# Acme Aroma Workforce Planning Project

Data Scientist

**Chih-Han (Dora) Yeh**

**Date: 2023/04/18**

# Table of Content

Source [3]: PAIRL

# PROBLEM

## Problems Acme Aroma Is Facing

- **Increasing turnover rate:** The employee turnover rate has been increasing since 2020. As shown in the visualization 'Acme Aroma Employee Turnover Rate', in 2020, the turnover rate was 4%, but in 2022, it was 16%. The increase rate is 300%.

- **Increasing acquisition cost:** The internal data shows the acquisition cost per employee goes from ₹15,000 in 2021 to ₹ 30,000 in 2022. The increase rate is 100%.

- **Decreasing job applications:** The amount of job applications per position is going down significantly. The number of job applications per position went from 17 in 2020 to 8 in 2022, the decrease rate is 53%.

- **Decreasing job satisfaction:** Job satisfaction is going down significantly. In the last 3 years, job satisfaction went from 3 to 2.7, the decrease rate is 10%.

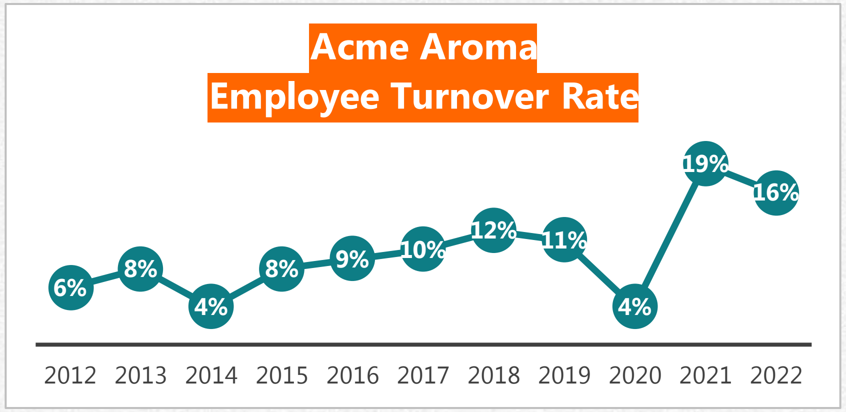
## The Need For Root Association And Predictive Insights

The high turnover rate is the major and core problem. It is adding costs and causing inefficiency to the company.

The high turnover rate impacts company production and sales process negatively due to the inefficiency of inexperienced employees. The increased acquisition cost adds costs. Besides, the decreasing job applications and job satisfaction are also impacting the company negatively. Therefore, the retention of employees is important.

To make it, understanding the root causes of the turnover problem is critical. It will help the company to understand what variables have a high association with employee turnover, to create strategies that help to reduce the turnover rate, and plan the workforce effectively.

Besides, gaining predictive insights can help us plan strategies in advance as well for the company to lower the turnover rate, which can help improve manufacturing efficiency, and optimize acquisition cost, so the company can become more competitive in the market.



**"Employee Turnover Rate has been increasing in recent 2 years."**

Source [1] Acme Aroma Project Introduction

# APPROACH

## Project Goal

The high turnover rate is adding costs and causing inefficiency to the company. To address this issue, our project aims to **understand the root causes** associated with high employee turnover and **produce a predictive model** to gain predictive insights into the problem.

## Data

- The data set for the project is from the human resources information system (HRIS).

- The dataset consists of 4410 employee records, and a sample record is presented in ***'Table 1' in Appendix***. Our objective is to utilize the dataset to gain insights into the **'*Attrition*' (the response variable)** field, which is a "Yes/No" variable that indicates whether an employee has left.

## Analytical Approach

We plan to utilize a logistic regression model in this project. The model provides benefits:

1) **enabling us to gain predictive insights**

2) **letting us understand the root causes associated with the turnover problem**

With Logistic regression, we can establish the relationship between the outcome variable ('leave/not leave') and factors (salary, job satisfaction level, etc.) that could potentially impact the outcome.

The model estimates the probability of the outcome variable given the predictor variables.

## Errors to Avoid

- **Type 1 error:** It is the error that we're predicting turnover for an employee who doesn't leave.

- **Type 2 error:** It is the error that we are failing to predict turnover for an employee who actually leaves.

It is important to identify Type 1 errors because they could result in incorrect responses given incorrect predictions. Type 2 errors are problems too because they can lead to missed opportunities to change.

In the evaluation process of a classification model, there is often a **trade-off between type 1 and type 2 errors**. Increasing one type of error usually leads to a decrease in the other type of error. For example, if the model is too cautious and avoids making the error of predicting turnover for an employee who doesn't leave, it may classify many people who leave as not leaving, leading to a high type 2 error but a low type 1 error. And vice versa.

