Analysis of Employee Resignation

Shiyi Yang

12/07/2024

Introduction

Employee resignation is a critical challenge for organizations, often leading to disruptions in operations, increased recruitment costs, and reduced morale. Understanding the factors that drive resignation, predicting when employees are likely to leave, and identifying patterns within employee groups can empower organizations to take proactive measures to improve retention and foster a productive and supportive work environment.

This project uses the Employee Performance and Productivity dataset offered on Kaggle, which includes detailed metrics such as employee demographics, performance evaluations, satisfaction scores, and work experience. This dataset contains 100,000 employees and related 20 features (Employee ID, Department, Gender, Age, Job Title, Hire Date, Years at Company, Education Level, Performance Score, Monthly Salary, Work Hours per Week, Projects Handled, Overtime Hours, Sick Days, Remote Work Frequency, Team Size, Training Hours, Promotions, Employee Satisfaction Score, Resigned). For more detailed information about these features, please refer to the data page.

The study aims to explore the dynamics of employee resignation.

Outcome of Interest: - Resign (0: the employee hasn't resigned; 1: the employee has resigned); - Years At Company (the number of years the employee has been working for the company)

Three key questions:

1. What feature significantly affects employee resignation?

   Approach: Logistic Regression to identify which factors (e.g., satisfaction score, hours worked, etc.) significantly impact resignation.

2. When will an employee resign, and what factors speed up resignation?

   Approach: Apply Cox Proportional Hazards model to predict time-to-resignation and key predictors that accelerate the process.

3. Are employee groups distinguishable based on resignation?

   Approach: Apply UMAP to reduce dimensionality and cluster employees based on resignation, exploring patterns in these groups.

Exploratory Analysis

The dataset is highly imbalanced, with 10,010 employees who resigned and 89,990 who did not. To address this imbalance, I sampled 10,010 employees who did not resign and combined them with the 10,010 who did, creating a balanced subset for analysis.

Below is a descriptive summary of each feature by resignation status.

	Not Resigned	Resigned	Overall
	(N=10010)	(N=10010)	(N=20020)
Department			
Customer Support	1092 (10.9%)	1098 (11.0%)	2190 (10.9%)
Engineering	1121 (11.2%)	1057 (10.6%)	2178 (10.9%)
Finance	1096 (10.9%)	1180 (11.8%)	2276 (11.4%)
HR	1115 (11.1%)	1125 (11.2%)	2240 (11.2%)
IT	1097 (11.0%)	1064 (10.6%)	2161 (10.8%)
Legal	1145 (11.4%)	1136 (11.3%)	2281 (11.4%)
Marketing	1121 (11.2%)	1125 (11.2%)	2246 (11.2%)
Operations	1074 (10.7%)	1121 (11.2%)	2195 (11.0%)
Sales	1149 (11.5%)	1104 (11.0%)	2253 (11.3%)
Gender			
Female	4758 (47.5%)	4816 (48.1%)	9574 (47.8%)
Male	4842 (48.4%)	4807 (48.0%)	9649 (48.2%)
Other	410 (4.1%)	387 (3.9%)	797 (4.0%)
Age			
Mean (SD)	40.9 (11.2)	41.1 (11.2)	41.0 (11.2)
Median [Min, Max]	41.0 [22.0, 60.0]	41.0 [22.0, 60.0]	41.0 [22.0, 60.0]
Job Title			
Analyst	1437 (14.4%)	1450 (14.5%)	2887 (14.4%)
Consultant	1430 (14.3%)	1430 (14.3%)	2860 (14.3%)
Developer	1379 (13.8%)	1411 (14.1%)	2790 (13.9%)
Engineer	1379 (13.8%)	1393 (13.9%)	2772 (13.8%)
Manager	1445 (14.4%)	1470 (14.7%)	2915 (14.6%)
Specialist	1492 (14.9%)	1427 (14.3%)	2919 (14.6%)
Technician	1448 (14.5%)	1429 (14.3%)	2877 (14.4%)
Years at Company			
Mean (SD)	4.50 (2.88)	4.48 (2.87)	4.49 (2.88)
Median [Min, Max]	5.00 [0, 10.0]	4.00 [0, 10.0]	4.00 [0, 10.0]
Education Level			
Bachelor	4996 (49.9%)	5003 (50.0%)	9999 (49.9%)
High School	2969 (29.7%)	2999 (30.0%)	5968 (29.8%)
Master	1503 (15.0%)	1487 (14.9%)	2990 (14.9%)
PhD	542 (5.4%)	521 (5.2%)	1063 (5.3%)
Employee's Performance Rating.			
1	2000 (20.0%)	2034 (20.3%)	4034 (20.1%)
2	2068 (20.7%)	1992 (19.9%)	4060 (20.3%)
3	1988 (19.9%)	2019 (20.2%)	4007 (20.0%)
4	1975 (19.7%)	2031 (20.3%)	4006 (20.0%)
5	1979 (19.8%)	1934 (19.3%)	3913 (19.5%)
Monthly Salary			
Mean (SD)	6390 (1380)	6400 (1370)	6390 (1370)
Median [Min, Max]	6500 [3850, 9000]	6500 [3850, 9000]	6500 [3850, 9000]

