Forecasting Employee Attrition with Advanced Machine Learning Models

Introduction

Data-driven journey to develop a predictive model using historical data.

This model aims to forecast employee attrition by analyzing a rich dataset encompassing employee demographics, job-related details, and more.

Our presentation encapsulates the analytics process, findings, and model performances that underpin informed decisions to mitigate attrition and enhance workforce stability.

Numerical Variables

1. Age
2. MonthlyIncome
3. Bonus
4. DistanceFromHome
5. TrainingTimeLastYear
6. YearsAtCompany
7. YearsSinceLastPromotion

Categorical Variables

1. BusinessTravel
2. JobSatisfaction
3. Department
4. Education
5. EducationField
6. EnvSatisfaction
7. Gender
8. JobRole
9. MaritalStatus
10. PerformanceRating
11. OverTime

Analytics Process

Data Exploration: We began by thoroughly examining the dataset, analyzing various factors such as age, income, job satisfaction, and more.

Variable Investigation: We delved into each variable's impact on attrition, using statistical tests like t-tests and chi-squared tests.

Machine Learning Models: We developed predictive models such as Logistic Regression and Naive Bayes to forecast attrition based on data patterns.

Insightful Findings: Our analysis aimed to uncover key insights for retaining talent and reducing attrition.

Data-Driven Decisions: The results from our analytics process guided us in making informed decisions to address employee attrition.

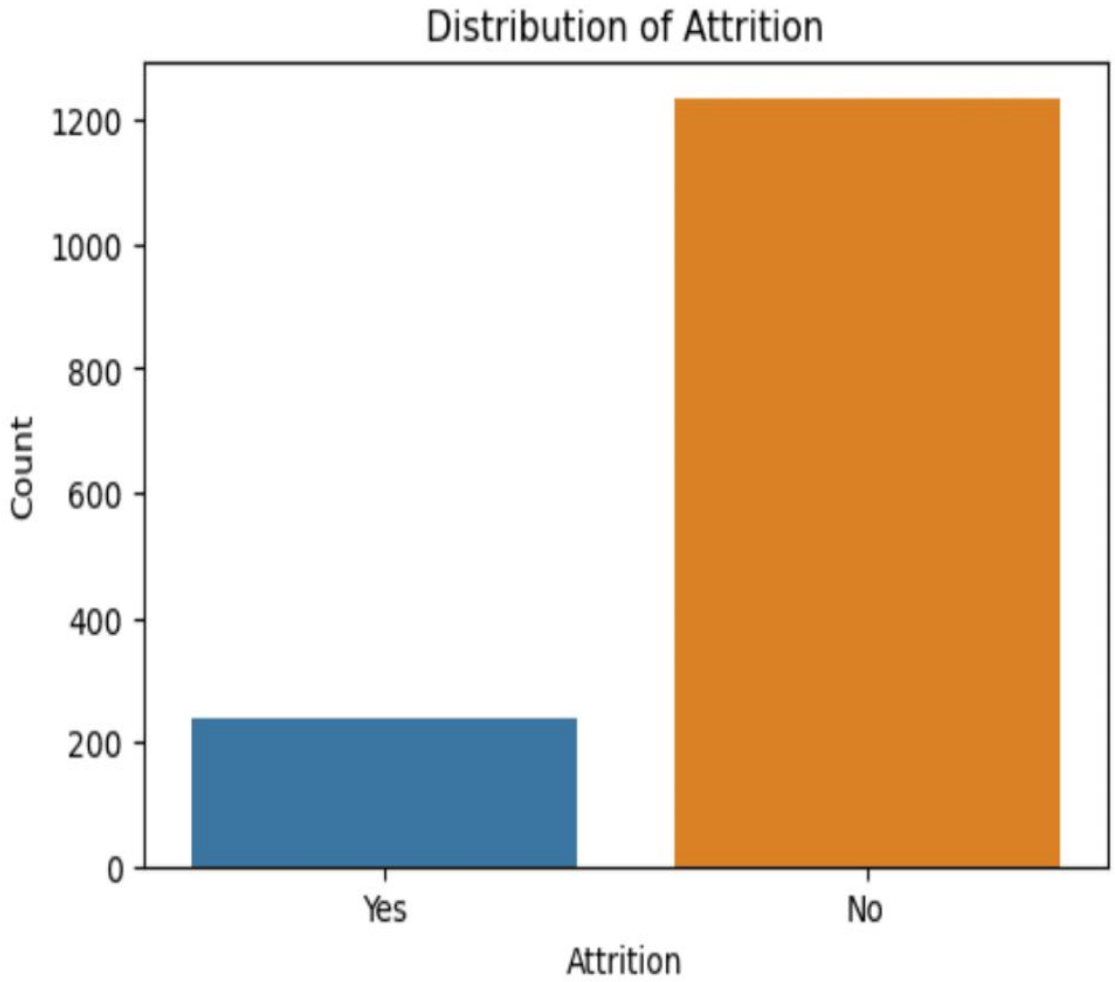
Univariate Analysis
Company's Attrition Rate

