Predicting Employee Attrition:

Data Driven Insights on patterns , predictors and prevention of employee attrition

Group 11- 16075, 16220 ,16333

11/4/25

Final Project

Abstract

This report presents an advanced analysis utilizing machine learning algorithms to predict employee attrition and performance outcomes. Building on insights gained from prior descriptive exploration, this study develops predictive models using a comprehensive human resources dataset sourced from Kaggle. The dataset encompasses various employee-related attributes, including demographic details, job roles, performance evaluations, and satisfaction levels. Key findings from both descriptive and predictive analyses provide actionable insights for HR managers, organizational leaders, and workforce strategists. The results support data-driven decision-making, optimized talent retention strategies, and a deeper understanding of the factors influencing employee turnover and performance.

Content

List of Figures

List of Tables

Introduction

The human resources function plays a pivotal role in shaping organizational success, especially in today’s competitive and rapidly evolving business landscape. As companies strive to attract, develop, and retain top talent, understanding employee behavior and performance has become more critical than ever. High employee attrition rates and inconsistent performance can lead to significant operational and financial challenges. Therefore, leveraging data-driven approaches to analyze workforce dynamics is essential for strategic decision-making. This study focuses on employee attrition and performance using a comprehensive HR dataset, aiming to uncover the key factors that influence these outcomes. Through careful analysis, this research seeks to provide meaningful insights that can enhance talent management, reduce turnover, and support the development of a more engaged and high-performing workforce.

Description of the question

The primary objective of this project is to develop a robust predictive model capable of accurately forecasting employee attrition. By analyzing a range of factors—including demographic information, job roles, performance metrics, and behavioral patterns—this model aims to identify employees who are at a higher risk of leaving the organization.

The central question this study seeks to answer is: *What are the key drivers of employee attrition, and how can predictive analytics help in identifying and mitigating potential turnover?* By addressing this question, the project not only enhances understanding of workforce dynamics but also empowers organizations to take proactive measures. Ultimately, the insights generated can support strategic HR planning, improve employee retention, and significantly reduce the costs associated with turnover.

Description of the data set

The "Employee Attrition and Performance" dataset, available on Kaggle, offers comprehensive information on various factors influencing employee turnover and job performance within an organization. The dataset includes detailed records for 1,470 employees, encompassing 35 attributes related to demographics, job roles, compensation, satisfaction levels, performance ratings, and work environment. This rich dataset enables in-depth analysis of the underlying patterns and predictors of employee attrition, making it a valuable resource for developing data-driven HR strategies.

Attribute	Description
PerformanceID	Unique identifier for each performance review.
EmployeeID	Unique identifier for the employee being reviewed.
ReviewDate	The date of the performance review.
EnvironmentSatisfaction	Rating of the employee's satisfaction with their work environment.
JobSatisfaction	Rating of the employee's satisfaction with their job.
RelationshipSatisfaction	Rating of the employee's satisfaction with workplace relationships.
TrainingOpportunitiesWithinYear	Number of training opportunities available to the employee within the year.

TrainingOpportunitiesTaken	Number of training opportunities the employee has taken.
WorkLifeBalance	Rating of the employee's work-life balance.
SelfRating	The employee's self-assessment rating.
ManagerRating	The manager's rating of the employee's performance.
Attrition	Whether a employee has left the company or not

Table 1: Some Attribute Information

<u>Data Preprocessing</u>

To ensure the predictive model was efficient and easy to interpret, a thoughtful selection of variables was carried out during data preprocessing. The main objective was to keep features that offered valuable insights while removing those that added noise, redundancy, or potential ethical issues. This process was guided by factors such as the relevance of each variable, its potential overlap with others, and expert domain knowledge.

Unique identifiers like *EmployeeID*, *FirstName*, and *LastName* were removed since they are specific to individuals and do not provide meaningful predictive value. Including them could also lead to overfitting or raise privacy concerns. Similarly, *PerformanceID* was excluded due to its unclear meaning and possible redundancy.

