Final Project: Predicting Possible Causes of Employee Attrition using the IBM Human Resources Employee Analytics Data

Brian M. Jones

Belhaven University

2025FA MDS620: Making Data Decisions

Dr. Brett Andrews

October 4, 2025

Phase 1: Problem Statement

Employee turnover is a major challenge for organizations because the costs of replacing staff often outweigh the costs of retention. The IBM Human Resources dataset provides an opportunity to examine how salary and educational attainment relate to attrition, or the likelihood that an employee will leave the company. Specifically, this project investigates to answer how does average monthly income differs across levels of education and whether this variation plays a role in employee attrition.

The target variable in this analysis is Attrition, a binary outcome indicating whether an employee has left the organization. Key predictors include Monthly Income and Education, with the potential to incorporate additional factors such as job satisfaction, job tenure, or work-life balance. Understanding how these variables interact is important because salary structures often reflect educational background, and inequities in pay can contribute to job dissatisfaction and higher turnover.

The results of this study can provide actionable insights for HR managers and business leaders. By identifying patterns in income and education that predict attrition, organizations can refine compensation policies, address pay desires and implement targeted retention preventatives. In conclusion, these strategies can reduce turnover costs and promote a more stable and productive work family.

Phase 2: Data Preprocessing and Exploratory Data Analysis (EDA)

Data Cleaning and Preparation

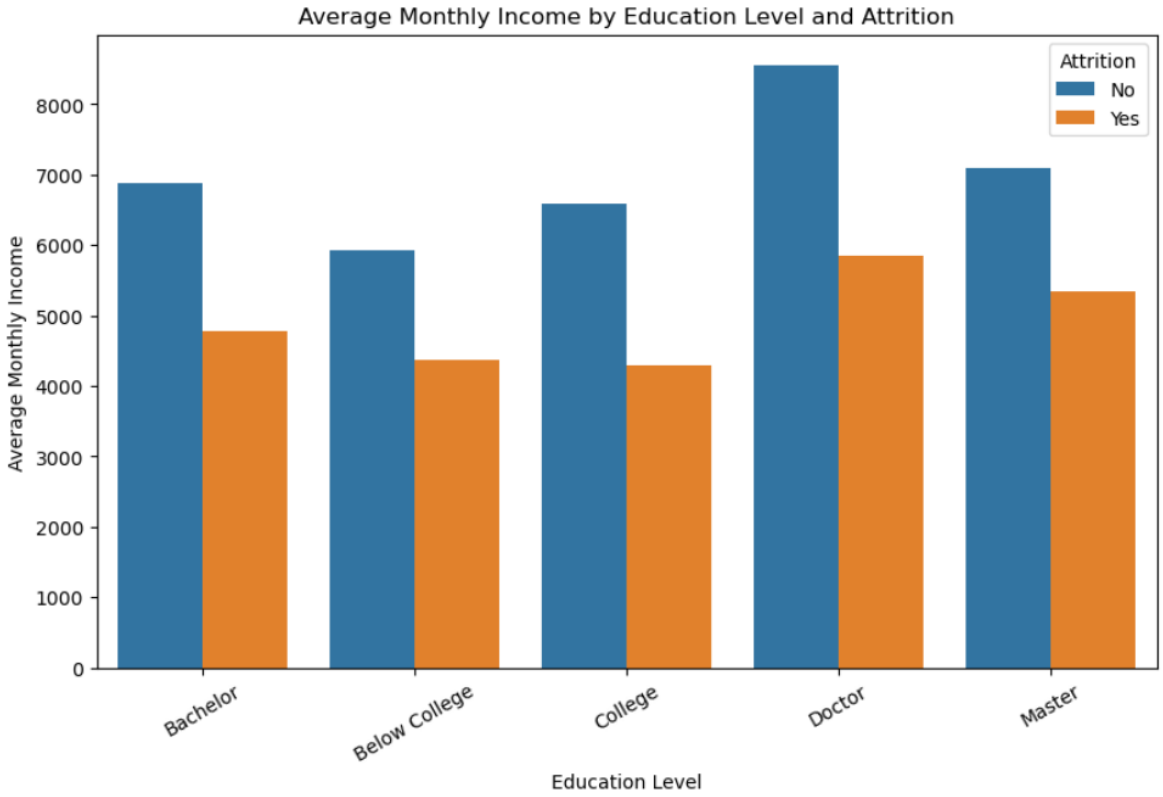
The IBM HR dataset contained no missing values, and all variables were formatted appropriately. Two categorical features required preprocessing for analysis: the Attrition variable, which was recoded into a binary datatype (Yes = 1, No = 0), and the Education variable, which was mapped to descriptive labels (Below College, College, Bachelor, Master, Doctor) for interpretability. No extreme outliers were observed in the target variables of interest, and the data was ready for exploratory analysis.

