Liam Kennedy
Professor Srinivasan
Statistical Modeling
December 11th, 2025

Predicting Attrition Report

Employee turnover or attrition can be a major challenge for organizations because replacing workers is costly, time consuming, and disruptive to productivity. Understanding which factors predict attrition can allow companies and managers to design more effective retention strategies, allocate resources more effectively, and improve employee satisfaction. The goal of this analysis was to build statistical models that could predict if an employee is likely to leave and identify which features play the largest role in these decisions. I used logistic regression and random forest classifiers to model attrition to best find the "story" behind employee churn. With so many initial variables, I divided the data into separate groupings reflecting time/tenure, compensation, life/environment, and the job field/role. This allowed me to evaluate whether certain categories of features show any predictive value and how they contributed to attrition.

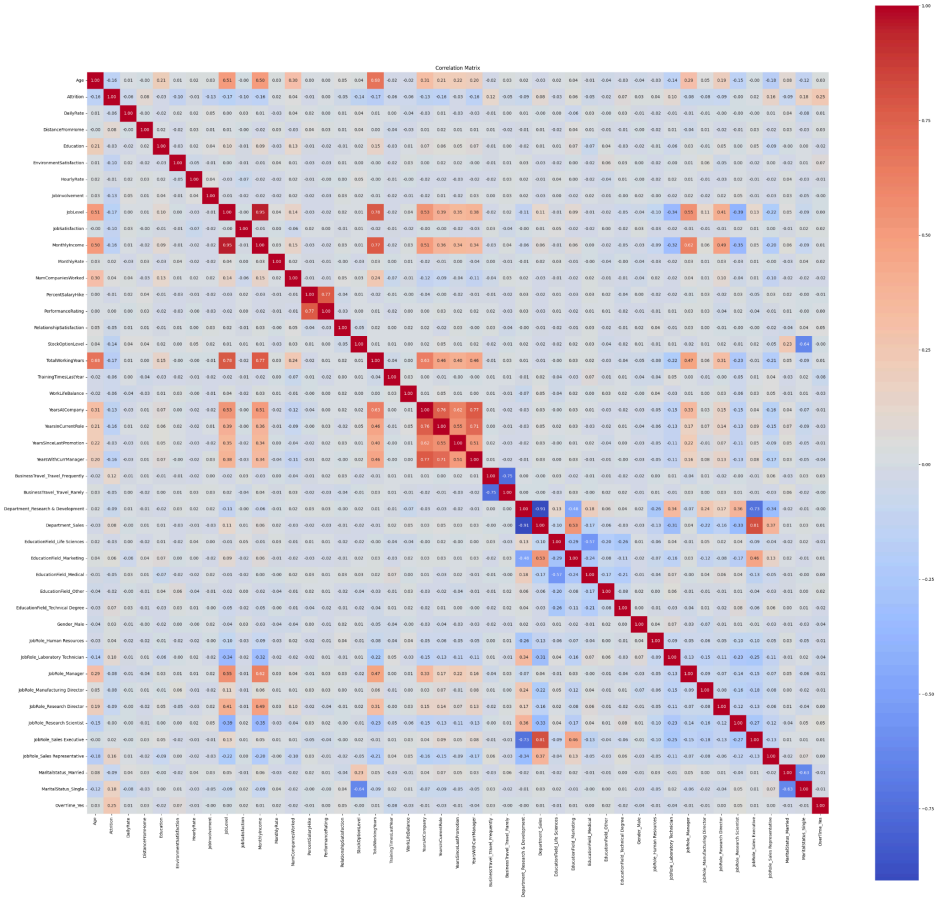
The key questions I aimed to answer:

- Which characteristics were most influential in predicting churn?
- Would separately grouping variables lead to successful predictive results?
- How accurately could workers be classified as likely to stay or quit?
- What actionable insights can organizations make from these findings to reduce attrition?

If successful, this type of analysis can provide leadership with actionable guidance about workplace conditions, priorities, and compensation policies to reduce turnover.

The Data Set

Overall, this dataset consisted of 1,470 employees, each described by a wide range of demographic, workplace, and performance variables. These include Time/Experience, Compensation, Life/Environment, and Job Role/Education. The target variable for this dataset was Attrition_Yes.



Preprocessing Steps

- Encoded all categorical variables to prevent multicollinearity.
- Removed irrelevant constants like employeecount, over18, standardhours, employeenumber which carried no predictive information.
- Scaling was applied to logistic regression models.
- Train/test split was 70% training and 30% testing and was stratified by attrition as well.
- No major concerns with data quality as well as no missing values.

