Denise Dodd

Predicting Attrition

White Paper

Business Problem

"Reducing employee attrition even by a small amount can have significant and meaningful time and cost savings for an organization, once recruiting, training, burnout, and other hidden costs are factored in" (Wallace, 2023). When an employee leaves a company, the business must invest time and money to post a job, filter through resumes, conduct interviews, hire, and onboard a new employee, train the replacement, and transfer duties to the new hire. Additionally, the historical knowledge that the original employee had is a resource that the company cannot regain. Therefore, it is to the company's benefit to determine what contributes to attrition and predict future attrition so they can budget and plan accordingly.

Background/History

For this study, I will be utilizing a dataset titled "Employee Attrition" (Kaggle, 2022). This is a dataset of about 4,500 employees of which roughly 16% are no longer with the company. There is a mix of 24 numerical and categorical variables. Numerical variables include data such as the employee's age, commute distance, years with the company, and salary. Categorical variables include what department the employee works in, marital status, job role, and gender.

In an effort to lower the cost of replacing an employee, preserve historical knowledge, and limit workflow disruption, the company would like to decrease this attrition rate. Throughout the course of this study, I am hoping to identify which factors contribute the most to attrition and predict if an employee is or is not likely to attrite.

Data Explanation (Data Prep/Data Dictionary/etc)

When reviewing for nulls in the dataset, I found 19 nulls in the "NumCompaniesWorked" col and 9 nulls in the "TotalWorkingYears" column. If this is an employee's first job and/or they have not spent a full year in the workforce, these null values make sense. Therefore, I will be replacing these null values with zero.

I then created histograms of all numerical values and bar graphs of categorical data to determine the spread of the data. In doing so, I found that the "Over18", "EmployeeCount" and "StandardHours" columns all have the same response for each entry. Additionally, there is an "EmployeeID" column with a unique number for each employee. These columns will not provide any context to the predictive model so they will be dropped from the dataframe.

Dummy variables were made of categorical columns so they can be passed to the classification models. The first column of each dummy variable is dropped to prevent multicollinearity. The dummy variables are joined with the numerical variables. Attrition is designated as the target variable and all other columns are designated as features variables. The features and target variables are split into 80/20 training and testing sets.

Methods/ Analysis

I placed several classification models (Decision Tree, K-Nearest Neighbors, Logistic Regression) through a loop which fit the models, made predictions, and printed a classification report with Precision,

Recall, F1-Score and Accuracy metrics. Using these metrics, it was determined that the Decision Tree Classifier was the best fit for this data as explained below (Kanstren, 2023):

> **Precision:** With a precision score of 1.00, all instances where the model predicted 0 (no attrition) were also a 0 in the test data. With a precision score of .96, 96% of the instances where the model predicted 1 (yes attrition) were also a 1 in the test set. (True Positives/All Predicted Positives)
>
> **Recall:** With a recall score of .99, this model correctly identified 99% of no attrition instances. With a recall score of 1.00 this model accurately identified all of the yes attrition instances. (Correctly Predicted Positives/Actual Positives in Test Set)
>
> **F1-Score:** The F1-Score is an average of the precision and recall scores.
>
> **Accuracy:** With an accuracy score of .99, 99% of this model's predictions are correct.