Exploratory Data Analysis

1. Monthly Income and Attrition by Education

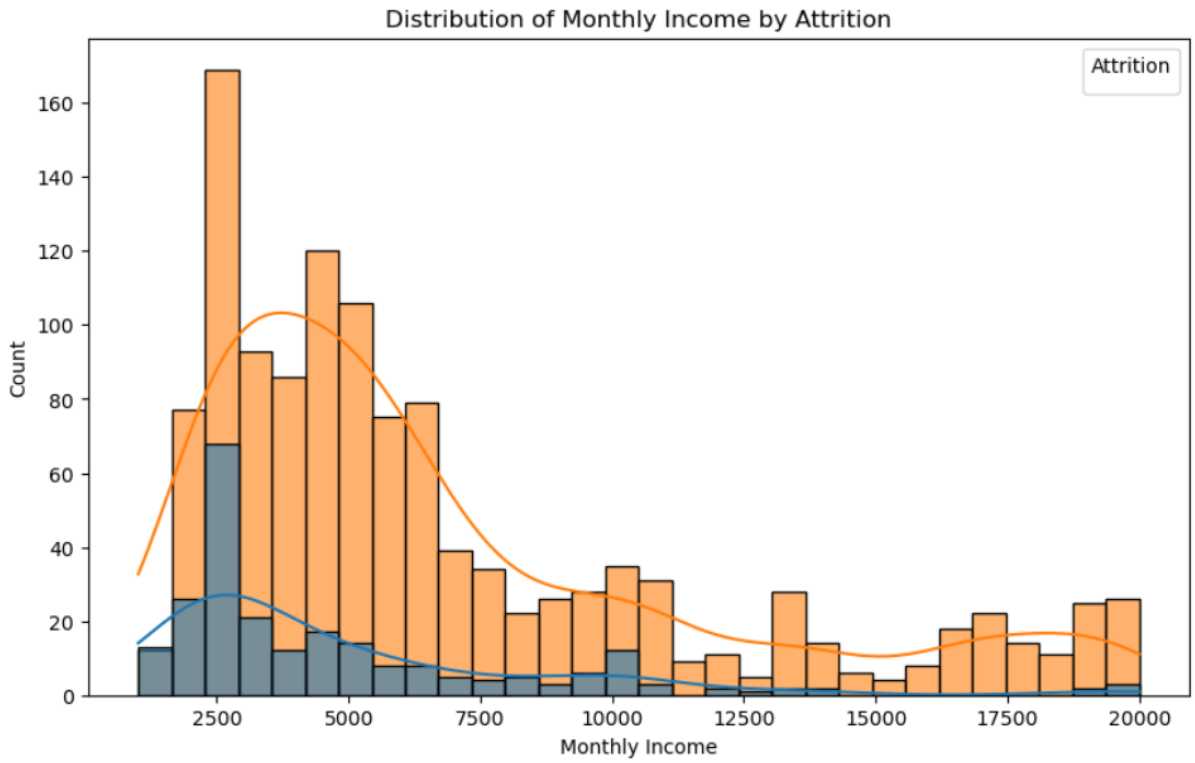
A grouped bar chart of average monthly income by education level and attrition revealed a consistent pattern. Across all five education levels, employees who did not leave the company earned higher average monthly incomes than those who left. For example, Bachelor's degree holders who stayed had an average income of approximately $6,883, compared to $4,770 for those who left. This pattern held even at the doctoral level, where those who stayed earned around $8,560 compared to $5,850 among those who left.

This finding suggests that lower compensation is strongly associated with higher attrition risk, regardless of education level.



