Predicting Employee Attrition

Business Problem of Employee Attrition

Employee attrition, the loss of employees through resignation or termination, is a critical issue for organizations. High attrition rates can lead to increased recruitment and training costs, reduced productivity, and loss of organizational knowledge. Predicting which employees are at risk of leaving can help companies implement strategies to retain valuable talent, thereby minimizing these costs and disruptions. This project will employ dashboards to better understand attrition within a company. It will also examine machine learning models to predict attrition rates based on existing employee data. Questions this project will seek to answer include the following:

What does attrition look-like within the company now?

What factors appear to influence attrition based on the employee data?

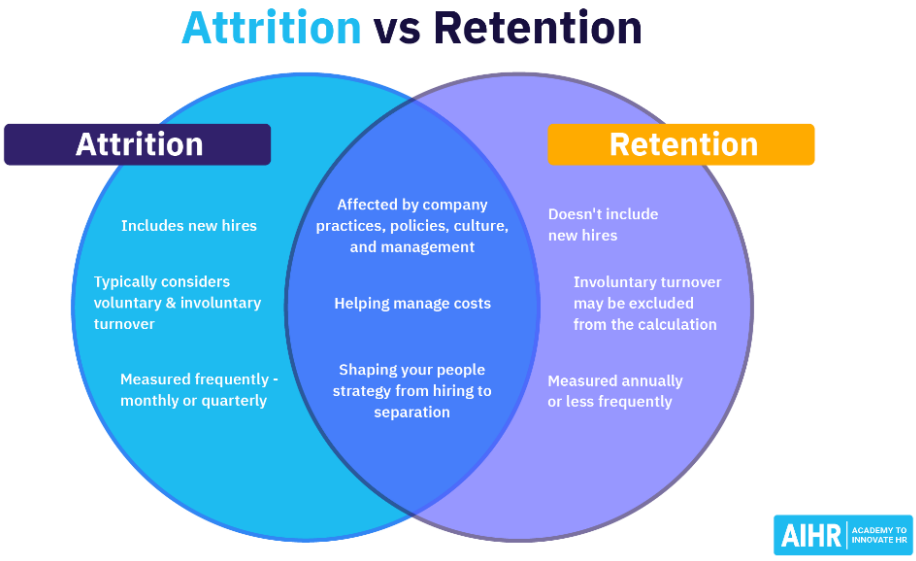
Can a machine learning model accurately predict attrition?

What modeling approach appears to work the best?

Background/History

Employee attrition has been a persistent issue in various industries, exacerbated by factors such as job dissatisfaction, better opportunities elsewhere, poor management, and lack of career development. As opposed to retention, attrition looks at all employee gains and losses whether

employees leave voluntarily or not. The figure below outlines the differences and convergence of these similar concepts.



While the direct effects of employee loss are intuitive, the impact of employee-loss also has important ripple effects for those who stay with an organization. One survey of "burnt-out' employees found 41% of them reported the source as employee shortages, other surveys suggest the influence of employee shortages on burn-out is even higher (Assemble, 2023).

Historically, companies have relied on traditional methods like exit interviews to understand attrition, but these are reactive and offer limited foresight (Why an Exit Interview Won't Help You Reduce Attrition, 2022). Machine learning methods allow for a more proactive approach, leveraging employee data to predict attrition and address issues before they lead to resignations.

Data Explanation

To examine the questions of this project a labeled dataset was sourced from kaggle.com. It included over 16k rows of employee data with 34 column features including our target feature,

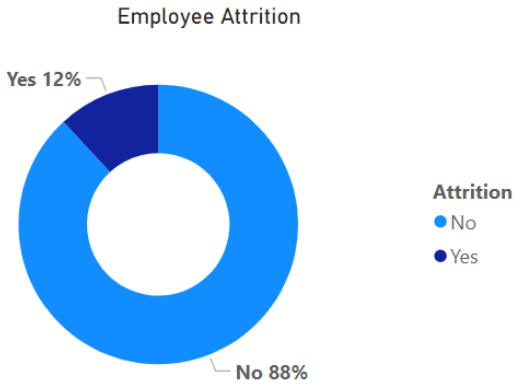
“attrition”. The dataset included numeric, categorical, and binary datatypes representing employee survey results for measures of things like satisfaction as well as demographic data for measures of education, gender, age, and more. These features help provide a comprehensive view of factors that could influence an employee's decision to leave the company. A complete list of features and their definitions is included in Table 1 of the appendix.