## Evaluation Methods

- **Recall:** It is a measure of completeness. It's saying how many of those who leave do the model capture. For example, "Recall for this model is 60%. This means 60% of actual attritions is identified correctly in the model."

- **Precision:** It is a measure of exactness. It measures the proportion of attrition identifications by the model that is correct. For example, "Precision for this model is 80%. This means when the model predicts the employee will leave, it is correct 80% of the time."

- **ROC curve & AUC:** the ROC is a visualized representation of a classification model performance. An AUC (Area Under the ROC Curve) score of 1 means perfect prediction model performance and a score of 0 means the worst.

We can evaluate the model by the above metrics to check the model performance, and the ability of the model to correctly distinguish between positive and negative instances.

## Data I Plan to Use

According to ***'Table 1' in Appendix***, the response variable for our analysis is *'Attrition'*. I ranked the other variables by the degree of correlation with '*Attrition*' and the highest correlated variables with *'Attrition'* are the potential key predictor variables I plan to use.

Besides, we also need to include variables related to the **Initiative Options provided by HR** to help us make recommendations.

Source [2]: Introduction to Logistic Regression

# INSIGHTS

## Model Description

We build a logistic regression model for the project. As shown in ***'Table 2'***, the response variable for our analysis is *'Attrition'*. The potential predictor variables are ranked by the degree of correlation with *'Attrition'*.

We do modeling with variables highlighted which are highly correlated with *'Attrition'*, with the magnitudes of correlation over 0.15. *'Distance From Home'* and *'Total Working Years'* are dropped because it doesn't contribute much to our model performance, that is, they don't help us much in making the prediction.

Variables *'Training Times Last Year'*, and *'Monthly Income'* are also included to make sure we evaluate all the variables related to the Initiative Options provided by HR to help us make recommendations.

We are using those variables highlighted in ***'Table 2'*** to **gain predictive insights** into the turnover problem, and also do **root association analysis** to identify those variables with a high degree of association with employee turnover.

## Variables Standardization

To make it clearer, in order to prevent the model impacted by the different scales in different variables, we standardize the variables before modeling. And **the output we see in the following pages is the results that we convert back to the natural units** by using the size of variables' corresponding standard deviations to help communication. Therefore, the model is robust and not impacted by the scale difference.

## More In-Depth Discussion In p.6 & p.7

In the following 'INSIGHTS' section, we will cover:

- the **performance evaluation** using key metrics of the model
- **root association analysis** that covers the concrete association between these key predictor variables and the outcome variable 'Attrition'.

Ranked by Magnitude of Correlation with 'Attrition'

| **Key Predictor Variables to Predict *'Attrition'*** |
|---|
| ▨ *Work Life Balance* |
| ▨ *Percent Salary Hike* |
| ▨ *Job Satisfaction* |
| ▨ *Environment Satisfaction* |
| *Distance From Home* |
| *Total Working Years* |
| ▨ *Age* |
| ▨ *Years With Current Manager* |
| *Years At Company* |
| ▨ *Training Times Last Year* |
| *Num Companies Worked* |
| ▨ *Years Since Last Promotion* |
| ▨ *Monthly Income* |

▨ Key Predictor Variable Used in Model

***(Table 2)***

Source [1] Acme Aroma Project Introduction

# INSIGHTS

We are going to discuss the final model performance results using the key performance measures mentioned in the ‘APPROACH’ section.

## Type 1 Error and Type 2 Error

It is important to identify Type 1 errors and Type 2 errors because they could result in incorrect responses.

- **Type 1 errors in our model:** The error that we're predicting turnover for an employee who doesn't leave is 64 out of 661 in the test set. (Test set is a set of data extracted from the dataset especially used for evaluating model performance.)
- **Type 2 errors in our model:** The error that we are failing to predict turnover for an employee who actually leaves is 13 out of 661 in the test set.

The Type 1 errors are almost 5 times more than Type 2 errors. This implies that **our final model is more aggressive** than cautious that has more tendency to make the error of predicting turnover for an employee who does leave, it may classify more people who are not leaving as leaving.