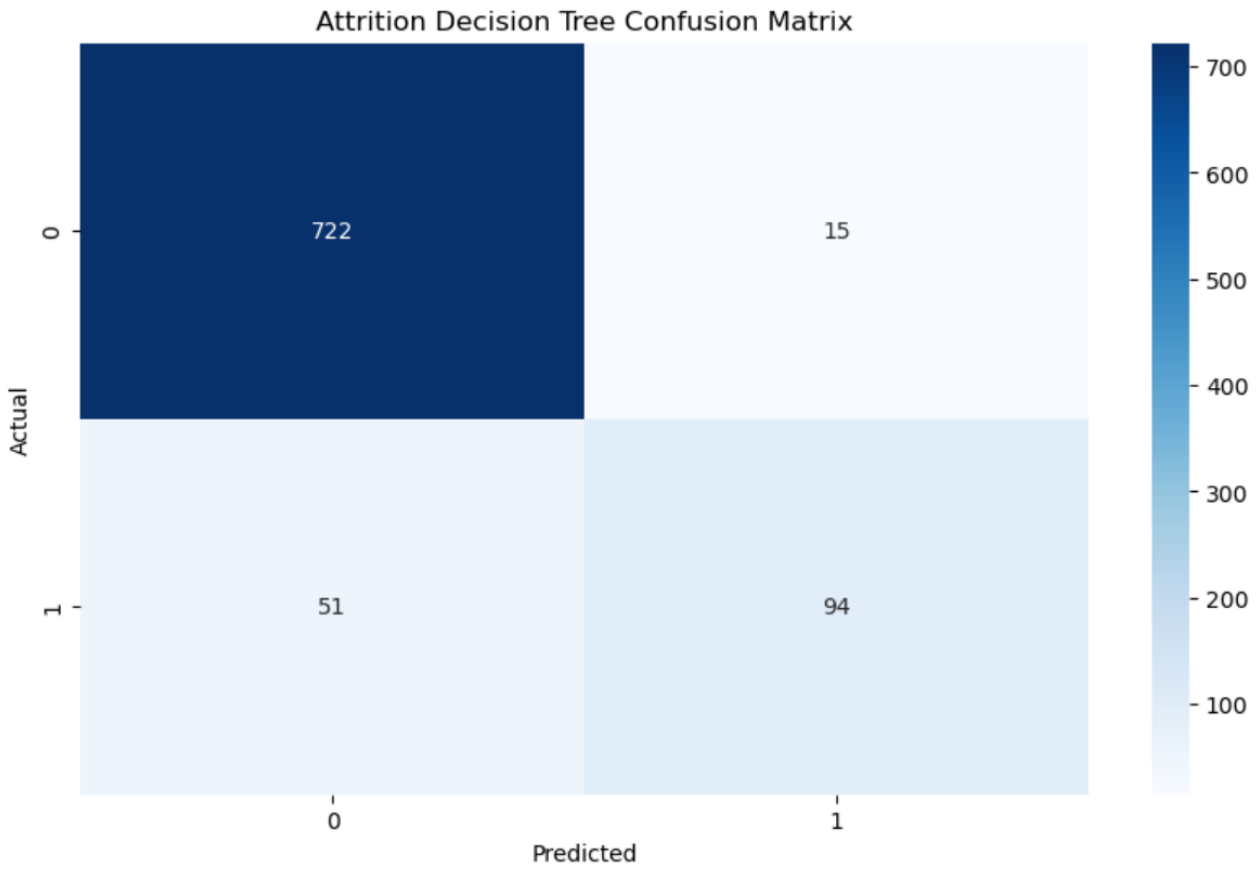
These metrics are promising, but they are so good they raise concerns of multicollinearity. Multicollinearity is when two variables in the dataframe are highly correlated with one another. For example, it can be anticipated that the "Age" variable is correlated with the “TotalWorkingYears” variable. The older one is, the longer they have probably worked. To account for the possibility of multicollinearity, I used a grid search to find the best hyperparameters for my Decision Tree Classifier. I then fit and made predictions based on a hyptertuned Decision Tree Classifier. My metrics dropped a bit, but I still had 93% accuracy and higher metrics than the K-Nearest Neighbors and Logistic Regression models in my loop.

> **Precision:** With a precision score of .93, 93% of the times of the model’s prediction of 0 (no attrition) were also a 0 in the test data. With a precision score of .86, 86% of the instances where the model predicted 1 (yes attrition) were also a 1 in the test set. (True Positives/All Predicted Positives)
>
> **Recall:** With a recall score of .98, this model correctly identified 98% of no attrition instances. With a recall score of .65 this model accurately identified 65% yes attrition instances. (Correctly Predicted Positives/Actual Positives in Test Set)
>
> **F1-Score:** The F1-Score is an average of the precision and recall scores.
>
> **Accuracy:** With an accuracy score of .93, 93% of this model's predictions are correct.

The hypertuned Decision Tree Classifier alleviates my concerns of multicollinearity, and still provides metrics which give me confidence in my model.