Methods

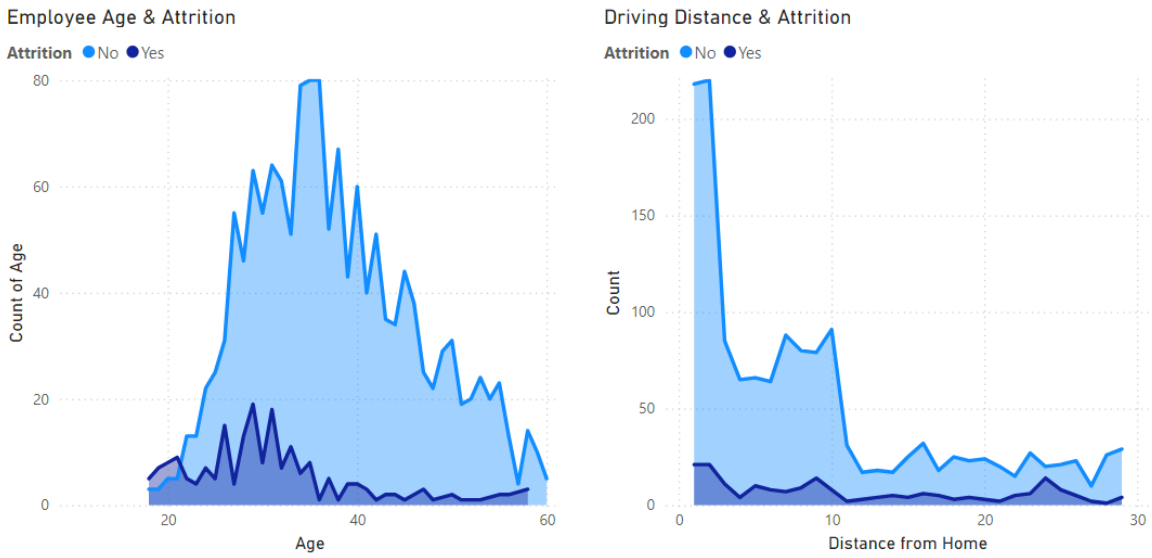
Analysis began with a view of value counts, data types, and visualization using Power BI and seaborn. Distributions and bivariate relationships between features and the target variable, attrition were plotted. For preprocessing, categorical variables were dummy coded to make a format suitable for Recursive Feature Elimination (RFE) with logistic regression and modeling. Data was first split into an 80:20 ratio for training and testing. The RFE was performed on full range of n test set features and plotted against accuracy results to identify the optimal number of features to include for modeling. During the model selection and tuning phase, GridSearchCV was implemented to tune hyperparameters for three models: Decision Tree Classifier, Gradient Boosting Classifier, and Multi-Layer Perceptron (MLP). Finally, the models were evaluated using a combination using accuracy scores, classification report, and confusion matrices.

Analysis

Initial analysis showed that the target class for the training data was highly imbalanced with only 12% of the data reflecting employees that had left the company. While this low percentage is good for a company, it would likely provide a challenge for prediction accuracy.

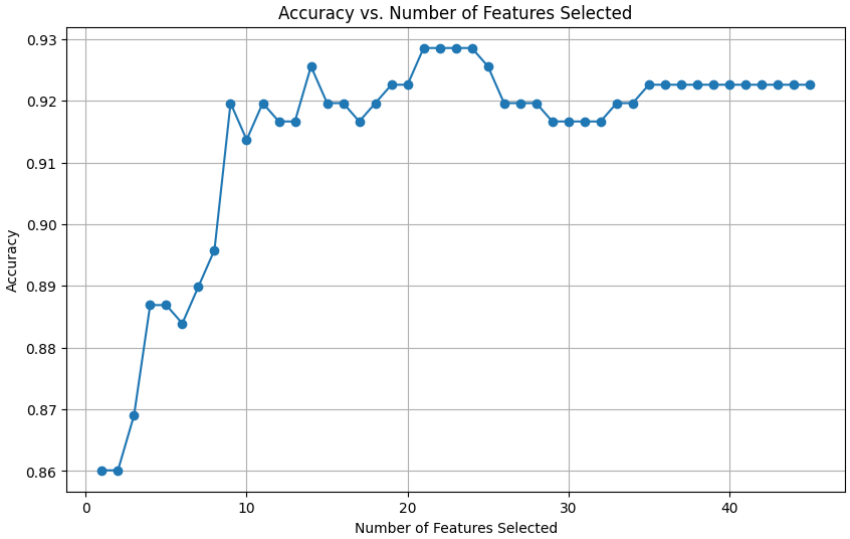


Most variables showed, if any, subtle relationships to the target. One of note included employee age, where the youngest employees had higher a proportion of attrition. The distributions for age were layered across the target, yes/no, for attrition. The attrition-positive distribution had lower minimum and maximum. Driving distance was another feature of interest. Visualization showed that the largest pool of employees that stayed with the company drove 10 miles or less to get to work, with the highest count of those staying driving less than 5 miles.