Some categorical features, such as *Gender*, *Ethnicity*, *EducationField*, and *JobRole*, were replaced with grouped versions (e.g., *Gender_Grouped*, *Ethnicity_Grouped*) to reduce complexity and improve interpretability. These grouped forms were created based on domain knowledge to retain key distinctions while simplifying the data structure.

Additionally, certain tenure-related variables like *YearsAtCompany*, *YearsInMostRecentRole*, and *YearsWithCurrManager* were removed because they overlapped heavily with *YearsSinceLastPromotion*, which was kept as the most informative feature. Removing these overlapping variables helped reduce multicollinearity, contributing to a more stable and reliable model.

Important results of the Descriptive analysis

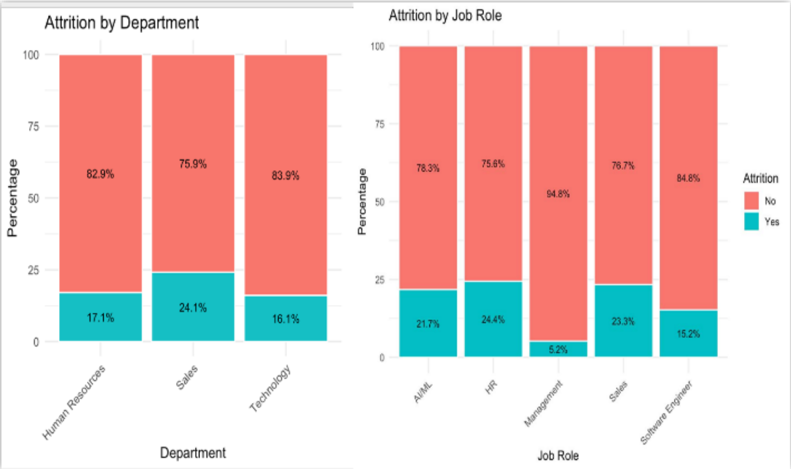

Fig 1: Attrition rate by department and job role

Attrition rates vary significantly across departments and job roles. Sales and Human Resources roles appear to have higher turnovers. perfection. Department and job roles are among the strongest predictors of attrition. This is intuitive as roles like sales or customer service are naturally high-pressure target drives roles that can lead to higher burnout and turnover. carried out factor analysis for mixed data and found that additionally, **HR** often experiences internal restructuring, which could explain its elevated attrition. In contrast, technical departments such as R&D or IT may offer more stable, long-term growth paths and less direct customer pressure, correlating with lower turnover.

Employees in specialized technical roles may also perceive fewer external job opportunities or may be more embedded in teams, decreasing the likelihood of leaving. Conversely, sales roles often offer externally transferable skills and a more aggressive job market, contributing to higher attrition.

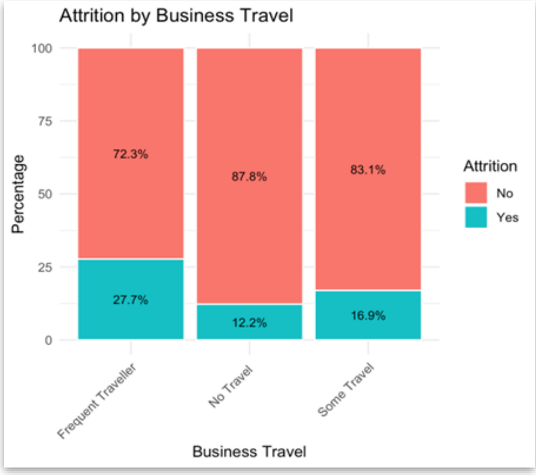

Fig 2: Attrition rate by business travel

Employes who travel frequently show notably higher attrition. Frequent business travel is disruptive to work-life balance and can strain personal relationships. Employees constantly on the road may feel disconnected from their teams, fatigued, or stressed. This cumulative stress can lead to job dissatisfaction even if other factors like pay or benefits are favorable. Interestingly, those who travel *occasionally* might strike a beneficial balance between variety and routine, potentially showing moderate attrition. This suggests that not just the **presence**, but the **intensity** of travel influences turnover.