Visualizations/Analysis

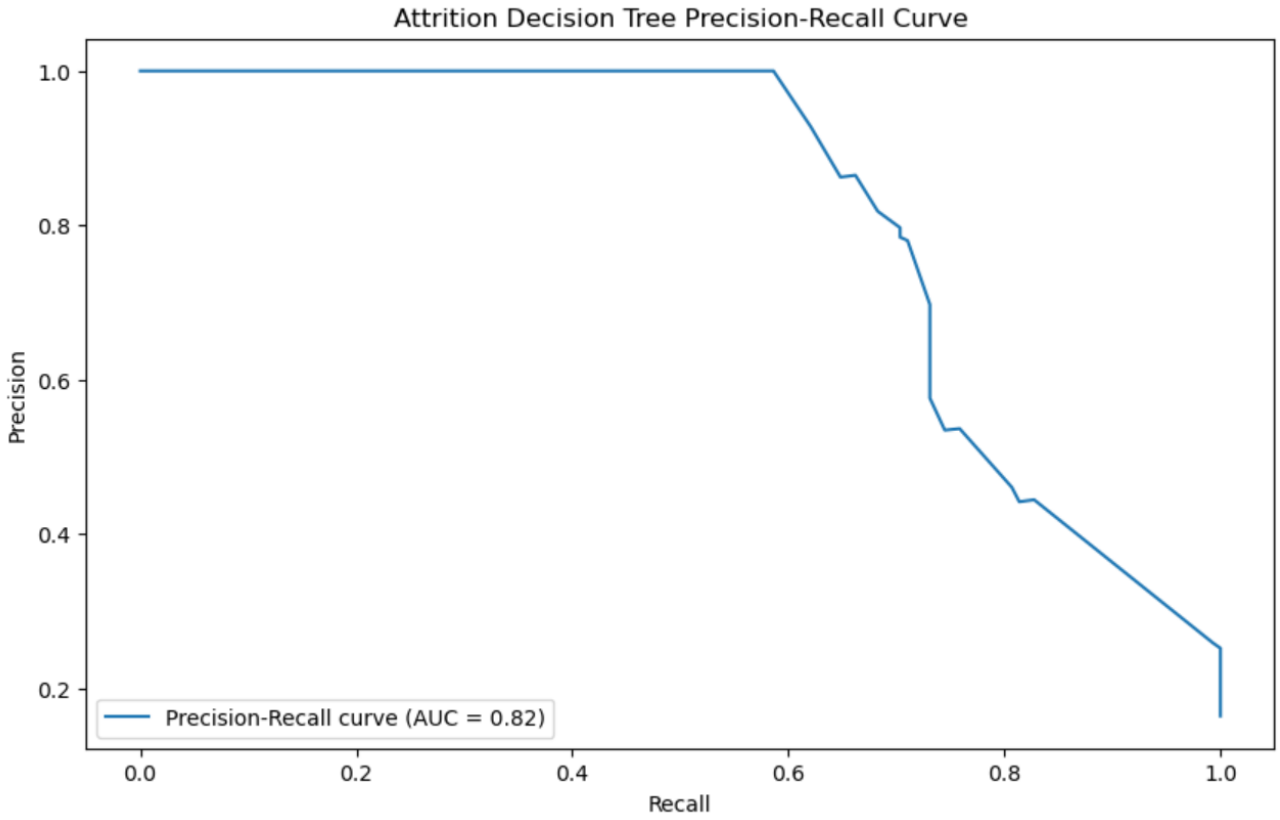
Confusion Matrix - To better visualize the classification efficiency of my model, I created a Confusion Matrix shown below.



This confusion matrix shows that the Decision Tree classified:

