Employee Turnover:

Understanding the drivers, Predicting at-risk employees, Acting to mitigate

Peter G. King

21 Apr 2025

Acme Aroma

Workforce Planning Project

Table of Contents

1 Problem

2 Approach

3-5 Insights

6 Recommendations

7 Limitations

The Problem: Employee Turnover is Increasing...

Problem summary

Acme Aroma is suffering an increasing trend in both turnover rate (currently 16%) and absolute employee departures (~634 in 2022). Inexperienced employees are jeopardizing production and sales. The costs of turnover (hiring and retraining) are also Increasing, amplifying the cost of each lost employee.

The company wants to understand what is causing the increase in turnover so that it can implement mitigating actions.

Business challenges

Production less efficient and sales jeopardized by inexperienced employees

Acquisition costs now significant, from 15K rupees per new hire in 2021 to 30K rupees in 2022

Job satisfaction down last 5 years, from 3.4 to 2.7

Applications per job opening down from 27 in 2017 to only 8 in 2022

Hiring costs: Doubled last year

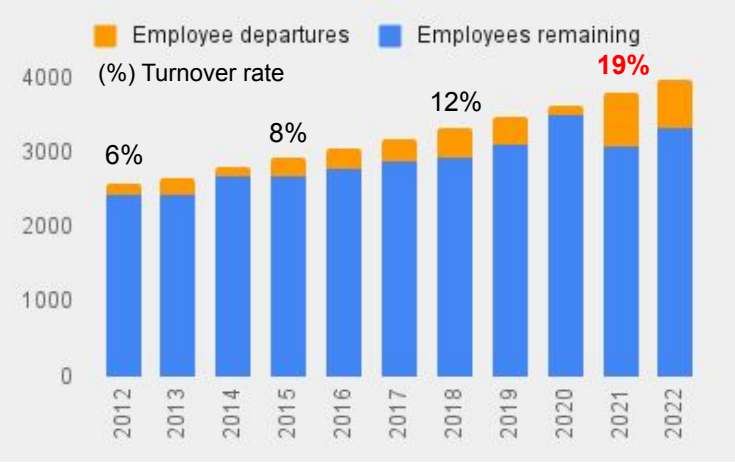
Hiring and training costs for each new employee are increasing.

- **15,000 rupees** per new hire in 2021
- **30,000 rupees** per new hire in 2022

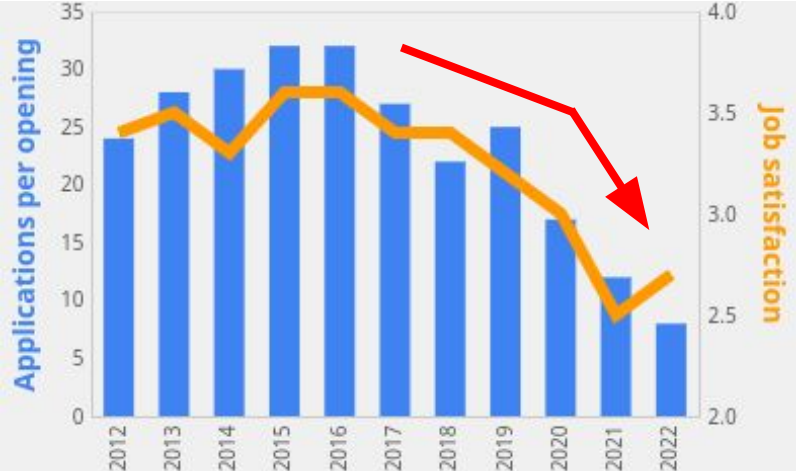
Total cost to replace ~634 employees in 2022:

19M rupees

Turnover: Tripled last 10 yrs



Worrisome downward trends



Project goals

Learn why employees are leaving.

Gain **predictive insights** into employee turnover to **improve workforce planning**.

If possible, predict which employees are at high risk for turnover.

Recommend course(s) of action to **lower employee turnover**, **reduce associated costs**, and **stabilize performance**, particularly in production and sales.

Our Approach: Data and Modeling Considerations

What information is available?

We will use internal employee data from Pegasus to **understand the factors driving attrition**: what makes an employee decide to leave the company?