	Not Resigned	Resigned	Overall
	(N=10010)	(N=10010)	(N=20020)
Work Hours per Week			
Mean (SD)	45.0 (8.96)	45.0 (8.98)	45.0 (8.97)
Median [Min, Max]	45.0 [30.0, 60.0]	45.0 [30.0, 60.0]	45.0 [30.0, 60.0]
Projects Handled			
Mean (SD)	24.7 (14.4)	24.4 (14.4)	24.5 (14.4)
Median [Min, Max]	25.0 [0, 49.0]	24.0 [0, 49.0]	25.0 [0, 49.0]
Total Overtime Hours Worked in the Last Year.			
Mean (SD)	14.5 (8.63)	14.6 (8.66)	14.5 (8.65)
Median [Min, Max]	14.0 [0, 29.0]	15.0 [0, 29.0]	15.0 [0, 29.0]
Sick Days			
Mean (SD)	6.99 (4.32)	7.03 (4.36)	7.01 (4.34)
Median [Min, Max]	7.00 [0, 14.0]	7.00 [0, 14.0]	7.00 [0, 14.0]
Percentage of Time Worked Remotely.			
0	2014 (20.1%)	1881 (18.8%)	3895 (19.5%)
25	2034 (20.3%)	2046 (20.4%)	4080 (20.4%)
50	1993 (19.9%)	1971 (19.7%)	3964 (19.8%)
75	2015 (20.1%)	2074 (20.7%)	4089 (20.4%)
100	1954 (19.5%)	2038 (20.4%)	3992 (19.9%)
Team Size			
Mean (SD)	10.0 (5.48)	10.0 (5.48)	10.0 (5.48)
Median [Min, Max]	10.0 [1.00, 19.0]	10.0 [1.00, 19.0]	10.0 [1.00, 19.0]
Training Hours			
Mean (SD)	49.4 (28.8)	49.6 (28.8)	49.5 (28.8)
Median [Min, Max]	49.0 [0, 99.0]	50.0 [0, 99.0]	50.0 [0, 99.0]
Promotions			
0	3344 (33.4%)	3344 (33.4%)	6688 (33.4%)
1	3279 (32.8%)	3343 (33.4%)	6622 (33.1%)
2	3387 (33.8%)	3323 (33.2%)	6710 (33.5%)
Employee Satisfaction Rating.			
Mean (SD)	3.00 (1.16)	2.99 (1.15)	2.99 (1.15)
Median [Min, Max]	2.99 [1.00, 5.00]	2.99 [1.00, 5.00]	2.99 [1.00, 5.00]

Logistic Regression

To identify the factors influencing employee resignation, a logistic regression model was employed to assess the effects of Department, Gender, Age, Job Title, Years at Company, Education Level, Performance Score, Monthly Salary, Work Hours per Week, Number of Projects Handled, Overtime Hours, Sick Days, Remote Work Frequency, Team Size, Training Hours, Promotions, and Employee Satisfaction Score. Employee ID and Hire Date were excluded from the analysis.

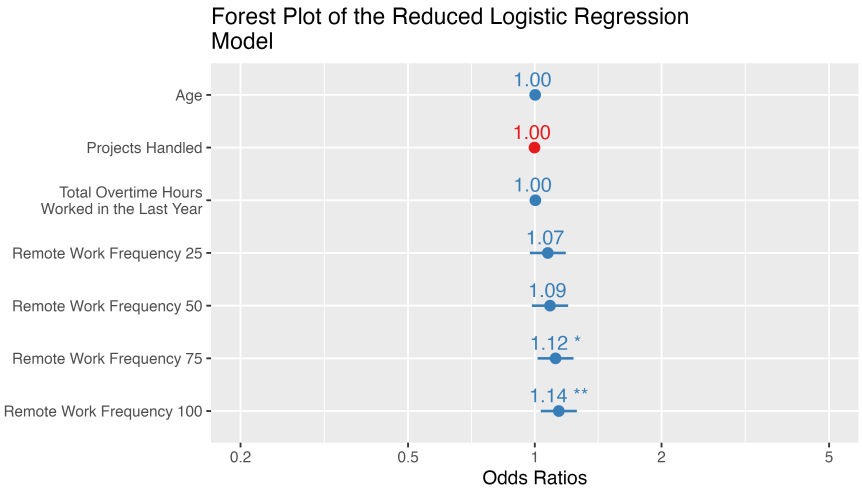
A final, reduced predictive model was derived using backward variable selection based on the Akaike Information Criterion (AIC) starting from the full model. Results are presented as Odds Ratios.