Total Employees = 1470
Employees who left the company = 237
Employees who stayed = 1233

Attrition Rate = 16.12%

With an attrition rate of 16.12%, the company is experiencing a turnover that is slightly higher than what is commonly observed (10-15%).

While it's not alarmingly high, it's worth assessing the reasons behind the attrition and considering strategies to mitigate it further, especially if the company aims to achieve a more stable and long-term workforce.




```
No     1233
Yes     237
Name: Attrition, dtype: int64
```

Factors Affecting Attrition

Based on the analysis of numerical and categorical variables, we have identified a total of 14 variables that show significant relationships with employee attrition.

- Among these, the top 5 most significant variables affecting attrition are Bonus, MonthlyIncome, Age, YearsAtCompany, and OverTime.

We have also identified that PerformanceRating, Education, Gender, and YearsSinceLastPromotion do not significantly affect employees attrition.

- While PerformanceRating does not directly affects attrition; we suspect that it might have a relationship with employees' level of Job Satisfaction and Environment Satisfaction (feeling unmotivated, unfulfilled)
- With that being said, this is a crucial section to pay attention higher performance employees benefit the company; and it might reduce the dissatisfaction.
- Though this is an analysis solely based on intuition and domain knowledge

ML Models
Logistic Regression vs Naive Bayes

	precision	recall	f1-score	support
No	0.90	0.69	0.78	199
Yes	0.27	0.61	0.38	38
accuracy			0.68	237
macro avg	0.59	0.65	0.58	237
weighted avg	0.80	0.68	0.72	237

Naive Bayes Model:
- Performs reasonably well in predicting employees leaving ("Yes") with higher recall (0.61) compared to Logistic Regression (0.13).
- Captures more actual "Yes" cases.
- Precision for "Yes" class is relatively lower at 0.27, indicating not all "Yes" predictions are accurate.
- Balanced by higher precision (0.90) and lower recall (0.69) for the "No" class.

	NO	YES
NO	TN 138	FN 15
YES	FP 61	TP 23

	precision	recall	f1-score	support
No	0.86	0.99	0.92	199
Yes	0.71	0.13	0.22	38
accuracy			0.85	237
macro avg	0.79	0.56	0.57	237
weighted avg	0.83	0.85	0.81	237

Logistic Regression Model:
- Excels in predicting employees not leaving ("No").
- Achieves impressive precision of 0.86 for "No" class, indicating highly accurate predictions.
- Exceptional recall for "No" class at 0.99, capturing almost all actual "No" cases.
- Recall for "Yes" class is relatively low at 0.13, missing a significant portion of actual "Yes" cases.

	NO	YES
NO	TN 197	FN 33
YES	FP 2	TP 5

Conclusion:

Focusing on minimizing false negatives (missing employees leaving):
- Logistic Regression model seems more suitable due to high recall for "No" class.
- Despite lower recall for "Yes," higher precision ensures more reliable positive predictions for employees leaving.

In scenarios where missing employees intending to leave is a critical concern (common in attrition prediction), Logistic Regression's high recall for "No" class makes it a strong choice.

Conclusion

Advanced machine learning models can help organizations forecast employee attrition with greater accuracy and develop targeted interventions to retain employees. While there are challenges associated with implementing these models, organizations can overcome them by addressing issues such as data quality and privacy concerns. By using these models, organizations can reduce turnover rates and increase productivity.

Thanks!

Do you have any questions?
radin.amr@gmail.com

Radify Analytica