Previous conjoint analysis identified several key variables that management can influence:

- Monthly Income
- Training TIme
- Environment Satisfaction
- Work-Life Balance
- Job Satisfaction

key variables

We also want to **predict which employees are at risk of leaving** the company. Our model will use these additional factors to aid in prediction:

- Business Travel
- Distance From Home
- Job Involvement
- Job Level
- Performance Rating
- Percent Salary Hike
- Years at the Company
- Years since Last Promotion

employee-specific risk factors

What factors are important?

Because an employee's decision to leave or stay is a binary choice, **logistic regression** provides an excellent way to model how underlying decision factors affect an employee's decision. It is the most mathematically appropriate model we can use to both **understand** and **predict** employee turnover.

With logistic regression, we estimate the relative effect of each factor on an employee's decision, while holding other factors constant, in terms of an **odds ratio**: leave (L) vs. stay (S), or L:S. Our primary goal is to identify a key variable that management can influence that will have the greatest reduction in the odds of an employee leaving.

We can use this same mathematical model to predict which employees are at higher risk of attrition by including employee-specific risk factors in our analysis.

Choosing a good model

Our model will estimate how selected variables influence attrition. We consider attrition "positive" when an employee leaves the company, and "negative" if the employee stays.

No mathematical model of a complex, real-world process will yield perfect results. Part of model selection involves deciding where to balance the **tradeoff** between **Type I** and **Type II errors**. A model tuned to be highly sensitive to detecting positive cases will typically produce more Type I errors (false positives). By contrast, a model tuned for high **precision** will typically produce more Type II errors (false negatives).

Because part of our goal is to identify at-risk employees, it is more acceptable for the model to incorrectly label an employee as leaving (a false positive, or Type I error) than for it to incorrectly label an employee as staying (a false negative, or Type II error). We want to minimize false negatives, so we will focus on **recall** (or sensitivity) as our primary model evaluation metric.

Focusing on recall will minimize the number of times the model fails to correctly predict an employee's decision to leave. The downside is that we may incorrectly predict that an employee will leave when they will actually stay, but this kind of error is less problematic for our workforce planning effort.

Exploratory Insights

Data exploration yielded a few surprising discoveries

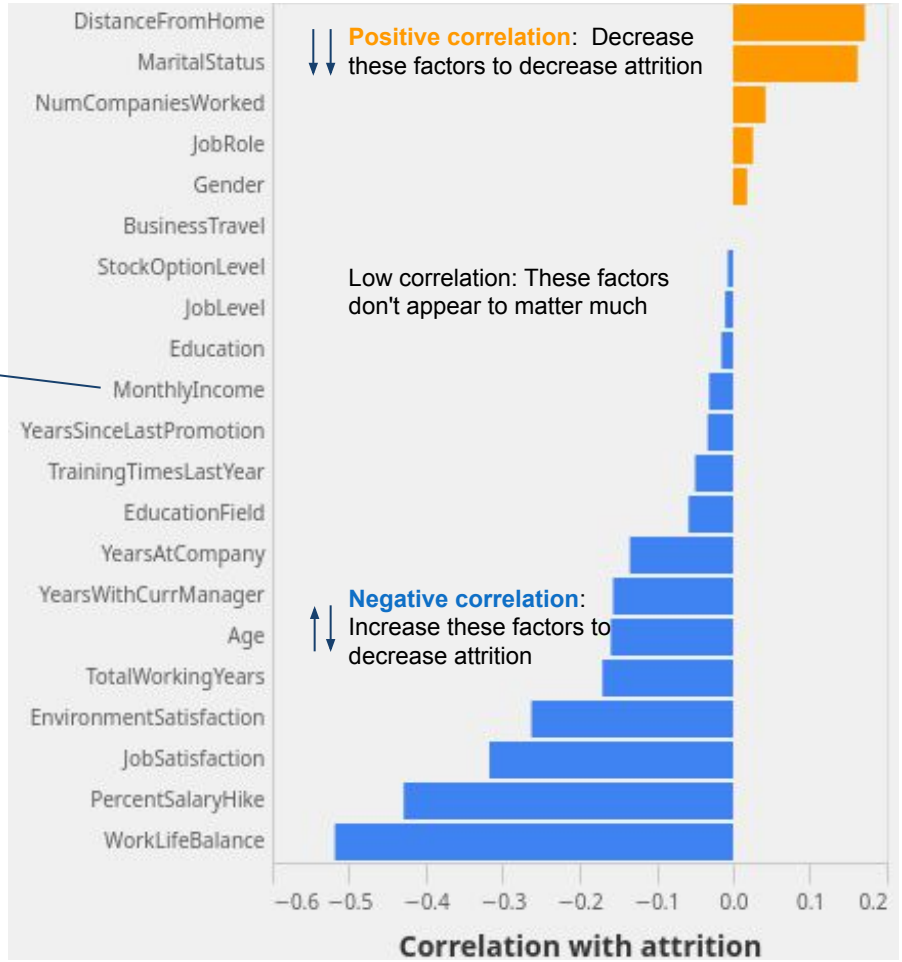
Data from Pegasus was in good condition with only a few instances of missing variables. We have high confidence in the fidelity of the data behind this study.