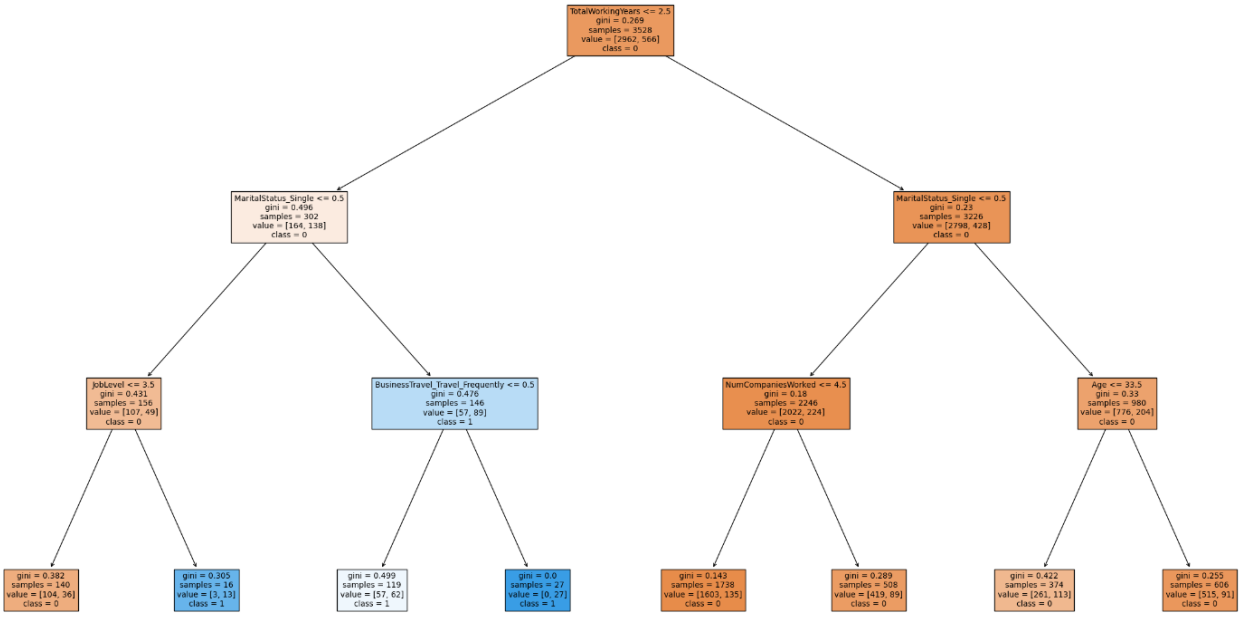
- 722 of the employees who did not attrite correctly. (TP)
- 94 of the employees who did attrite correctly (TN)
- 15 of the employees who did not attrite incorrectly (FP)
- 51 of the employees who did attrite incorrectly (FN)

Precision-Recall Curve - Because there is an imbalance in the number of "Yes" (16%) and "No" (84%) observations for the Attrition variable, I will use a Precision-Recall curve to evaluate how well my model classified attrition entries as opposed to a ROC Curve.

