Nick Rolen

Professor Walker

SDS 570

05/04/2025

The Attrition Archetype

Employee attrition is a consistent enemy to organizational efficiency across multiple industries. This research aims to provide a comprehensive review of the factors related to attrition in the workplace today. There could be several different reasons for employee turnover. Through the analysis of publicly available HR datasets containing large amounts of anonymous employee information, many factors will be considered to create data-driven behavioral profiles of employees who are more (and less) likely to leave their organizations. These profiles will be comprised of a combination of characteristics, workplace experiences, and job conditions that are found to contribute to high or low attrition rates and will help HR professionals and organizational leaders better understand the predictors of attrition and develop strategies to improve retention and reduce turnover. Furthermore, the primary objective of this project is to examine how factors such as age, department, business travel frequency, overtime status, job satisfaction, job performance, commute length, compensation, and more correlate with employee attrition. The project aims to answer the following research questions: What are the key demographic, job-related, and organizational factors that predict employee attrition? And how do different combinations of these factors form distinct "attrition archetypes" within organizations?

The existing literature on employee attrition highlights both its cost and complexity. Al-Suraihi et al. (2021), for example, underscore the negative consequences of attrition on productivity, sustainability, and overall organizational performance, largely due to the high costs associated with replacing talent. Furthermore, Haldorai et al. (2019) introduce the concept of "push" and "pull" factors, noting that low pay, long hours, emotional labor, and poor work-life balance increase turnover, while factors like travel opportunities and community fit help retain employees.

A growing body of research focuses on the predictive modeling of attrition. Jain et al. (2020) propose models that assess individual attrition risk based on a range of variables, while Manafi Varkiani et al. (2025) explore the use of machine learning techniques to identify important features related to attrition. Other studies provide contextual insights: Sriram et al. (2019) find organizational culture and leave policies significantly influence attrition, while Taye and Getnet (2020) identify the lack of decision-making participation and career development as additional drivers. Yucel (2021) adds a leadership perspective, suggesting that transformational leaders can reduce employee departure rates by fostering a motivating work environment.

Despite these advances, there remains a gap in the literature when it comes to clustering-based behavioral profiling. Clustering-based behavioral profiling offers a distinct advantage in understanding attrition by grouping employees into meaningful, interpretable segments based on shared characteristics. Unlike predictive modeling, which focuses on estimating the probability of individual departure, clustering highlights broader structural patterns and behavioral archetypes within the workforce. This allows organizations to recognize at-risk groups without relying solely on individual-level predictions, which can be less transparent or actionable. By uncovering these clusters, HR teams can design tailored interventions for entire employee subgroups (such as early-career employees in customer-facing roles) rather than responding reactively on a case-by-case basis. Most studies rely on predictive modeling without segmenting employees into interpretable groups. This project addresses that gap by applying unsupervised learning to create data-driven archetypes of employee attrition, combining numerical and categorical insights to inform targeted HR strategies.

The dataset used in this analysis is the IBM HR Analytics Employee Attrition dataset, obtained from Kaggle. It contains 1,470 rows and 35 columns, each representing a unique employee and a range of variables related to demographics, job role, compensation, and satisfaction. Importantly, the dataset has no missing values, which facilitated a clean and uninterrupted analysis. Variables include demographic indicators (age, gender, marital status), job-specific features (job role, department, business travel frequency), and behavioral markers (job satisfaction, performance rating, overtime status). The primary target variable is a binary attrition indicator (Yes/No). This dataset is well-suited to the project goals because it captures a wide spectrum of factors known to influence employee turnover, and its structure enables both exploratory and unsupervised analysis.

The analysis began with exploratory data analysis (EDA), where numerical distributions and categorical breakdowns were visualized using boxplots, histograms, and pie charts. Early visual inspection revealed clear trends: attrition was more common among younger employees, those with fewer years at the company, lower monthly incomes, and lower job satisfaction scores. Following EDA, the data was preprocessed for clustering. All categorical variables were encoded using label encoding, and numerical features were standardized using `StandardScaler` from scikit-learn. The attrition column was removed from the dataset to ensure the clustering was unsupervised. KMeans clustering with two clusters was implemented, with the assumption that one cluster would broadly align with attrition-likely employees and the other with those more likely to stay. A Principal Component Analysis (PCA) was used to reduce the dataset to two dimensions for visualization. Once clusters were assigned, they were relabeled based on actual attrition rates within each group to reflect our two behavioral archetypes: attrition-likely and retention-likely.

The enriched dataset, now including PCA components and cluster labels, was exported to Tableau for visual storytelling. Three dashboards were created to explore different dimensions of the cluster profiles. The first was a general overview of our behavioral archetypes premise, introducing the PCA-based separation and showing attrition rates across the two clusters. The second dashboard focused on comparing numerical variables across the two clusters, highlighting differences in factors like age, tenure, job level, commute distance, and compensation. And finally, the third focused on comparing categorical variables across the two clusters, examining differences in department, job role, marital status, travel frequency, and overtime. This multi-stage approach not only allowed for the identification of meaningful employee clusters, but also enabled clear visual communication of their characteristics and attrition risk.

The cluster analysis successfully revealed two distinct behavioral profiles related to employee attrition. In Dashboard 1, the PCA projection clearly showed separation between the clusters, validating the KMeans approach. Cluster 0, identified as "attrition-likely," showed an internal attrition rate of approximately 30%, compared to just 9% in Cluster 1. Dashboard 2 highlighted key numerical differences. The attrition-likely cluster tended to be younger and earlier in their careers. They had lower job levels and shorter tenure, and monthly income was also significantly lower on average. These findings align with previous literature identifying career stage, compensation, and satisfaction as primary attrition drivers. Dashboard 3 added further depth by examining categorical attributes. Employees in the attrition-likely cluster were more likely to be single, work in roles like Sales Executive or Research Scientist, and travel frequently for business. These lifestyle and job-structure factors may reduce stability and increase burnout risk.

Together, these findings support and extend prior research. This project not only confirms well-known attrition drivers but also introduces a novel, interpretable clustering of employees based on their behavioral and structural profiles. These archetypes can help organizations move beyond individual predictors and toward pattern-based intervention strategies.

This project applied exploratory and unsupervised machine learning methods to create behavioral archetypes of employee attrition using a real-world HR dataset. Through clustering, PCA visualization, and storytelling with dashboards, the analysis revealed two distinct profiles that differ significantly in age, career satisfaction, compensation, job structure, and work-life dynamics.

By synthesizing the findings with prior literature, this report demonstrates that attrition is not only predictable but also profile-based. Rather than treating employees as isolated cases, organizations can benefit from understanding the broader patterns that lead to turnover. Future research could refine these archetypes further or apply similar methods in other industries to test generalizability. Ultimately, this project offers a practical framework for organizations to diagnose attrition risk at scale and intervene more proactively and equitably.