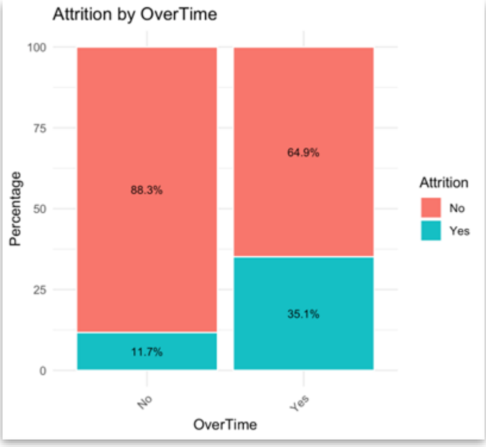

Fig 3: Attrition Rate by Over Time

Employees who work overtime are significantly more likely to leave the organization. This variable has one of the **most obvious** and **direct** correlations to attrition. Working overtime often leads to **burnout**, **chronic stress**, and a **disrupted work-life balance**. Employees with consistently long hours may not feel adequately compensated or appreciated for the extra effort, making them more likely to leave for roles with healthier boundaries.

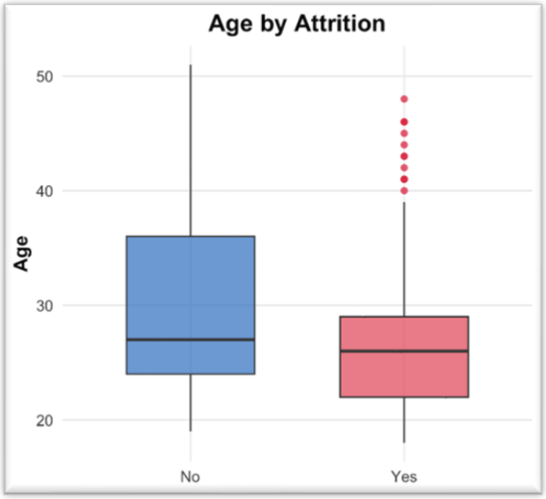

Fig 4: Attrition Rate by Age

Younger employees show significantly higher attrition compared to older employees. Age has a clear trend — younger employees tend to leave more often. This is likely driven by life stage: younger professionals are exploring opportunities, changing industries, or relocating more frequently. Older employees may value stability, benefits, or retirement plans, making them less likely to leave without strong push factors

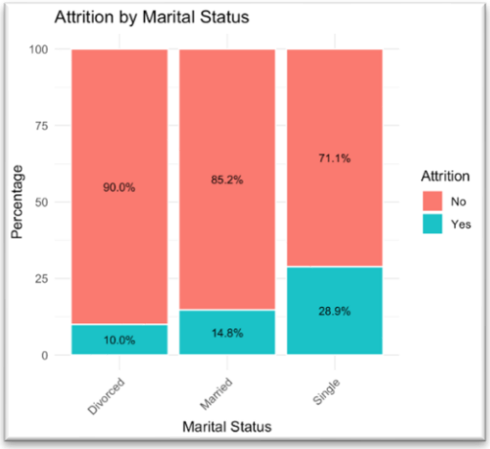

Fig 5: Attrition rate by Marital Status

Single employees are more likely to leave compared to their married counterparts. Married employees often prioritize job stability, particularly if they have dependents. Single employees, with fewer familial constraints, may be more willing to pursue risky or adventurous career changes. This dynamic is essential for understanding employee mobility and developing personalized engagement strategies.