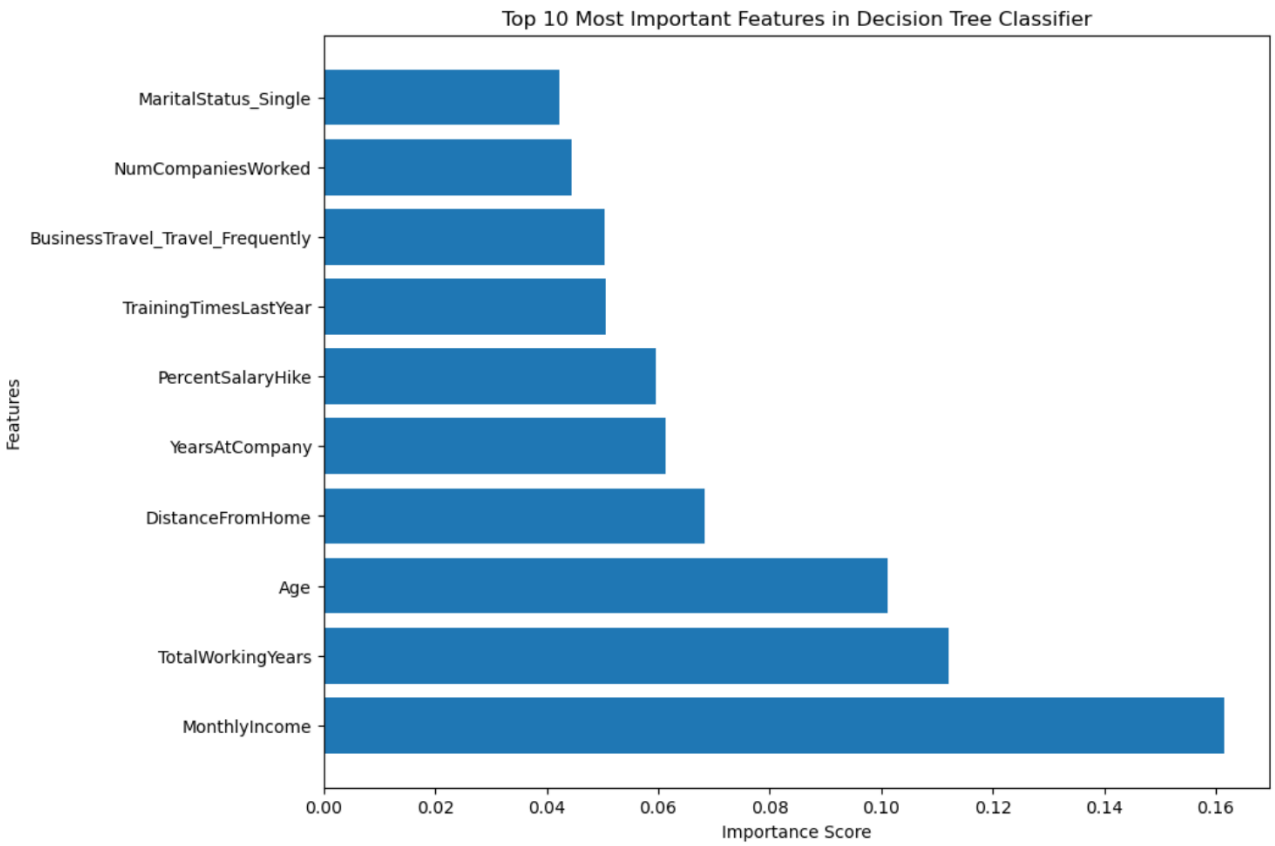


In this graph, the flat line at the top of the Precision-Recall curve indicates that for low recall values, the model achieves almost perfect precision. This means that when the model makes predictions at these thresholds, it is correctly identifying the positive instances (yes to attrition) while making very few false positive predictions. However, as the recall increases above 0.6, the precision starts to drop. This suggests that as the model tries to capture more of the positive instances in the dataset, it also starts to include more false positive predictions. The model becomes less conservative and more inclusive in its predictions, resulting in more false positive errors (Steen, 2020).

Decision Tree Visualization - The below tree is a visualization of the behind-the-scenes reasoning that occurs when the Decision Tree deploys its model. The hyper-tuned Decision Tree has a depth of 10 which results in multiple branches and nodes that clutter the visualization. Therefore, I have trained a model with a depth of 3 to provide an example of the logic a decision tree performs to complete its classification.



Features Importance – The bar graph below shows the ten variables with the highest importance score in the Decision Tree Classifier. This shows that “MonthlyIncome” is by far the greatest factor in attrition. The next closest factors are “Age” and “TotalWorkingYears.” It is not surprising that these variables track closely together as previously mentioned. It is worth noting that these are variables which the company has no control over. The business will have to focus their efforts on variables that the business can impact.



Conclusion

This study utilized dummy variables to create a hyper-tuned Decision Tree Classifier that will predict if an employee will attrite with 93% accuracy. Additionally, this study provided visualizations detailing the model's classification process and success rate at classifying employees that will and will not attrite. Finally, the importance scores were extracted informing the viewer which variables had the largest impact when the model performed its classification algorithm.

Assumptions

This study assumes that all employees weigh factors contributing to attrition in the same regard. However, each employee has outside factors that can contribute to how they feel about each variable. For example, one employee might enjoy travel and find frequent business travel to be a positive aspect of the business. Another employee might find that travel is time-consuming and draining and would prefer a role which allows them to limit travel.

Additionally, this study only applies directly to this company and this dataset. It cannot be assumed that the findings in this study can be applied universally across all fields of business in every location.

Limitations

A challenge with the dataset is that I do not have any information about the company that this data comes from. Not knowing the field of business makes recommendations and implementations difficult.

For example, without additional context about the company and the field of business, it is difficult to know if offering a “work from home” option might be a viable recommendation. It would also be valuable to have a location variable added to the dataset. If there are several offices in multiple locations it would be interesting to determine how the attrition rate varies by location. If all employees work in the same office, it would be interesting to compare the attrition rates of this company to the attrition rates of a similar company in the same location.