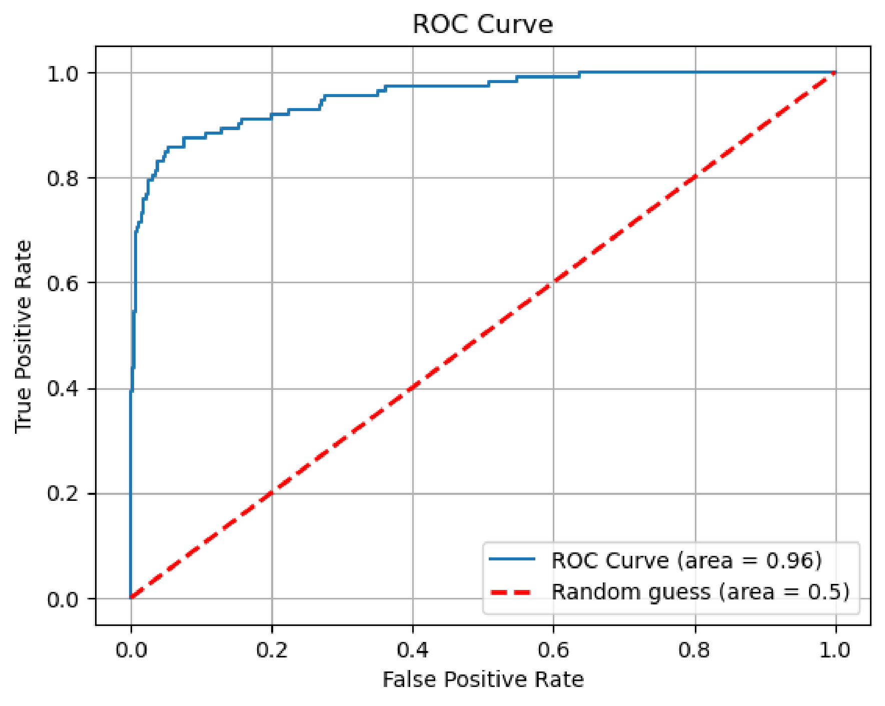
## Evaluation Methods

- **Recall:** It is a measure of completeness. It's saying how many of those who leave do the model capture. In this model, Recall for this model is 89%. This means 89% of actual attritions are identified correctly in the model.
- **Precision:** It is a measure of exactness. It measures the proportion of attrition identifications by the model that is correct. In this model, Precision for this model is 92%. This means when the model predicts the employee will leave, it is correct 92% of the time.
- **ROC curve & AUC:** the ROC is a visualized representation of a classification model performance. An AUC (Area Under the ROC Curve) score of 1 means perfect prediction model performance and a score of 0 means the worst. In our model, the AUC score is 0.96, which makes mild mistakes, but the performance is overall great.

We can see from the ROC Curve below that the model do much better than random guess and not far from perfection (1.0).

To make some side notes, the ROC curve is made of

- ‘True Positive Rate’ on the y-axis: the proportion of actual positives was identified correctly in the model.
- ‘False Positive Rate’ on the x-axis: the proportion that the model labels a class is positive when the true label is actually negative.



Source [2]: Introduction to Logistic Regression

# INSIGHTS

## Root Association Analysis

Our model not only enables us to gain predictive insights but also let us understand the root causes associated with the turnover problem. It helps us do the root association analysis.

## Coefficients-Based Analysis

With Logistic regression, we can establish the relationship between the outcome variable ***'Attrition'*** ('leave/not leave') and predictor variables. We can analyze the association with coefficients in the model. The original coefficients are transformed into odd ratios and probability to help us interpret the root association between the predictor variables and *'Attrition'*.

The magnitude of association between the predictors and the outcome variable, and corresponding interpretations are illustrated in the below table.

The below predictor variables are listed by the ranking of change of probability to help us see clearly the association strength between predictor variables and outcome variables. It helps us identify those variables with a high degree of association with employee turnover: *Training Times Last Year, Environment Satisfaction, Job Satisfaction, and followed by Work Life Balance and Years With Current Manager.*