Unsurprisingly, salary is a core driver of employee retention. Employees in the lower salary bands are more likely to seek external opportunities for better compensation, especially in high-demand industries. Interestingly, past a certain salary threshold, attrition tends to stabilize — suggesting that beyond competitive pay, other factors like growth, leadership, and recognition play a stronger role in retention. This points to a layered retention approach: for lower-salaried employees, focus on financial incentives; for higher-paid ones, emphasize career development and culture

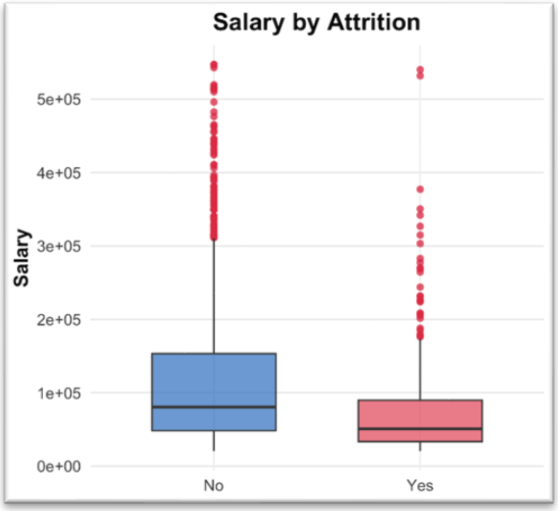

Fig 6: Attrition rate by Salary

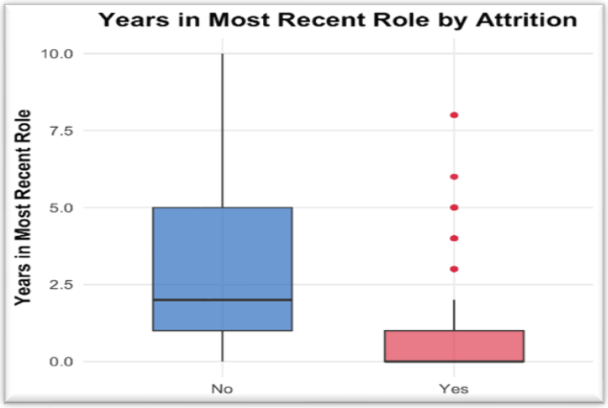

Fig 7: Attrition rate by Years in Most Recent Role by Attrition

Employees who have spent more years in the same role show increasing attrition. Staying too long in the same position without advancement can lead to career stagnation. Employees may perceive a lack of growth or recognition, prompting them to look elsewhere. Interestingly, this factor ties into job satisfaction and ambition — highlighting the need for regular performance reviews, clear promotion paths, and internal mobility options. Even loyal employees will leave if they feel they've hit a ceiling. Regular title upgrades or lateral moves can break the stagnation pattern.

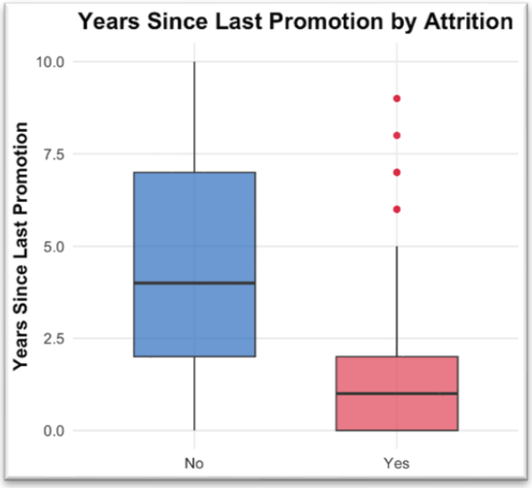

Fig 8: Attrition rate by Years since last promotion

A longer duration since the last promotion correlates with higher likelihood of attrition. Career progression is a fundamental human motivator. Employees who feel stuck or overlooked often interpret the lack of promotion as a sign that their contributions aren't recognized — leading them to explore external options.

This reinforces the importance of a transparent promotion system and development-focused culture. Even small internal changes, like title adjustments or skill expansions, can keep employees engaged and reduce turnover.

Factor Analysis for Mixed Data

The graph represents the result of a Factor Analysis of Mixed Data (FAMD), a method used to reduce dimensionality in datasets that include both numerical and categorical variables. Individuals are plotted based on their coordinates along the first two dimensions, which highlight the main sources of variation in the dataset. This visual representation helps identify patterns, similarities, and potential groupings among individuals.