With the additional variables created after dummy coding, the RFE was useful in subsetting a smaller, effective group of features to use for prediction. A plot of the accuracy across n

variables for inclusion found that 21 features was the optimal number. The exact list of the features chosen for modeling can be found in table 2 of the appendix.



After features were selected, null accuracy was calculated on the test data set at 86% which provided a comparison for model evaluation. The first model trained and tested was a decision tree classifier. A gridsearch was used to identify ideal hyperparamters and the resulting accuracy score was 89 %. Again, a gridsearch was used to tune and train a gradient boosting classifier and a multilayer perceptron (MLP) model. The resulting accuracies were 90%, and 92% respectively. While there did not appear to be a dramatic improvement in the overall accuracy across models, classification reports revealed a large improvement in the F1 scores related to the minority target class (Attrition-Yes) from .46 to .58 to .68. In fact, the MLP model had the highest scores for precision, recall and f1 scores for the minority target class compared to the other. The classification report is illustrated below and reports for all three models can be found in tables 3,4,and 5 in the appendix.

```
Accuracy: 0.9226190476190477
Confusion Matrix:
[[283   6]
 [ 20  27]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.98      0.96       289
           1       0.82      0.57      0.68        47

    accuracy                           0.92       336
   macro avg       0.88      0.78      0.82       336
weighted avg       0.92      0.92      0.92       336
```

Conclusion

All three models struggled to accurately predict the minority target class for attrition. With the first 2 of 3 models mis-predicting the majority of employees that left the company. The MLP model with a single layer of 100 neurons, however, was able to predict more of this class correctly than incorrectly. While a score of 92% accuracy appears very high, and it is higher than the 87% null accuracy that was calculated, the accuracy of predicting which employees left is still low. When examining a confusion matrix on the best model, the ratio of correct to incorrect is merely 27 to 20, a less than impressive result. However, it is important to acknowledge that reasons for leaving a company are complex, so a high performing accuracy in this context may look different than in other contexts.

Assumptions

The major assumption of this project is that the data is representative of the entire employee population. Another general assumption of attrition prediction is that historical data is an accurate predictor of future attrition. In a similar vein, there is an assumption that the relationships between features and attribution are stable over time, which is unlikely.

Limitations & Challenges

While this data had no missing values and appeared to have no incomplete data, the imbalance in the target class makes it more difficult for a model to be trained to predict attrition. Another important limitation is that using MLP model cannot provide direct insights into how features impact attrition. It was hoped that a decision tree classifier would suffice for modeling, as it can provide a descriptive outline of how features are handled. In this project accuracy, in effect, has been traded for interpretability.

Future Uses/Additional Applications

As a company grows deploying real-time prediction can be advantageous to monitor and address issues as they come. Over time, models can be adjusted, updated, and even customized to subgroups of employees/departments/ or roles as a company sees fit. As more data sources become available, modeling will only have more potential to improve its predictions. Meanwhile, insights gained through dashboarding and monitoring of employee metrics can help this company understand what inspires employees to stay vs go.

Implementation Plan

- Pilot Testing- applying this model can initially be limited to a subset of the company to help validate findings regarding measures that inspire employees to leave vs stay. As the model can demonstrate validity, it can be further integrated.

- Integration - At this stage the MLP prediction model can be integrated into the current HR systems to highlight risk features related to attrition. Fold insights into hiring and exit interviews.
- Training - Conduct training sessions for HR personnel to interpret model outputs and take necessary actions.
- Feedback Loop Establish a feedback mechanism to continuously improve the model based on new data and outcomes. Again, data gathered from surveys, hiring and exit interviews can extend and improve modeling.

Ethical Assessment

Because there is a range of protected class-related data often in employment data, it is preferable features related to these classes are not included in the analysis (i.e., gender, race, ethnicity). Proxies for these measures should also be avoided, such as zip code. These features were not included to ensure the model does not disproportionately affect any group of employees. Regular audits and fairness metrics should be implemented to prevent unethical use of employee data because the potential is there of course.

