES

- C. Trevor and Y. Ko, "Impending-exit period and employee performance: Rethinking human capital disruption," *J. Manage.*, 2025. [Online].
- Available: https://doi.org/10.1177/01492063251316464
- C. Li *et al.*, "A study in the relationship between supportive work en- vironment and employee retention: Role of organizational commitment and person–organization fit as mediators," *SAGE Open*, vol. 10, 2020. [Online]. Available: https://doi.org/10.1177/2158244020924694
- S. Yadav, A. Jain, and D. Singh, "Early prediction of employee attrition using data mining techniques," in *Proc. 8th Int. Adv. Comput. Conf. (IACC)*, 2018, pp. 349–354. [Online]. Available: https://doi.org/10.1109/IADCC.2018.8692137
- N. Bandyopadhyay and A. Jadhav, "Churn prediction of employees using machine learning techniques," *Tehn. Glas.*, vol. 15, no. 1, pp. 51–59, 2021. [Online]. Available: https://doi.org/10.31803/tg-20210204181812
- R. Gupta, A. S. Chand, S. Solanki, T. Gautam, and N. Garg, "A comparative analysis of logistic regression and SVM for employee churn prediction," in *Proc. ICSES*, 2024, pp. 1821–1827. [Online]. Available: https://doi.org/10.1109/ICSES63445.2024.10762955
- S. Agarwal *et al.*, "AI based employee attrition prediction tool," *Lect. Notes Comput. Sci.*, vol. 14078 LNAI, pp. 580–588, 2023. [Online].
- Available: https://doi.org/10.1007/978-3-031-36402-0 54
- V. Lalitha, S. P. Raj, and P. Lokesh, "A novel analysis of employee attrition rate by maneuvering machine learning," *Adv. Sci. Technol.*, vol.
- 124 AST, pp. 469–477, 2023. [Online]. Available: https://doi.org/10. 4028/p-sw6mmb
- S. Madane and D. Chitre, "A survey of employee and customer churn prediction methodologies," *Adv. Math. Sci. J.*, vol. 9, no. 6, pp. 3955– 3962, 2020. [Online]. Available: https://doi.org/10.37418/amsj.9.6.76
- A. V. Pachghare, S. Deshmukh, and S. Salunkhe, "Employee churn prediction using machine learning techniques: A systematic review," in *Proc. 2nd IEEE Int. Conf. Data Sci. Netw. Secur. (ICDSNS)*, 2024. [On- line]. Available: https://doi.org/10.1109/ICDSNS62112.2024.10691112
- J. Limjeerajarat and D. Naparat, "Preserving talent: Employee churn prediction in higher education," in *Proc. Int. Conf. Inf. Syst. (ICIS)*, 2022. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85192542390
- D. K. Srivastava and P. Nair, "Employee attrition analysis using predic- tive techniques," *Smart Innov. Syst. Technol.*, vol. 83, pp. 293–300, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-63673-335
- K. M. Mitravinda and S. Shetty, "Employee attrition: Prediction, analysis of contributory factors and recommendations for employee retention," in *Proc. IEEE Int. Conf. Women Innov. Technol. Entrep. (ICWITE)*, 2022. [Online]. Available: https://doi.org/10.1109/ICWITE57052.2022.10176235