To evaluate model fit and predictive performance, the dataset was partitioned into training and testing sets at an 80:20 ratio. The model was trained on the training set and subsequently evaluated on the testing set.

Fitting Results

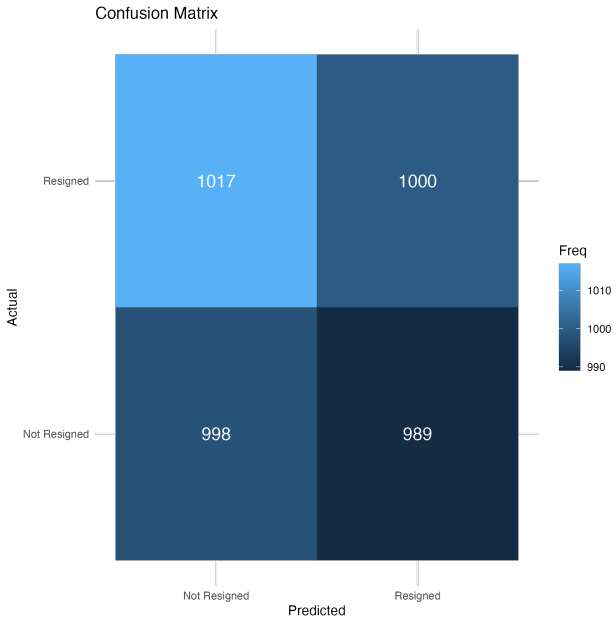
In the final model, Age, Projects Handled, Overtime Hours, and Remote Work Frequency were retained as predictors. The forest plot below displays the estimated effects and corresponding 95% confidence intervals

for these variables, based on the reduced logistic regression model fitted to the training set. Among these factors, only Remote Work Frequency demonstrated a statistically significant association with employee resignation.



Prediction

Below is the confusion matrix for the predictions on the testing set. The model's performance is extremely poor. Both the recall for the positive class (resigned) and the overall accuracy are around 50%, which is essentially equivalent to random guessing.

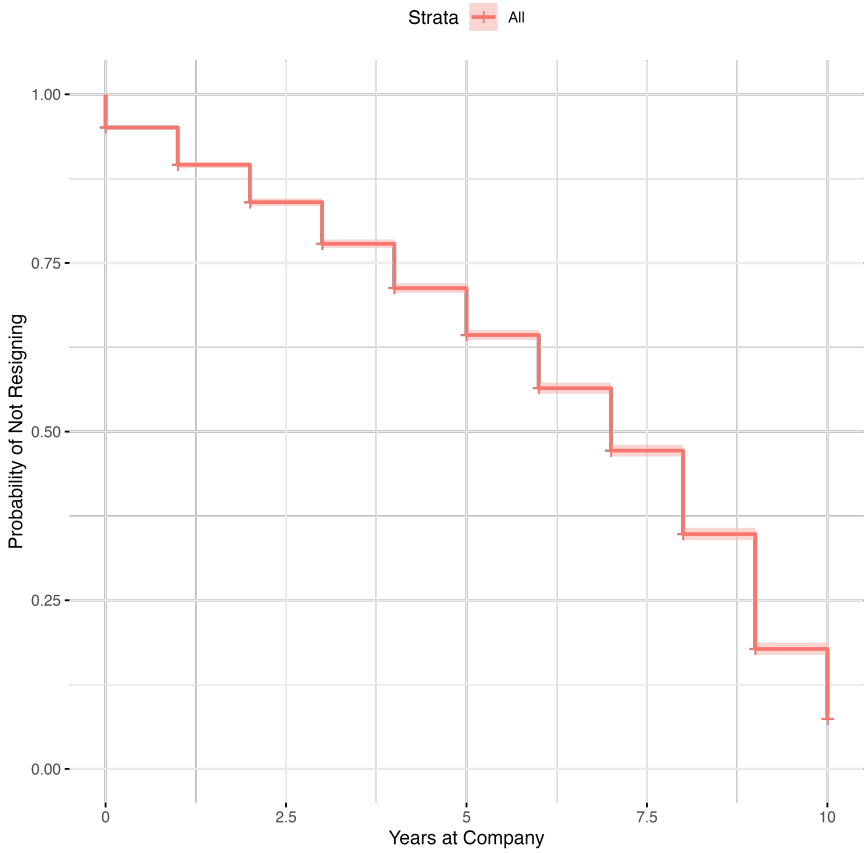


Survival Analysis

Given that the dataset includes each employee's tenure (in years) at the company, employee resignation can be modeled as a time-to-event outcome, with non-resigned employees treated as right-censored observations.