■ Highly Associated Predictor Variable

| PREDICTOR VARIABLES | ODD RATIO BY UNITS | PROBABILITY BY UNITS | INTERPRETATION OF ASSOCIATION WITH ODD RATIO | INTERPRETATION OF ASSOCIATION WITH PROBABILITY |
|---|---|---|---|---|
| ***Training Times Last Year*** | **0.69** | **0.41** | **Increasing *Training Times Last Year* by 1 time, will drop the odds of attrition by 31%** | **Increasing *Training Times Last Year* by 1 time, will drop the probability of attrition by 41%** |
| ***Environment Satisfaction*** | **0.50** | **0.33** | **Increasing *Environment Satisfaction* by 1 degree, will drop the odds of attrition by 50%** | **Increasing *Environment Satisfaction* by 1 degree, will drop the probability of attrition by 33%** |
| ***Job Satisfaction*** | **0.38** | **0.23** | **Increasing *Job Satisfaction* by 1 degree, will drop the odds of attrition by 62%** | **Increasing *Job Satisfaction* by 1 degree, will drop the probability of attrition by 23%** |
| *Work Life Balance* | 0.19 | 0.16 | Increasing *Work Life Balance* by 1 degree, will drop the odds of attrition by 81% | Increasing *Work Life Balance* by 1 degree, will drop the probability of attrition by 16% |
| *Years With Current Manager* | 0.19 | 0.16 | Increasing *Years With Current Manager* by 1 year, will drop the odds of attrition by 81% | Increasing *Years With Current Manager* by 1 year, will drop the probability of attrition by 16% |
| *Age* | 0.07 | 0.07 | Increasing *Age* by 1 year, will drop the odds of attrition by 93% | Increasing *Age* by 1 year, will drop the probability of attrition by 7% |
| *Percent Salary Hike* | 0.04 | 0.04 | Increasing *Salary Hike* by 1 percent compared with last year, will drop the odds of attrition by 96% | Increasing *Salary Hike* by 1 percent compared with last year, will drop the probability of attrition by 4% |
| *Monthly Income* | 0.02 | 0.02 | Increasing *Monthly Income* by ₹1000, will drop the odds of attrition by 98% | Increasing *Monthly Income* by ₹1000, will drop the probability of attrition by 2% |

Source [5] 2_employee_turn_model_metrics

# RECOMMENDATIONS

## Suggesting Initiative Based On Cost Effectiveness

The Initiative Options provided by HR are listed in ***'Table 3'.*** To give recommendations for how to lower employee turnover from the potential initiative list, we utilize the model coefficients to calculate expected values and get the change in probability of *'Attrition'* to estimate how many fewer attritions would happen if each initiative implemented.

Given Acme Aroma's internal data, "the cost of recruitment, training, and general onboarding is approximately 50 - 75% of a full-time annual salary." We estimate the presumed savings through 50% ~ 75% of the mean annual salary in the history data from HRIS. We also estimate the anticipated decline in the number of attrition by the total number of employees who haven't left (3699 people in the dataset) with the corresponding decrease in probabilities of attrition of each initiative.

In ***'Table 3',*** we can see the initiative 'Workplace flexibility' have the greatest decrease in the probability of attrition and has the greatest presumed saving of ₹ 185.0 ~ ₹ 277.5 millions. Given the estimated savings, we would recommend implementing this initiative that allows work-from-home.

## Recommend Using the Final Model

I would recommend implementing the turnover model which has excellent performance of an AUC score of 96%. It could help the company by

- **enabling us to gain predictive insights.** We could feed the model of the predictor variables and get the prediction about '*Attrition*', about whether the employee is leaving or not. It will help the company to plan ahead and mitigate the inefficiency problems.
- **letting us understand the root causes** associated with the turnover problem. As shown in ***'INSIGHTS'***, we understand the impact the change of predictor variables could have on the outcome variable. This will help us create strategies to deal with the high turnover problem.
- **helping us evaluate the impact and cost savings** of potential initiatives to address the turnover problem. As demonstrated with the Initiative Options, the model gives precise suggestions and cost impact estimations.

Therefore, I recommend implementing our turnover model given its great performance and functionality on the company's needs.

| INITIATIVE | DESCRIPTION | EFFECT | DECREASE IN PROBABILITY OF ATTRITION |
|---|---|---|---|
| **Workplace flexibility** | **Allow work-from-home** | ***Environment Satisfaction*** **increases by 0.75** | **0.2475** |
| Training | Provide additional professional development | Moves *Training Times Last Year* up 0.5 on average | 0.2 |
| Pay base increase | Increase base pay for all employees | Increase *Monthly Income* by 7500 | 0.15 |
| Employee Appreciation | Various employee appreciation initiatives | *Job Satisfaction* increases by 0.5 | 0.115 |
| Limit business travel | Reduce the amount of required business travel | *Work Life Balance* increases by 0.3 | 0.048 |

***(Table 3)***

🟧 Suggested Initiative

Source [5] 2_employee_turn_model_metrics

# LIMITATIONS

## Correlation is not Causation

The Logistic regression model estimates the probability of the outcome variable given the predictor variables, but it **does not establish a causal relationship**.

In logistic regression, we cannot conclude from a strong correlation between a predictor variable and the outcome variable that the predictor variable causes the outcome variable. There could be other factors (**confounding variables**) that are not included in the model that affect the outcome. The correlation between the predictor and outcome variables may be resulted from the associations with the confounding variables.

**To confirm the causal relationship** between predictor and outcome variables**, a randomized controlled experiment is needed** as other confounding variables are controlled for through randomization.