References

Assemble. (2023, March 21). Five hidden costs of employee attrition. *Forbes*. https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/?sh=7812695062f4

Dataset: *Employee attrition for healthcare*. (2023, February 15). Kaggle. https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86. https://doi.org/10.3390/computers9040086

Why an exit interview won't help you reduce attrition. (2022, August 8). Workday Blog. https://blog.workday.com/en-us/2021/exit-interview.html

Appendix

Table 1. All dataset column names

Variable Name
EmployeeID
Age
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
EducationField
EmployeeCount
EnvironmentSatisfaction
Gender
HourlyRate
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked
Over18
OverTime
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
Shift
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Table 2 – RFE selected features for analysis

Feature	Selected
Age	True
DistanceFromHome	True
EnvironmentSatisfaction	True
JobInvolvement	True
JobSatisfaction	True
MonthlyIncome	True
NumCompaniesWorked	True
TotalWorkingYears	True
WorkLifeBalance	True
YearsAtCompany	True
YearsInCurrentRole	True
YearsSinceLastPromotion	True
BusinessTravel_Travel_Frequently	True
Department_Cardiology	True
EducationField_Human Resources	True
JobRole_Admin	True
JobRole_Administrative	True
JobRole_Therapist	True
MaritalStatus_Divorced	True
MaritalStatus_Single	True
OverTime_Yes	True

Table 3 – Evaluation Measures for Decision Tree

```
Accuracy: 0.8898809523809523
Confusion Matrix:
[[283   6]
 [ 31  16]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       289
           1       0.73      0.34      0.46        47

    accuracy                           0.89       336
   macro avg       0.81      0.66      0.70       336
weighted avg       0.88      0.89      0.87       336
```

Table 4 – Evaluation Measures for Gradient Boosting Classifier

```
Accuracy: 0.9077380952380952
Confusion Matrix:
[[284   5]
 [ 26  21]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95       289
           1       0.81      0.45      0.58        47

    accuracy                           0.91       336
   macro avg       0.86      0.71      0.76       336
weighted avg       0.90      0.91      0.90       336
```

Table 5 – Evaluation Measures for MLP Classifier

```
Accuracy: 0.9226190476190477
Confusion Matrix:
[[283   6]
 [ 20  27]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.98      0.96       289
           1       0.82      0.57      0.68        47

    accuracy                           0.92       336
   macro avg       0.88      0.78      0.82       336
weighted avg       0.92      0.92      0.92       336
```

10 Questions

1. What is precision?

It is the ratio of true positive predictions to the total number of positive predictions (both true positives and false positives). In other words, it measures how many of the predicted positive cases are actually positive. The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

2. What is recall?

Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify all positive instances. It is the ratio of true positive predictions to the total number of actual positive instances (true positives and false negatives). The formula for recall is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3. What is an f1 score?

- *The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is especially useful when you need to take both false positives and false negatives into account. The formula for the F1 score is:*

$$F1\,\text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- *An F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.*

4. What is a gridsearch?

GridSearch is a technique used to tune hyperparameters of a machine learning model. It performs an exhaustive search over a specified parameter grid, evaluating the performance of the model for each combination of parameters using cross-validation. The goal is to find the best parameter set that maximizes the model's performance.

5. What were the best parameters chosen for the MLP classifier?

In your project, the best parameters chosen for the MLP classifier through GridSearch might include settings like the number of hidden layers, the number of neurons in each layer, the

learning rate, activation function, and the solver type. Specific values depend on the results of your GridSearch process.

6. Why was a gradient boosting classifier chosen?

- *It is highly effective for classification tasks, especially with complex datasets.*
- *It builds an ensemble of weak learners (usually decision trees) that are optimized to correct the errors of the previous learners.*
- *Gradient Boosting is known for its high performance and ability to handle both numeric and categorical features well.*
- *It often provides better accuracy than single models because it combines the strengths of multiple models.*

7. What exactly is null accuracy?

Null accuracy is the accuracy that can be achieved by always predicting the most frequent class in the dataset. It serves as a baseline to compare the performance of your model. If your model's accuracy is not significantly better than the null accuracy, it indicates that the model may not be capturing meaningful patterns in the data.

8. Why does imbalance in the dataset target classes matter?

Imbalance in the target classes means that some classes are underrepresented compared to others. This imbalance can lead to biased models that favor the majority class, resulting in poor performance on the minority class. It can also affect the accuracy, precision, recall, and F1 scores, making them misleading if not properly addressed.

9. What can the company do to address the class imbalance in the data?

- *Use techniques like oversampling the minority class or undersampling the majority class.*
- *Implement more sophisticated methods like SMOTE (Synthetic Minority Over-sampling Technique).*
- *Adjust the class weights in the model to give more importance to the minority class.*
- *Collect more data to ensure a balanced representation of classes.*

10. What can the company do to improve or supplement data used for this analysis?

- The company can gather more data points to enhance the dataset's size and diversity.
- Incorporate additional relevant features that could improve the model's predictive power.
- Ensure data quality by cleaning and preprocessing to remove inconsistencies or errors.
- Use external data sources if available and relevant to the prediction task.
- Continuously monitor and update the dataset to reflect any changes in the underlying processes or factors influencing employee attrition.