Challenges

A personal challenge was working with dummy variables and blending numerical and categorical data. I had a challenging time keeping track of the binary variables and noting what “0” and “1” represented. Most of my previous projects have involved predicting a continuous numerical variable so I had to spend a good amount of time reviewing classification models and their corresponding metrics. I found the confusion matrix to be extremely helpful in overcoming this challenge and understanding the Precision, Recall, F1, and Accuracy scores.

Future Uses/Additional Applications

There are two additional uses to which this study would provide useful insight.

> ***Benchmark*** – Once mitigating efforts have been in place for a length of time, the company can assess if attrition rates have declined and rerun this study with an updated dataset to determine if the cause of attrition has changed. For example, one of the important features for the Decision Tree Classifier used in this study is “DistanceFromHome.” If the company begins to offer remote or hybrid work, does this variable drop out of the top ten most important features in a future study?
>
> ***Compare to Other Businesses*** – It would be interesting to run similar studies on a variety of other businesses to determine if they have similar attrition rates and causes of attrition, it they can train a classifier with a similar accuracy rate, and if they have alternate data points available. I use the broad phrase “a variety of other businesses” because I would be interested to compare the company in this study to a similar company in the same area, a larger/smaller/different field of company in the same area, and companies in other states. It would be interesting to see what variables are important to attrition across all companies and what variables vary depending on location, size of company, field of business, etc.

Recommendations

When reviewing the features with the greatest importance to the Decision Tree Classifier, there are features which the business can impact and features which the business has no control over. It is recommended that the business focuses on the features that it can manage. Variables in the top ten features that the company has control over include: MonthlyIncome, DistanceFromHome, PriceSalaryHike, TrainingTimesLastYear, and BusinessTravel_Travel_Frequently. Recommendations for each of these variables are below:

> ***MonthlyIncome/PriceSalaryHike*** – These variables both pertain to the associate’s salary. It is recommended the company conducts an industry analysis to determine if their compensation structure matches industry standards.

DistanceFromHome – The company cannot dictate where an employee lives, but they can offer options such as remote work, hybrid work, or a satellite office.

TrainingTimesLastYear – Offer an optional mentorship program. This will allow employees overwhelmed with training to opt out and employees seeking additional training and guidance the opportunity to learn more about their job and the industry from a more experienced coworker.

BusinessTravel_Travel_Frequently – Modern technology allows for many meetings to be done virtually via a video conferencing system. Depending on the field of business, it is possible that this will not eliminate all business travel, but it can lessen the amount from “Frequently” to “Rarely.”

Implementation Plan

There are consequences of every business decision, and the above recommendations are no different. Each of these recommendations could have positive effects on preventing attrition, but there could be additional effects such as additional expense and effort on the company’s behalf. Management will have to work with the company’s Finance team to determine if the cost of implementing the above recommendations is less of a burden than the cost of attrition.

To implement the actual classification model, the company can run this on a yearly basis with their current employee dataset and use it to budget accordingly for the following year. As noted at the beginning of this paper, “One estimate places the cost to replace an employee at three to four times the position’s salary” (Wallace, 2023). If the finance team determines that 5 of x position and 3 of y position are expected to attrite in the coming year, they can factor the expense of filling that position into the yearly budget.

Ethical Assessment

There are several ethical concerns with how a company might implement this model. The first concern is the fear of discrimination. If a hiring manager knows that a certain gender or age group attrites more than another, they cannot take this information into consideration when making hiring decisions. A second ethical concern pertains to equity in the workforce. If an employee is about to attrite and the company offers to alter their features that lead to attrition such as business travel or distance from home, those same opportunities should be offered to other employees even if they are not at risk of attrition. For example, if an employee plans to attrite but the company allows them to work remotely to alter their distance to work, the opportunity should then be offered to all employees in that position or department. This model can be beneficial to a company to predict attrition and factors contributing to attrition on a company or even a department level, but any individual offers or opportunities should be made based on an employee's merits and quality of work, not based on their attrition classification.