2. Distribution of Monthly Income by Attrition

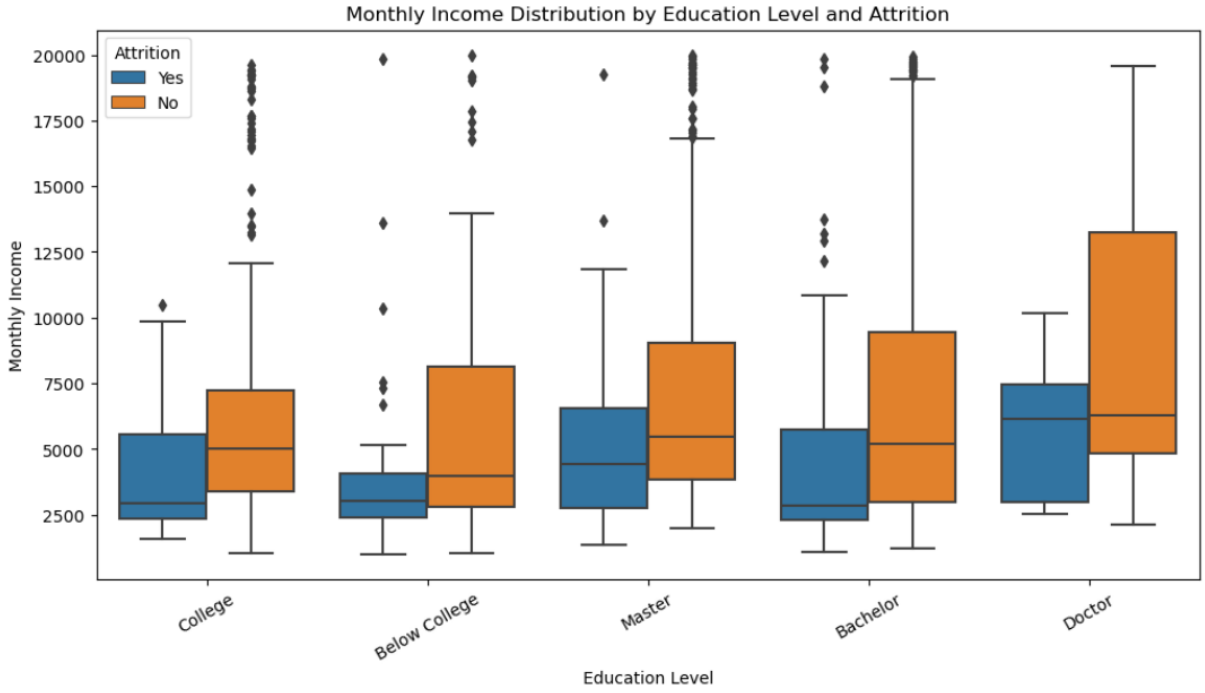
A distribution plot showed that employees who left the organization tended to cluster in the lower end of the salary distribution, while those who stayed were more evenly spread across higher income levels. This reinforces the conclusion that lower-paid employees are more likely to leave.



3. Boxplots of Monthly Income by Education and Attrition

Boxplots provided additional detail on the spread of salaries within each education group. Employees who stayed generally showed higher medians and upper ranges of income. Those who left clustered more tightly at the lower end of the salary distribution. This visual highlighted that attrition is not random but is **systematically related to lower pay within each education tier**

.



---

4. Correlation Analysis

A correlation heatmap of numerical features revealed several weak but meaningful associations with attrition. Monthly Income (-0.160), Job Level (-0.169), and Total Working Years (-0.171) all showed negative correlations with attrition, suggesting that employees who are more experienced, hold senior roles, or earn higher salaries are less likely to leave. Similarly, measures of satisfaction such as Job Satisfaction (-0.103), Environment Satisfaction (-0.103), and Job Involvement (-0.130) were also negatively associated with attrition.

In addition, Distance From Home (0.078) and Number of Companies Worked (0.043) showed small positive correlations, suggesting that employees who commute farther or have a history of job-hopping are slightly more likely to leave.

Although none of these variables displayed strong correlations on their own, the results support the idea that attrition is a multifactor outcome influenced by compensation, experience, satisfaction, and stability within the organization.

Summary of Observations from EDA

- Employees with lower monthly incomes are more likely to leave, regardless of education level.
- Satisfaction and involvement measures are associated with lower attrition risk.
- Experience and tenure also reduce the likelihood of attrition.
- No single variable strongly predicts attrition, reinforcing the need for multivariate modeling to understand turnover risk.

These findings provide a strong foundation for the modeling phase, where logistic regression and clustering will be applied to quantify and segment attrition risk.