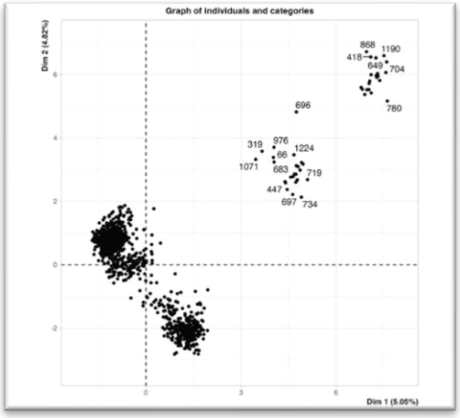

Fig 9: Factor Analysis For Mixed data

Most individuals appear closely grouped near the center of the plot, suggesting that they share similar characteristics. However, several individuals (e.g., IDs 868, 704, 780, and 1190) are located farther from these central clusters, particularly in the top-right quadrant. These individuals may be considered outliers or distinct profiles, representing unique or less common patterns in

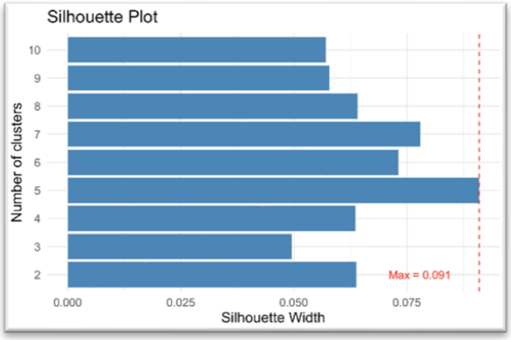

Fig 10 Silhouette Plot

the data. They may warrant further analysis to understand what differentiates them from the rest of the population.

FAMD plots like this are useful for visually detecting possible clusters. To confirm whether clear groupings exist, k-means clustering was applied using enough dimensions to explain 70% of the variance. The analysis showed no clear clustering structure, indicating that the individuals in this dataset do not naturally separate into well-defined groups.

Handling Multicollinearity

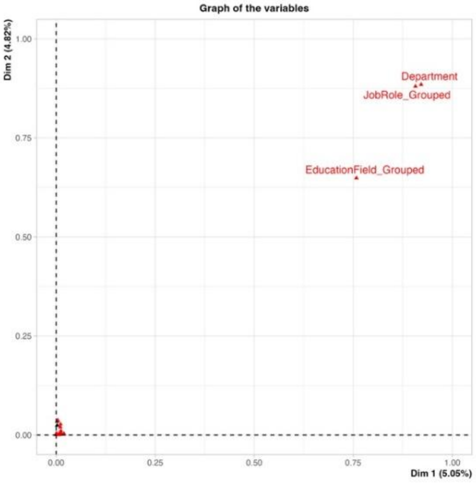

Fig 11: Variable Plot

This variable plot reveals that Department, JobRole_Grouped, and EducationField_Grouped are strongly aligned and located closely in the same direction, indicating a high degree of correlation between them. This suggests the presence of multicollinearity, where multiple variables carry overlapping or redundant information.

Because logistic regression is sensitive to multicollinearity, including these highly correlated variables can lead to unstable coefficient estimates, inflated standard errors, and difficulty in interpreting the effect of individual predictors. As a result, logistic regression may not be the most suitable modeling approach for this dataset without first addressing the multicollinearity, for example, through dimensionality reduction or variable selection. Alternative models like tree-based methods (e.g., XGBoost or Random Forest) may be better suited as they are more robust to multicollinearity.