CITATIONS

Brownlee, J. (2023, October 10). *How to use ROC curves and precision-recall curves for classification in Python*. MachineLearningMastery.com. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Kaggle. (2022, November 3). *Employee attrition*. Kaggle. https://www.kaggle.com/datasets/ajayganga/employee-attrition

Kanstren, T. (2023, August 4). *A look at precision, recall, and F1-score*. Medium. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

Steen, D. (2020, September 20). *Precision-recall curves*. Medium. https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248

Wallace, L. (2023, March 21). *Forbes EQ Brand Voice: Five hidden costs of employee attrition*. Forbes. https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/?sh=482ecd3562f4

APPENDIX

Appendix A – Factors of Attrition

Title: "5 Key Factors Impacting Employee Attrition and Retention"

Publisher: Perceptyx

Published: February 5, 2024

URL: https://blog.perceptyx.com/5-key-factors-impacting-employee-attrition-and-retention

5 Key Factors Impacting Employee Attrition and Retention

Given the rapidly evolving dynamics of the modern workplace, understanding the factors that influence employee attrition and retention is critical. Despite the attention garnered by layoffs, often within high-profile organizations, the broader labor market presents a complex picture of concurrent hiring spurts and strategic workforce adjustments across various sectors.

At Perceptyx, our research and consulting teams have dedicated themselves to collaborating with leading organizations across the globe. Together, we've developed sophisticated employee listening and action strategies that are purpose-built for specific industries and specific talent challenges. Our listening strategies are designed not only to gauge the pulse of the workforce but also to foster an environment where employees feel valued, heard, and motivated to grow alongside their employers.

This article draws on the collective expertise of our team members, who share their insights on the evolving dynamics of attrition and retention. In the sections that follow, we look at some of the strategies that can empower employers to effectively monitor, understand, and influence these critical aspects of their organizational health in 2024 and beyond.

1. The Importance of Feeling Valued

Michael Mian, Ph.D., Principal Consultant: “In customer-focused roles, the feeling of being valued significantly impacts an employee's decision to stay with an organization. This feeling of value extends beyond mere financial compensation. It encompasses factors such as employee empowerment, the opportunities provided for career growth, and the investment an organization makes in the development of its employees.”

“Employees' perceptions about their compensation, work/life balance, and the flexibility of their schedules are crucial elements in their overall job satisfaction. In the modern work environment, where the boundary between professional and personal life is increasingly blurred, the ability to maintain a balance is vital. Employees greatly value flexible scheduling, which directly affects their quality of life and overall satisfaction with their job. It's important for employers to recognize that these factors are not perks but necessities in cultivating a motivated and committed workforce.”

“Another critical aspect is the alignment between job expectations set during the hiring process and the actual role once an employee joins the organization. A significant mismatch between these can lead to early turnover. New hires often enter a role with certain expectations, and if the reality falls short, it can

lead to feelings of disillusionment or even betrayal. Similarly, performance expectations need to be clear and realistic. When employees feel that what is expected of them is either not communicated well or is unattainable, it can lead to frustration and a decline in job satisfaction, which often precedes the decision to leave."

2. Compensation Is Just the Tip of the Iceberg

Crystal Perel, M.A., Senior Consultant: "When I attended the IPMI conference in May, a key theme emerged from discussions with Chief Human Resources Officers (CHROs) across various industries. It was unanimously recognized that feeling valued and recognizing employees are the primary methods to enhance engagement and retention in the workforce."

"In my work with different organizations, I've consistently found that compensation is often cited as the top reason employees choose to leave a job, and usually by a substantial margin. This is an interesting observation because, while it highlights the importance of fair and competitive remuneration, it also tends to mask deeper, more complex issues within the workplace. Compensation is frequently just the tip of the iceberg. Beneath it are more intricate challenges that contribute to an employee's decision to leave an organization."

"These deeper challenges include a heavy workload, the all-too-common issue of burnout, leadership that fails to inspire or support, and a lack of clear pathways for professional development and career progression. These factors collectively create an employee experience that can be unsatisfying and demotivating. When employees face these challenges on a daily basis, the appeal of better compensation from another employer becomes not just about the money, but also about the potential for a more rewarding and balanced professional experience."

"It's important to address these underlying issues to create a more fulfilling and engaging work environment. It's not just about increasing salaries; it's about understanding and improving the overall employee experience. This approach is essential for organizations looking to retain their best talent and foster a productive, committed workforce."