Therefore, it is important to interpret the results of logistic regression analysis carefully and avoid assuming there are causal relationships based solely on correlation findings.

## Issues with the imbalance in the dataset

**Unbalanced data can lead to biased result and produce a model that underrepresents the minority class.** As we can see from *the bar chart,* there is much less outcome variable 'attrition' compared with 'non-attrition of an employee'. More specifically, **only 16% of attrition consists in our dataset.**

The imbalanced nature of the dataset makes the accuracy score a poor choice to evaluate our model performance. The number of attritions is small when compared to the number of non-attrition, and this leads to a **class imbalance problem**, that is one class is much more than the other.

If we have 16% of attritions and 84% of non-attritions, a classifier that predicts any input variable as a 'non-attrition' would get 84% accuracy even though it doesn't predict right on any of the attrition. Therefore, the raw percent accuracy score alone is misleading under this kind of situation. **With Precision and recall, we can better evaluate the model**.

Also, to deal with the imbalance problem in the dataset, we use a technique named **SMOTE** to generate new synthetic data points to adjust the class distribution of a data set. The technique aims to deal with the imbalance problem, but it might lead to overfitting, which means the model cannot generalize well to new and unseen data. Therefore, we need to pay attention to the risk.

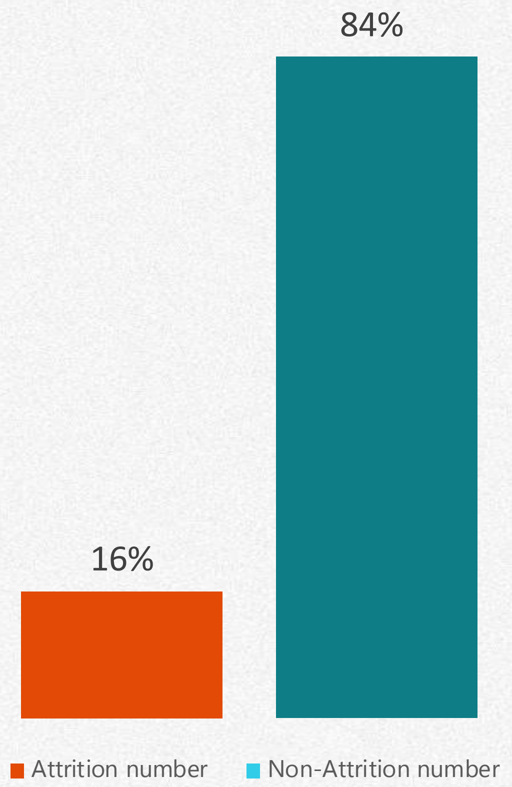
## Issues with data quality/missing data

Our dataset is not complete **as we have two variables with missing values** (Number of Companies Worked: missing 0.43% and Total Working Years: missing 0.2%). We have tackled the missing data by filling them with the mean numbers, and it should not impact the model a lot. However, this reminds us that we should be aware of the dataset quality is not perfect in the first place. We might want to look back if our analysis found something illogical.

## Given the limitations,

My recommendation will carefully **emphasize the correlation instead of causation relationship between variables** at this point since this is what the logistics regression model can tell us.

## Imbalance of Attrition and Non-Attrition Number in Dataset



Source [4] 1_employee_turn_data_prep.ipynb

# APPENDIX

**Complete Sample Record with Variables Ranked By Correlation with 'Attrition'**

| ***Variable*** | ***Sample value*** |
|---|---|
| Attrition | 0 |
| Work Life Balance | 4 |
| Percent Salary Hike | 11 |
| Job Satisfaction | 2 |
| Environment Satisfaction | 2 |
| Distance From Home | 2 |
| Total Working Years | 13 |
| Age | 38 |
| Years With Current Manager | 5 |
| Years At Company | 8 |
| Training Times Last Year | 5 |
| Num Companies Worked | 3 |
| Years Since Last Promotion | 7 |
| Monthly Income | 83210 |
| Gender | Male |
| Education | 5 |
| Marital Status | Married |
| Job Level | 3 |
| Stock Option Level | 3 |
| Education Field | Life_Sciences |
| Job Role | Human_Resources |
| Business Travel | Non-Travel |

*(Table 1)*

■ Response Variable

□ Potential Key Predictor Variable

Source [1] Acme Aroma Project Introduction

| **Reference List** | | |
|---|---|---|
| Source | [1] | Acme Aroma Project Introduction |
| Source | [2] | Introduction to Logistic Regression |
| Source | [3] | PAIRL |
| Source | [4] | 1_employee_turn_data_prep.ipynb |
| Source | [5] | 2_employee_turn_model_metrics |