Prior to modeling, we looked at the pairwise Pearson correlation between attrition and potential explanatory variables. Pearson correlation can range from -1 (perfect negative correlation) to +1 (perfect positive correlation). Variables with a correlation coefficient of zero are essentially unrelated. We were surprised to find that **monthly income shows practically zero correlation with attrition**. When it comes to a choice of whether or not to leave the company, it doesn't seem to matter how much or little you pay an employee.

Demographic variables like age and gender showed less significant correlation with attrition (age more so than gender), and attrition rates appeared balanced among various categories.

We found several factors highly correlated with attrition

- **Work-life balance**
- **Salary hike last year (percentage of salary)**
- **Job satisfaction**
- **Environment satisfaction**



Model Selection & Performance

A description of our logistic regression model

Our logistic regression model estimated the relative importance of selected variables in explaining the observed pattern of employee attrition decisions. Mathematically, it asked a question like, "how important is *salary* compared to *job satisfaction* in explaining why an employee decides to leave or stay?" Regression coefficients reflect this relative importance. Although the exact mathematical relationship is more complicated than a simple ratio, we can still use the coefficients to rank various factors.

Our final model included these variables*, listed in order of their relative importance by **odds ratio**:

• Work-Life Balance (sd)	1 : 6.4	scale of 1-4 (4 = best, 1 = worst)
• Performance Rating (sd)	1 : 4.0	scale of 1-4 (4 = outstanding, 1 = low)
• Job Involvement (sd)	1 : 3.9	scale of 1-4 (4 = very high, 1 = low)
• Job Satisfaction (sd)	1 : 2.6	scale of 1-4 (4 = very high, 1 = low)
• Percent Salary Hike (sd)	1 : 2.6	quantitative: percent salary increase last year
• Environment Satisfaction (sd)	1 : 1.8	scale of 1-4 (4 = very high, 1 = low)

**(sd) indicates that the variable was standardized before being introduced to the model*

Selecting the best model based on performance

We wanted to minimize the number of times the model would fail to predict an employee's decision to leave (a false negative), so we chose recall (or sensitivity) as our primary selection metric.

Accuracy: 0.93 **Precision: 0.75** **Recall: 0.89**

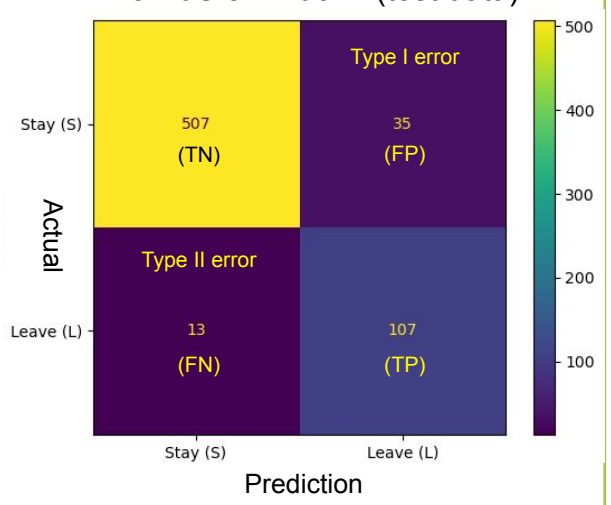
From the table, you can see that our model correctly identified roughly 9 of every 10 employees in the set of unseen test data who decided to leave last year (recall = 0.89).