Phase 3: Machine Learning Models (Logistic Regression and Cluster Analysis)

Logistic Regression

A logistic regression model was developed to predict employee attrition (Yes/No) using factors such as monthly income, education, job satisfaction, environment satisfaction, job involvement, overtime, distance from home, and job role. The model performed well, with a validation accuracy of 85% and an area under the ROC curve (AUC) of 0.83. These values indicate that the model was effective at distinguishing between employees who stayed and those who left.

The regression coefficients revealed meaningful insights. Employees who worked overtime were significantly more likely to leave ($\beta = 1.06$). Similarly, being single ($\beta = 0.43$), or working in roles such as laboratory technician ($\beta = 0.31$) and sales representative ($\beta = 0.26$), increased the probability of attrition. On the other hand, higher job satisfaction ($\beta = –0.18$), environment satisfaction ($\beta = –0.34$), and job involvement ($\beta = –0.27$) were associated with lower attrition risk. While monthly income was negatively related to attrition, the effect was small once other variables were included. This suggests that salary differences are important but act in combination with other factors such as satisfaction and workload.

These results tie directly to the problem statement by showing that while education and salary contribute to attrition risk, they cannot fully explain turnover on their own. Instead, attrition emerges from a combination of pay, role-related stressors, and satisfaction levels, indicating that compensation policies must be integrated with broader engagement strategies.

Cluster Analysis

To complement the predictive modeling, clustering analysis was performed using both Ward's hierarchical method and KMeans. Features included monthly income, job satisfaction, environment satisfaction, work-life balance, and distance from home. The goal was to identify natural employee groups that might shed additional light on attrition patterns linked to income and education.

Ward's method revealed three primary clusters: (1) moderately paid employees with high satisfaction but long commutes, (2) low-income employees with the lowest satisfaction but good work-life balance, and (3) high-income employees with poor work-life balance. KMeans also produced three clusters, one of which highlighted a particularly high-risk group of low-income and highly dissatisfied employees.

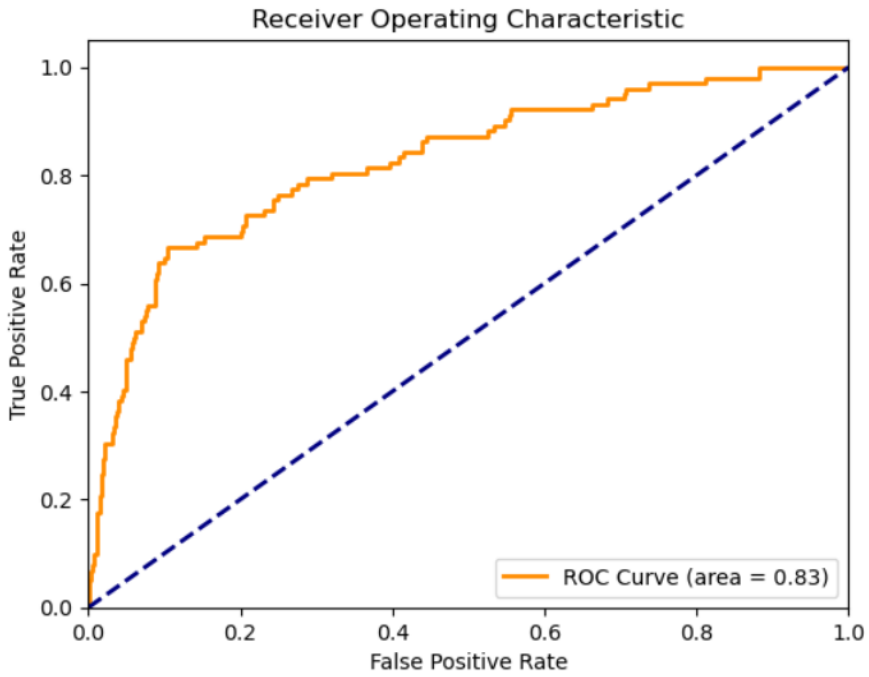
These groupings reinforce the problem statement by showing that employees in lower income brackets, often linked to lower educational attainment, are more vulnerable to attrition when dissatisfaction or poor work-life conditions are also present. Clustering provided a complementary perspective by highlighting how salary, satisfaction, and lifestyle factors interact to define at-risk employee groups.

