Addressing Employee Attrition

Utilizing Logistic Regression to determine causes of employee attrition and the effectiveness of retention initiatives

Nathan Cantwell

Table of Contents

01 Problem Statement (I)
Descriptive and quantitative assessment of employee attrition at Acme Aroma

02 Modeling Approach (II)
Explanation of the logistic regression model and how it will be evaluated

03 Insights (III-V)
Findings from exploration of the dataset, evaluation of model performance, and root association analysis

04 Recommendations (VI)
Action steps for implementation of the model, and efficacy of employee relief initiatives

05 Potential Limitations (VII)
Discussion of limitation of this analysis, from both the model and underlying data

Problem Statement

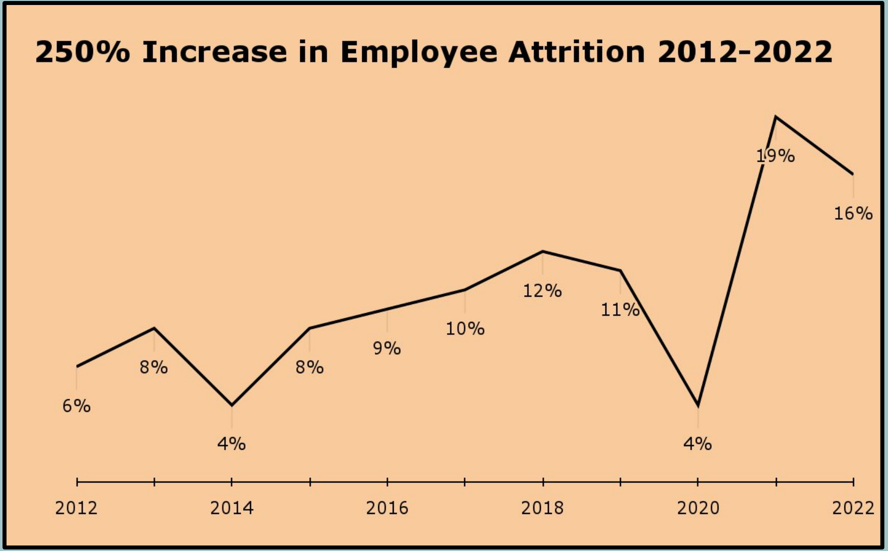
In the last ten years Acme Aroma has seen employee ***attrition rise from 6% to 16%***. Since 2021, rehiring due to attrition has ***cost the company ₹30MM***.

In the same ten year period there were ***66% fewer job applications*** per opening. To address this, predictive insights into individual attrition risk can help workforce planning lift employee retention.

Additionally, employee ***job satisfaction fell by 17%***. Insights into the root causes of dissatisfaction and attrition could reveal which retention initiatives would be most impactful.

For these purposes Acme Aroma would benefit from analysis of:

- Which employees have high attrition risk
- The root associations of attrition
- The predicted impact of different initiatives



250% Increase in Employee Attrition 2012-2022

Modeling Approach

For the three analysis goals, the data science team built a ***logistic regression model*** using HR employee data. The model uses variables such as: distance to work, gender, income, years at Acme, and survey scores to ***predict attrition for each employee***. The model also tell us ***which variables have the strongest relationship with attrition***.

For logistic regression we will ***evaluate model performance using false positives (FP) and false negatives (FN)***. In our case, an FP is when an employee stays at Acme that the model predicted would leave. An FN is when an employee is predicted to stay, but actually leaves.

We will tune our model to minimize FP's or FN's, but neither can be completely eliminated. Considering the recent ***₹30MM costs in rehiring, it is best to avoid false negatives*** where the model misses that an employee will leave. Therefore, we will tune our model by ***"recall" score: the percentage of lost employees correctly identified***.

<u>F1 Score</u>

Using only recall can lead the model to simply identify all employees as high risk; we will also evaluate performance using the "F1" score. Balanced models that ***correctly identify lost and retained employees at high percentages*** will have the highest F1 scores.

Insights, Data Exploration

Before statistical modeling, it is important to look into the dataset more closely.

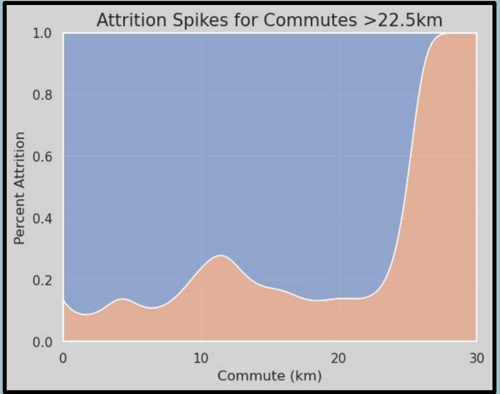
Analyzing our workforce, 70% of employees are below job level 3. Similarly, 68% of employees are under the age of 40. Considering Acme Aroma’s headcount has grown 53% in the last 10 years, it is not surprising that ***our workforce is relatively young and low level***.

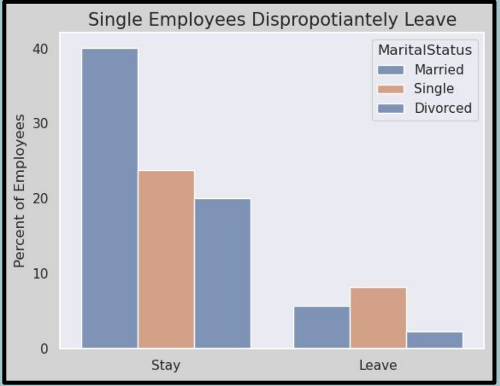
From career histories, variables detailing ***job role and field of education were not significantly different*** between employees who left or stayed.