3. The Need for Clarity of Direction, Both Personally and Professionally

Sarah Jorgenson, Senior Consultant: "Two critical factors can significantly influence an employee's decision to stay with an organization: feeling valued and having clear career and growth opportunities."

"One of the key issues I'm observing in many organizations is the lack of clarity in direction. Employees are increasingly finding it difficult to envision a long-term future with their current employers. This isn't just about the immediate role they're in; it's about their overall career trajectory within the organization. When employees can't 'see' themselves growing or advancing in the company, it starts to impact their perception of having a viable career path there. This lack of visibility and uncertainty about the future can be profoundly demotivating and is often a critical factor leading to an employee's decision to leave."

"Our internal research at Perceptyx supports this view. We've found a significant correlation between an employee's perception of their career development opportunities and their long-term commitment to the organization. This research isn't just about numbers and data; it's about understanding the human aspect of organizational life. When employees feel that their growth and career development are

supported and they have a clear sense of direction, they are more likely to feel connected and committed to the organization."

4. The Importance of Culture

Christian Roome, Senior Consultant: "Career opportunities are certainly a significant part of an employee's vision for their future within an organization. The chance for growth, the ability to progress, and the potential for personal and professional development are key drivers of employee retention. However, equally important is the organization's culture and its values. Employees need to feel that they are part of a culture that resonates with their personal values and beliefs. This cultural fit is what often makes the difference between an employee who is merely satisfied and one who is truly engaged and committed."

"In recent times, the traditional structure of organizations has been evolving. One notable trend is the stripping away of managerial levels, a change that challenges the traditional pathways employees envision for their career progression. This flattening of organizational hierarchies can lead to ambiguity and uncertainty about career advancement opportunities. It requires organizations to rethink how they define and present growth opportunities to their employees."

"Additionally, the widespread shift towards working from home has had a profound impact on how employees perceive and experience organizational culture and values. The physical separation from the workplace can lead to a sense of disconnection, making it more challenging for employees to feel a part of the organizational culture. This shift necessitates a reevaluation of how culture is communicated and sustained in a remote or hybrid work environment."

5. Help Employees Envision a Successful Future

Bradley Wilson, Ph.D., Principal Consultant: "The key to retention lies in the anticipation of success, both for individuals and the organization as a whole. Success, of course, is a subjective concept, varying significantly from one person to another. However, four elements play an important role for most employees: achievement, affiliation, affluence, and autonomy."

"Achievement and affiliation are often prioritized by organizations, particularly those that have experienced increased attrition following the implementation of new Return-to-Office (RTO) policies. While these elements are important, many leaders overlook the critical aspect of autonomy. Autonomy allows employees to feel in control of their work and their environment, which is especially significant in the context of recent shifts towards more flexible work arrangements."

"The ability of organizations to enable individuals to anticipate success, work in their areas of strength, and contribute to something larger than themselves is vital for driving retention. This is true across various business sectors and economic conditions. However, the high inflation rates seen over the past two years have added a new layer of complexity. This economic pressure has created a financial incentive for employees to change jobs in an attempt to maintain their purchasing power, which has been significantly impacted since 2022."

"When compensation adjustments fail to keep pace with the rising cost of living, employees effectively experience a reduction in pay. This is exacerbated when companies offer higher pay to new hires for the same roles, leading to a perception among existing employees that loyalty is undervalued and

eroding trust within the organization. In today's economic climate, helping employees manage their compensation and benefits is more important than ever. It's not just about providing a fair wage; it's about ensuring that employees feel their financial well-being is being considered and protected."

10.5 Questions

Question #1: What Department has the highest attrition rate?

Question #2: Job Role has the highest attrition rate?

Question #3: Do certain Job Levels attrite more than others?

Question #4: Does more training time lead to attrition or less training time?

Question #5: Does a higher level of education lead to more or less attrition?

Question #6: Does more travel lead to attrition or less travel?

Question #7: Is there a Monthly Income where attrition plateaus?

Question #8: How much has the attrition in the dataset cost the company?

Question #9: We have assumed that Age and Total Working Years are related. Can this be confirmed?

Question #10: Can this model be run using only variables that the company has control over?

Question #10.5: Will the updated model have the same metrics?