**Predicting Employee Attrition: A Data Science Approach**

**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. Question this project will seek to answer include the following:  
What does attrition look-like within the company now?

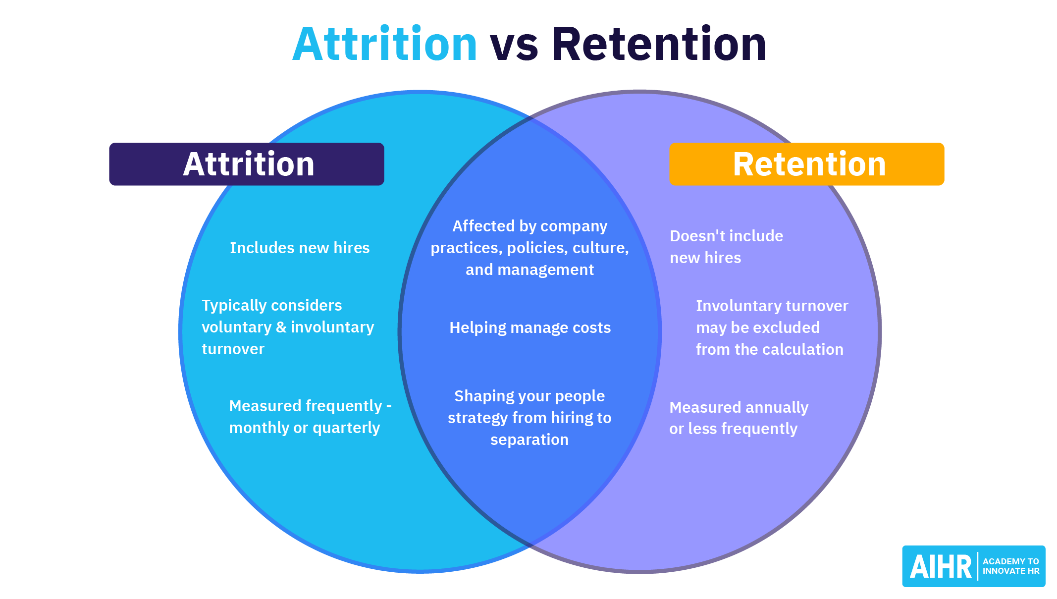
What factors appear to influence attrition based on the employee data?

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). The advent of data science allows 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 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 data exploration and visualization using Power BI to examine distributions and bivariate relationships between features and the target variable, attrition. For preprocessing, categorical variables were dummy coded make a format suitable for modeling and employed Recursive Feature Elimination (RFE) with logistic regression. The RFE was performed on full range of n features and plotted against accuracy results to identify optimal, minimal amount of features to include. Data was then split into an 80:20 ratio for training and testing. During the model selection and tuning phase, we implemented GridSearchCV to tune hyperparameters for three models: Decision Tree Classifier, Gradient Boosting Classifier, and Multi-Layer Perceptron (MLP). Finally, we evaluated the models using accuracy score, classification report, and confusion matrices.

Data Exploration and Visualization: Utilized Power BI to examine distributions and bivariate relationships between features and the target variable (attrition).

2. Preprocessing:

- Dummy coded categorical variables to convert them into a numerical format suitable for modeling.

- Used Recursive Feature Elimination (RFE) with logistic regression to identify the top 20 features.

3. Train/Test Split:Divided the data into an 80:20 ratio for training and testing.

4. \*\*Model Selection and Tuning:

- Implemented GridSearchCV to tune hyperparameters for three models: Decision Tree Classifier, Gradient Boosting Classifier, and Multi-Layer Perceptron (MLP).

5. \*\*Evaluation:\*\* Assessed the models using accuracy score, classification report, and confusion matrices.

## Analysis

- \*\*Decision Tree Classifier:\*\* Provided a clear and interpretable model but tended to overfit on the training data.

- \*\*Gradient Boosting Classifier:\*\* Achieved high accuracy and balanced performance, excelling in capturing complex relationships in the data.

- \*\*MLP Model:\*\* Showed potential with its ability to model non-linear relationships but required significant computational resources and fine-tuning.

## Conclusion

The Gradient Boosting Classifier emerged as the best model for predicting employee attrition, offering a good balance between accuracy and interpretability. By identifying at-risk employees, companies can proactively engage in retention strategies, thereby reducing attrition rates and associated costs.

## Assumptions

- The data is representative of the entire employee population.

- Historical data is an accurate predictor of future attrition.

- Relationships between features and attrition are stable over time.

## Limitations & Challenges

- \*\*Data Quality:\*\* Incomplete or inaccurate data could affect model performance.

- \*\*Feature Selection:\*\* Important variables might have been overlooked, impacting model accuracy.

- \*\*Model Complexity:\*\* More complex models, while accurate, can be difficult to interpret and implement.

## Future Uses/Additional Applications

- \*\*Real-time Prediction:\*\* Implementing real-time attrition prediction systems to monitor and address issues continuously.

- \*\*Cross-industry Applications:\*\* Applying similar models in different industries to address sector-specific attrition issues.

- \*\*Enhanced Features:\*\* Incorporating additional data sources like employee engagement surveys, social network analysis, and external labor market trends.

## Implementation Plan

1. \*\*Pilot Testing:\*\* Start with a small, manageable segment of the company to validate the model.

2. \*\*Integration:\*\* Embed the prediction model into existing HR systems for real-time monitoring.

3. \*\*Training:\*\* Conduct training sessions for HR personnel to interpret model outputs and take necessary actions.

4. \*\*Feedback Loop:\*\* Establish a feedback mechanism to continuously improve the model based on new data and outcomes.

## Ethical Assessment

- \*\*Bias and Fairness:\*\* Ensure the model does not disproportionately affect any group of employees. Regular audits and fairness metrics should be implemented.

- \*\*Transparency:\*\* Maintain transparency about how predictions are made and used in decision-making.

- \*\*Privacy:\*\* Safeguard employee data, ensuring compliance with data protection regulations and maintaining confidentiality.

---

This summary provides a comprehensive overview of your project, covering all the necessary aspects from problem identification to ethical considerations.

**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**

**A computer screen shot of a list of words

Description automatically generated**