From personal characteristics, two variables of note were ***commute distance and marital status***.

Attrition increases dramatically around a 22.5km commute, and ***all employees beyond 24km have left***. A marital status of “single” indicates a higher risk of attrition. While they make up only 28% of retained employees, ***single employees account for 50% attrition***.

Both of these variables will prove important to the model predictions in the following slides.





Insights, Model Performance

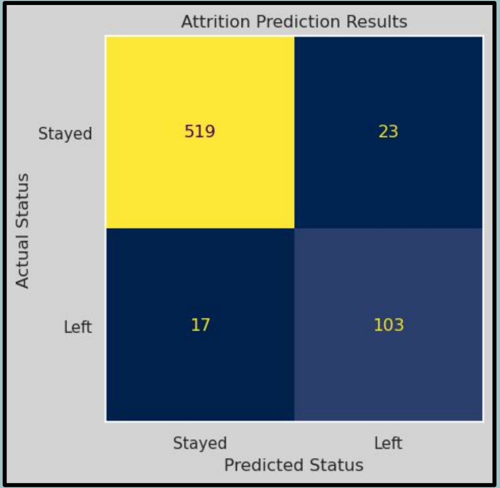
We have tuned the model to maximize our performance scores. Keep in mind, ***perfect scores are not obtainable***, and in fact would indicate a modeling error.

The final model used over 5,000 employee profiles during training and includes 36 different variables for predicting attrition.

In discussing our modeling approach we defined our ***key interest in minimizing FN's***, and selected recall and F1 scores to evaluate model performance.

When testing with employee profiles new to the model, it correctly identified employees staying in 519 of 542 cases and attrition in 103 of 120 cases. This translates to ***23 false positives (FP) and 17 false negatives (FN)***, respectively.

The final model had a ***recall score of 86% and an F1 score of 84%***. This means our model correctly identifies 86% of lost employees, while also predicting well for retained employees.



Insights, Root Associations

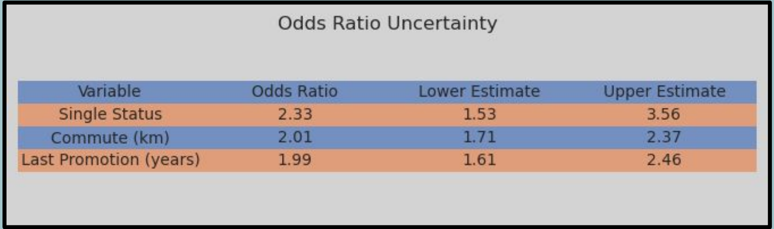
To better understand the root associations of employee attrition, we can investigate which variables the model found significant in making it’s predictions.

The ***3 top indicators*** for association with attrition were: ***single marital status, commute distance, and years since promotion***. Single status and long commutes are expected from exploratory analysis. Years since promotion is also an intuitive indicator of attrition.

Odds ratios, when controlling for other factors**:**

- ***Single employees*** are ***2.3x more likely*** to leave, when compared to married or divorced employees
- For each ***7.5km increase in commute***, an employee is ***twice as likely*** to leave
- For every ***3 years since a promotion***, an is employee is ***twice as likely*** to leave

Also noteworthy is a cohort of 14 ***employees without company history*** that the model found were ***5.4x more likely*** to leave.



Odds Ratio Uncertainty

Variable	Odds Ratio	Lower Estimate	Upper Estimate
Single Status	2.33	1.53	3.56
Commute (km)	2.01	1.71	2.37
Last Promotion (years)	1.99	1.61	2.46

Recommendations

The data science team would recommend implementing this logistic regression model. Recent ***hiring and training costs total to ₹30MM***, and reducing employee attrition can yield significant savings for Acme. With a recall score of 86%, the model can ***identify attrition in 5 out of 6 cases***.

From the root association analysis, we would recommend ***hiring individuals living within 22km of the office, and do not have single marital status***. More frequent promotions would be more costly and perhaps unsustainable for Acme.

Retention savings, estimated hiring and training savings from each employee retention initiative**:**

- Increased base pay would retain 1 employee and save ₹30k.
- Additional professional development would retain 4 employees and save ₹120k.
- Allowing WFH would retain 6 employees, ₹180k saved
- Employee appreciation initiatives would retain 9 employees and save ₹270k.
- ***Reducing required business travel would retain 15 employees and save ₹450k.***

While the WFH initiative is predicted to retain only 6 employees, this does not account for the possible effect on employee attrition due to commute distance. ***Employees with long commutes may stay with Acme if the commute is less frequent***.

Potential Limitations

Data limitations, a statistical model is only as good as the data it is trained on**:**

- ***Missing values*** in the number of companies worked for, and total working years variables.
- 16% of employees in the data leave, while 84% stay. This is a ***class imbalance***; model training prefers balanced classes (50/50).
- All employees in the data have filled out the survey. There may be ***unrepresented employees*** with incomplete surveys, depending on how this data was collected.

There are techniques to handle missing data and class imbalance. The possible exclusion of employees without survey data is more concerning.

Model limitations, a statistical model is an imperfect representation of reality**:**

- Logistic regression ***does not determine a causal relationship***, only what factors coincide with attrition. We should not assume causation from model results.
- This model is ***trained on historic data***. Significant industry shifts may change the factors for attrition and harm this model's performance.

Even with these limitations, attrition is a growing problem that needs to be understood. This model grants us insight into which factors are related to attrition and who is at risk.