Important results of the Advanced analysis

Model	Train Accuracy	Test Accuracy	Test Precision	Test Recall	F1 Score	AUC
Random Forest	0.957	0.878	0.723	0.553	0.627	0.8877864
XG Boost	0.880	0.875	0.693	0.575	0.628	0.8996522
Classification Tree	0.854	0.839	0.547	0.745	0.630	0.8742328
SVM	0.978	0.875	0.727	0.511	0.600	0.8734656
Logistic Regression	0.804	0.768	0.434	0.830	0.570	0.871456

Table 2: Evaluation metrics for Models

Best overall Model: XG Boost

- High Test Accuracy: 0.875 — on par with SVM and RF
- Balanced Precision and Recall: 0.693 / 0.575 → leads to an F1 of 0.628
- High AUC (89.97%) → shows strong class separation
- Less Overfitting: Training Accuracy (0.880) and Test Accuracy (0.875) are close, indicating good generalization

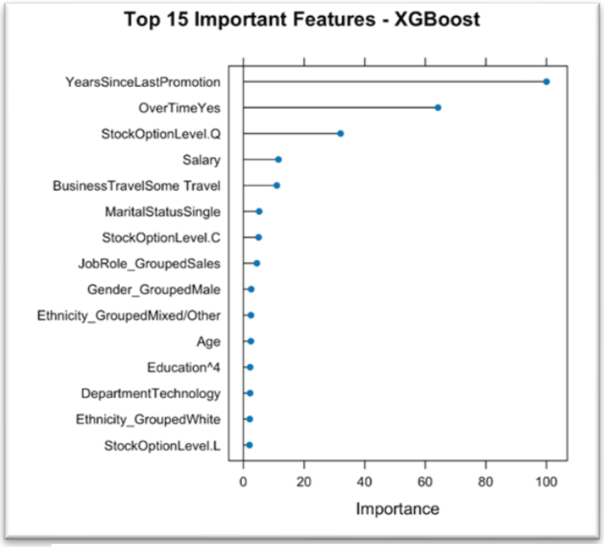

Fig 12: Feature importance plot of XGBoost

The XGBoost model highlights Years Since Last Promotion as the most important factor influencing employee attrition. This suggests that employees who haven't been promoted in a long time are more likely to leave the organization. A lack of career growth can lead to frustration and disengagement, making promotion opportunities a key area to address in retention efforts.

Another highly important factor is Overtime. Employees who regularly work overtime may experience burnout, stress, and an unhealthy work-life balance. These conditions often contribute to higher attrition rates, as overworked employees may seek better conditions elsewhere.

Stock Option Level, particularly the category "Q", also plays a significant role. This likely reflects the employee's perceived value or investment in the company. Those with lower or less attractive stock options might feel less committed or motivated to stay, while more favorable options can help improve retention. Other relevant factors include Salary, which directly impacts employee satisfaction and financial motivation. Business Travel frequency and Marital Status (e.g., being single) may also influence attrition, possibly due to how they affect personal stability and work-life balance. Together, these insights show that career advancement, fair compensation, manageable workloads, and employee recognition are critical in reducing turnover.

Discussion and Conclusion

This project demonstrates that machine learning, particularly the **XGBoost classifier**, is highly effective in predicting employee attrition. Among the five models evaluated—Logistic Regression, Classification Tree, SVM, Random Forest, and XGBoost—**XGBoost emerged as the most balanced**, with the highest AUC (89.97%), strong F1 score (0.628), and minimal overfitting, making it well-suited for real-world application.Feature importance analysis highlighted **Years Since Last Promotion**, **Overtime**, and **Stock Option Level** as key drivers of attrition. These insights align with known HR challenges, emphasizing the need for clear career advancement, workload management, and fair incentives. Other factors like **Marital Status**, **Salary**, and **Business Travel** also influence turnover risk. The XGBoost model can serve as a powerful **HR tool**, enabling early identification of at-risk employees and informing strategic interventions like career coaching and workload adjustments. Integrated into HR dashboards, it can guide data-driven retention strategies. In summary, this analysis offers actionable insights and a predictive model that organizations can use to reduce turnover and improve workforce planning. Future enhancements may include real-time data integration and feedback loops for model refinement

Appendix

- Link for the dataset: https://www.kaggle.com/datasets/mahmoudemadabdallah/hr-analytics-employee-attrition-and-performance?select=Employee.csv
- Link for the GitHub repository: https://github.com/JPasindu/Employee-Attrition-Analysis-and-Prediction