Type I vs. Type II errors

What exactly is a false negative? A false negative occurs when the model predicts that an employee will stay, but the employee actually decides to leave. In order to be effective in our workforce planning efforts, we need to be able to identify the employees at high risk for attrition. We don't want to miss an opportunity to provide an incentive to stay to an employee at risk of leaving. When we tune our model to be **sensitive** to an employee's decision to leave, we minimize these Type II errors.

By contrast, we are less concerned with "over-identifying" at risk employees. We would rather allow for less **precision** in our model, meaning that it is less important for our positive predictions to be correct. You can see from the confusion matrix (right) that our selected model makes fewer Type II errors than Type I errors.

Confusion Matrix (test data)



What factors appear to influence turnover?

- **Work-life balance** is king.
- **Hard working, highly rated** employees tend to stay.
- **Job involvement** is very important.
- **Job satisfaction** is also important.
- Base monthly income doesn't seem to matter, but a **salary hike** helps.
- **Environment satisfaction** matters, but not as much as you might think.

Note: Our study was not designed to prove causation, but our model was able to predict employee turnover with 93% accuracy using these key factors.

Our model estimated an odds ratio for each factor, which provides an indication of relative importance to an employee's decision to leave (L) or stay (S), when holding other factors constant. We estimated the odds ratio for work-life balance to be ~0.157, or roughly 1:6 (L:S). By contrast, the odds ratio for environment satisfaction was ~0.544, or roughy 1:2 (L:S) -- much closer to even odds of 1:1.

Image credit: inspiredpencil.com

Recommendations: What should we do?

	Reduction in turnover	Employees retained	Total savings (M)
Limit Business Travel	1.85 %	68	**2.79**
Employee Appreciation	1.67 %	62	2.54
Workplace Flexibility	1.36 %	50	2.05

Focus on work-life balance

We have **high confidence** in our model's ability to predict employee turnover based on its performance (accuracy 93%, recall 9 out of 10) using standardized variables.

Of the proposed workforce planning initiatives, **limiting business travel** yielded the greatest reduction in turnover (1.85%).

We recommend moving forward with an initiative to improve work-life balance by limiting business travel, which should result in roughly 68 fewer annual employee departures and a total cost savings of **2.79M rupees** annually.

Image credit: vecteezy.com

Limitations

Model Assumptions

A logistic regression model assumes that we can combine the individual effect of each explanatory variable in a linear, additive fashion. This assumption does not account for the possibility of interactions between variables.

For example, it is reasonable to believe that allowing work from home will have a greater effect on employees who live far from their work location, which could be expressed mathematically as an interaction between those two variables. However, the logistic regression model in its basic form does not allow for this kind of interaction. In order to give the model a way to account for interactions between variables, we can derive a new explanatory variable that reflects what we believe to be the mechanism of interaction between the original variables.

Limitations in the Data

Because our data is taken from the company's internal human resource information system, we can have high confidence in its fidelity, but errors still occur. We should assume that a small part of the data is problematic. For example, data may have been entered incorrectly, or some data may be missing.

If only a small portion of the data is somehow problematic, we have different options to deal with those problems depending on their nature. If we find that less than five percent of the data has values missing completely at random, we can simply remove those problematic data items from the analysis. When there is a pattern to the problems, we need to use more sophisticated techniques to prevent the final set of data from being skewed or biased.

Model Limitations

Our model estimates how different factors influence attrition **on average**. It can help us understand, for example, how raising base pay will **tend** to influence an employee's decision on whether to leave the company. We can also use the model to predict which employees are **at higher risk** of attrition... but nothing is certain in the individual case.

Different employees may be influenced by some factors more than others. Each employee's decision is also subject to factors for which we have no data and which are not included in the model. The variation we observe in individual decisions is reflective of this uncertainty. We should not expect our model to make perfect predictions.

We tuned our model to favor recall over precision, so we should expect more false positive results (Type I errors) than false negatives (Type II errors). Every model of a complex real-world process or system must balance between these two types of errors.

Final Thoughts

The limitations discussed here are known and do not change the recommendations we developed through this analysis. Examples of problems severe enough to alter our recommendations might include:

- Significant portion of data missing or corrupt
- No significant relationship between any of the explanatory variables and the response variable (attrition) -- in other words, none of the variables for which we have data influence an employee's decision to leave or stay in a meaningful way