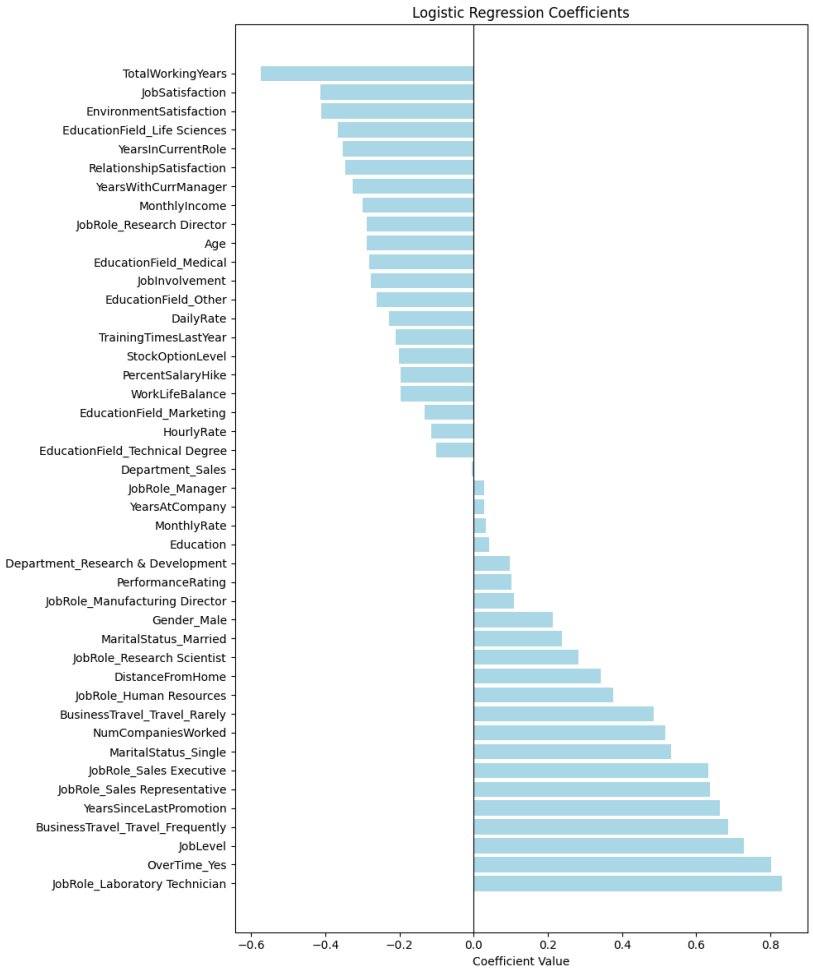
Model	# Accuracy	# Precision	# Recall	# ROC-AUC
Set 1: Time/Experience	0.61	0.23	0.63	0.67
Set 2: Compensation/Reward	0.68	0.29	0.72	0.75
Set 3: Life/Environment	0.66	0.26	0.62	0.69
Set 4: Job Role/Education	0.59	0.22	0.61	0.65
Full Logistic Regression	0.75	0.36	0.68	0.82
Random Forest	0.82	0.38	0.23	0.77

Results

Across all experiments, model performance varied depending on the predictor set used. The single set logistic regression models showed moderate accuracy but still struggled with precision and recall, while the full feature Logistic regression and Random Forest models performed notably better.

When looking specifically into the single set models, the Compensation set produced the strongest recall and the highest ROC-AUC, while the Experience set also showed high levels of recall despite the low precision. This suggests that these two variable sets did carry some signaling for identifying employees likely to quit, but each category alone is not sufficient for thorough prediction.

The full logistic regression model performed substantially better. Here, it achieved 0.75 accuracy, 0.36 precision, 0.68 recall, and a strong 0.82 ROC-AUC. This meant that the model captured almost 70% of true attrition cases and showed improvement in discrimination when all predictors were included together. Further inspecting the coefficients, features like Overtime or YearsSinceLastPromotion increased likelihood of quitting, whereas TotalWorkingYears, JobSatisfaction, and EnvironmentSatisfaction decreased this chance.



The Random Forest Model achieved the highest accuracy with a strong ROC-AUC value, but recall remained relatively low at 0.23. This meant that the RF model was very effective at correctly predicting employees who stayed but more conservative when predicting attrition. Feature importance and permutation highlighted consistent predictors like MonthlyIncome, TotalWorkingYears, Age, and YearsAtCompany all of which suggested that these factors play a major role in long term retention.

Discussion

This analysis showed that no single category of variables was sufficient to model attrition well. Each set captures one dimension of attrition behavior but only the full model is able to

effectively combine all predictors. My full logistic regression model provided clear interpretability and showed which direction variables influenced attrition. On the other hand, Random Forest captured the relative importance of features, showing which factors mattered most. Across all models, the results consistently suggest that employees are more likely to quit when they experience stagnation, instability, or weak working relationships. These insights can guide decision making by highlighting when and where interventions are important and how these issues can be solved.