Kaplan-Meier Surival Plot

Displayed below are the Kaplan–Meier survival curves. Based on these estimates, the median time until resignation is 7 years, with a corresponding 95% confidence interval of 7.00–7.00

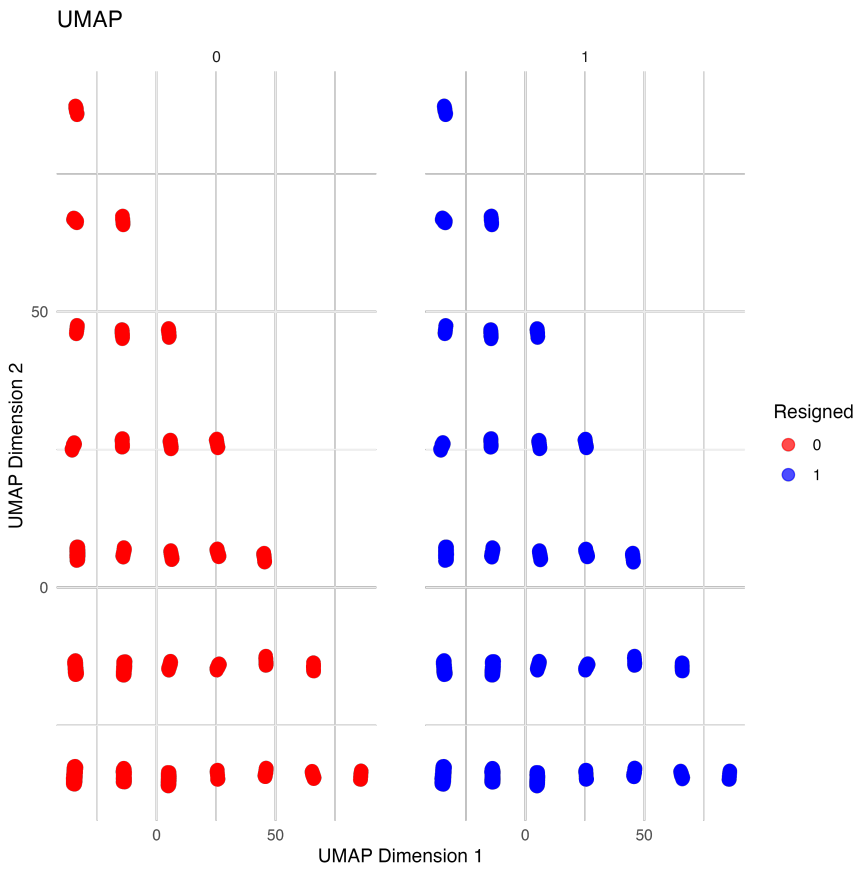


Cox Proportional Hazards Model Results

A procedure similar to that used in the logistic regression analysis was employed for the Cox proportional hazards model. After variable selection, only "Projects Handled" remained in the final model. The estimated hazard ratio for this variable is 0.9988, and its 95% confidence interval includes 1, suggesting that its effect on the hazard rate is not statistically significant.

Dimension Reduction

UMAP was employed to reduce the dataset to a lower-dimensional representation. Prior to this, all categorical variables were converted into dummy variables, resulting in a total of 40 columns. The visualization below shows the UMAP projection. Because the points representing "resigned" and "not resigned" employees overlap substantially in the original plot, separate visualizations for each group are presented to facilitate clearer interpretation.



Summary

The dataset, as provided, does not appear to contain strong or easily identifiable predictors of employee resignation. Although a few factors were initially considered, their effects proved to be statistically weak and offered negligible improvement over random guessing. The absence of meaningful patterns persisted even when approaching the problem from a time-to-event perspective and exploring the data through dimension reduction techniques. Consequently, the findings suggest that either the underlying factors driving employee turnover are not captured by the available features, or more sophisticated modeling techniques and additional data are needed to uncover informative patterns.