Phase 4: Model Evaluation

Logistic Regression Evaluation

The logistic regression model achieved a validation accuracy of 85%, high specificity (98%), and an AUC of 0.83. Precision for predicting attrition was 75%, but recall was low at 24%, indicating that while the model was strong at confirming employees who would stay, it missed many who actually left. From a human resources standpoint, this makes the model most useful as a **risk-ranking tool** for identifying potential attrition cases rather than a comprehensive predictor of all turnover.

```
Precision: 0.75
Recall: 0.23529411764705882
Specificity: 0.9835390946502057
```



Cluster Analysis Evaluation

Clustering performance was evaluated using the silhouette score. Both Ward's method and KMeans produced low silhouette values, suggesting weak separation between clusters. For three clusters, Ward's achieved a score of 0.169, while KMeans slightly outperformed it with 0.191. Increasing to four clusters improved Ward's score, suggesting that a more nuanced solution may provide better segmentation. Despite modest scores, the clusters revealed meaningful groups, such as employees with low pay and low satisfaction, who are the most vulnerable to attrition.

Comparison of Models

Comparing the two models shows that logistic regression provided stronger predictive power and clearer identification of individual drivers of attrition, while clustering offered exploratory insight into how employees group naturally. Both approaches support the original problem statement: salary and education influence attrition, but they must be understood in combination with satisfaction, workload, and engagement factors.

Summary of Evaluation

- Logistic Regression: Reliable model performance with strong specificity and interpretability of predictors. Monthly income had a limited effect when modeled alone but contributed indirectly through job level and satisfaction.
- Clustering: Modest silhouette scores indicated weak separation, but the analysis still highlighted vulnerable groups, especially employees with low income and low satisfaction.
- Integration with Problem Statement: Both models confirmed that income and education-related variables are important but insufficient by themselves. Instead, attrition is best understood when pay and education are considered alongside engagement, satisfaction, and workload factors.

Phase 5: Deployment and Insights

The final step in the project was to translate the findings from the models into actionable insights for organizational decision-making. Logistic regression provided a predictive framework for estimating the likelihood of attrition, while clustering offered a descriptive segmentation of employee groups. Together, the models highlighted both individual-level drivers of attrition and group-level risk profiles.

From a deployment perspective, a logistic regression model could be implemented in a human resources dashboard to flag employees with elevated attrition probabilities. Although the model demonstrated strong specificity and precision, its recall was limited. This means that while not every potential leaver would be identified, the model would reliably highlight those employees at the highest risk. Human resources could then target interventions such as retention bonuses, career development opportunities, or workload adjustments for those employees most likely to leave.

Clustering analysis can also be deployed in practice by segmenting employees into groups with shared characteristics. For example, the KMeans analysis revealed a high-risk cluster of employees who reported both low income and very low job satisfaction. Ward's method highlighted groups with long commutes or poor work-life balance. These clusters can help HR managers design tailored interventions: salary adjustments for low-income groups, satisfaction improvement initiatives for disengaged employees, and flexible work policies for those facing long commutes.

The combination of predictive modeling and segmentation creates a holistic strategy. Logistic regression identifies who is most at risk, while clustering provides context for why certain

groups are vulnerable. Together, these insights support targeted, data-driven HR policies designed to reduce attrition and improve workforce stability.

Conclusion

The purpose of this project was to evaluate whether education level and monthly income explain employee attrition, using the IBM HR dataset. Exploratory analysis demonstrated that employees with higher education generally earned higher average monthly incomes, which confirms that education contributes to compensation differences. However, the findings also revealed that income and education alone are not sufficient to predict attrition.

The logistic regression model showed that attrition is influenced by multiple interacting factors, including job satisfaction, environment satisfaction, job involvement, number of companies previously worked, job level, and distance from home. While higher income reduces attrition risk, the effect was modest once other factors were included. This highlights that turnover is not solely a function of compensation or education, but rather a multifaceted outcome shaped by satisfaction, workload, and stability.

The cluster analyses reinforced these conclusions. Both Ward's hierarchical method and KMeans clustering revealed groups of employees where low income coincided with low satisfaction, a combination strongly associated with higher attrition. These findings suggest that employees in lower salary brackets are more vulnerable, not only because of pay but also because of dissatisfaction and limited engagement.

Taken together, the results address the original problem statement by showing that education influences income, and income is correlated with attrition risk, but these variables cannot be viewed in isolation. Retention strategies must account for multiple factors simultaneously. Human resources leaders should therefore focus on integrated approaches that balance fair compensation with improved satisfaction, engagement, and workload management. By combining predictive models with employee segmentation, organizations can proactively identify high-risk groups and implement tailored interventions, reducing turnover and fostering a more stable workforce.