IBM Attrition Analysis by Machine Learning Methods

Advance Machine Learning Final Report

109306094 Yen Chu, Chen

Table Of Content

I. Introduction p. 3

II. Methodology p. 3

III. Modeling (ML) and Model Comparison p. 5

IV. Results p. 6

V. Conclusion and Business Insights p. 9

VI. Appendix: Plot For Employee Attrition Profile p. 11

VII. Reference p. 14

I. Introduction

IBM is a renowned American multinational corporation with operations spanning approximately 170 countries. Its core business verticals include computing, software, and hardware. In service-providing organizations like IBM, where trained and experienced personnel are invaluable assets, employee attrition poses a significant risk. Understanding the factors that influence employee attrition is crucial for developing effective retention strategies. This report aims to leverage machine learning techniques to identify and analyze these factors, providing insights that can help mitigate attrition and enhance workforce stability at IBM.

II. Methodology

Data Overview

In this report, I utilized a dataset on Kaggle called "IBM Attrition Dataset". It provides an observational view of human resources specialist, including 13 columns and 1470 rows. The description and original data type of the dataset is presented in Table 1.

Table 1. Column Explanation

Column Name	Data Type	Explanation
Age	Numerical	Age of Employee
Attrition	Boolean	Employee attrition status
Department	String	3 categories. Research & Development/ Sales/ HR
DistanceFromHome	Numerical	The distance between home/company
Education	Categorical	1-Below College; 2- College; 3-Bachelor; 4-Master; 5-Doctor;
EducationField	String	6 categories Life Sciences/ Medical/ Marketing/ Technical Degree/ Other/ Human Resources
Environment Satisfaction	Categorical	1-Low; 2-Medium; 3-High; 4-Very High
JobSatisfaction	Categorical	1-Low; 2-Medium; 3-High; 4-Very High

MaritalStatus	String	3 categories. Married/ Single/ Divorced
MonthlyIncome	Numerical	Income per month
NumCompaniesWorked	Numerical	Number of companies worked prior to IBM
WorkLifeBalance	Categorical	1-Bad; 2-Good; 3-Better; 4-Best;
YearsAtCompany	Numerical	Current years of service in IBM

Checking NA value

Before exploratory data analysis, we need to make sure our dataset is clean and balance. In Figure 1, we could see that there aren't any NA values in the dataset.

```
NA值檢查結果：
Age                       0
Attrition                 0
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EnvironmentSatisfaction   0
JobSatisfaction           0
MaritalStatus             0
MonthlyIncome             0
NumCompaniesWorked        0
WorkLifeBalance           0
YearsAtCompany            0
dtype: int64
```

Figure 1. NA Check

Checking Outliers

As for outliers, we could use box plot to check on our numerical columns. It is clear that column "Monthly Income" encountered outlier issues. Therefore, we use statistical method, deleting data which locate higher/ lower 1.5 IQR. The changing of box plot before and after is presented in figure 2.

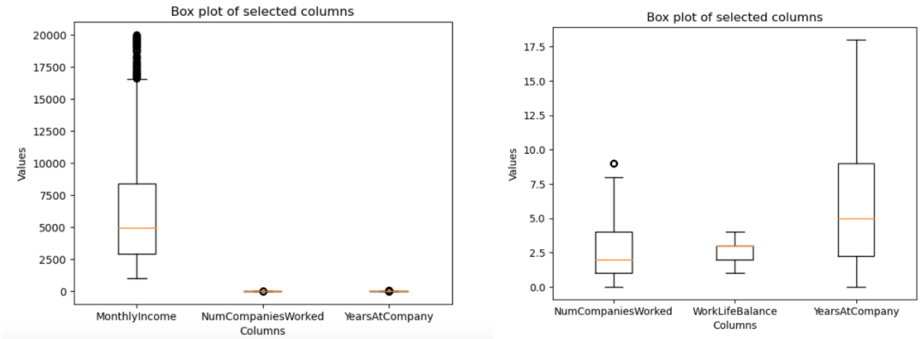


Figure 2. Box plot before and after cleaning outliers

Transforming Categorical Variables and Standardization

Categorical variables represent data types that can be divided into groups or categories. These variables often need to be transformed into a numerical format that can be interpreted by machine learning algorithms. In this study, categorical variables such as "Department," "EducationField," and "MaritalStatus" were transformed using a technique known as one-hot encoding. This process converts each category into a new binary column, allowing the model to interpret categorical data as numerical inputs without implying any ordinal relationship among categories.

As for numerical variables, we use standardization to ensure that each feature contributes equally to the model's learning process. In this study, numerical variables including "Age," "DistanceFromHome," "MonthlyIncome," "NumCompaniesWorked," "WorkLifeBalance," and "YearsAtCompany" were standardized. This transformation helps in achieving a more stable and faster convergence during model training.

Balancing Data

Before proceeding with modeling, it is essential to define our predictor variables (X) and the target variable (Y). This research focuses on a binary classification problem. Upon examining the dataset, we found that the target variable "Attrition" in the training set is imbalanced, with 784 instances of "No" and 172 instances of "Yes". To address this imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which balanced the dataset to 784 instances of both "No" and "Yes".

III. Modeling (ML) and Model Comparison

I had experimented six models, including logistic regression, KNN (K Nearest Neighbor), Random Forest, Adaboost, VotingClassifier which combining both KNN and decision tree and deep learning model ANN. All models are using 80% of data as training set, and 20% of data as testing set. To address the class imbalance issue within the dataset, the Synthetic Minority Over-sampling

Technique (SMOTE) was applied to the training set. However, to preserve the original data distribution, SMOTE was not applied to the testing set.

The final evaluation metrics on testing set to all models are presented in table2.

Table 2. Model Comparison

	Acuracy	**Recall**	**precision**	**F1**
logistic regression	0.67	**0.69**	0.24	0.36
KNN	0.68	**0.67**	0.24	0.36
Random Forest	**0.86**	0.30	**0.46**	0.36
Adaboost	**0.84**	0.42	**0.40**	**0.41**
VotingClassifier	0.77	0.58	0.31	**0.41**
ANN	0.80	0.4	0.31	**0.42**

IV. Results

If we want to choose the best model for this question, we could use ROC curves to help us to make the selection.

There are numerous metrics to help us to determine the best model. Although Random Forest got the highest accuracy and precision scores among others, it got lowest score in Recall and f1.

In Figure3, we could see that read line keeps being the top of all other lines. Furthermore, it also gains the highest AUC scores among all other models. Therefore, I decided to take Adaboost as our best model. If the goal is to find a balance between precision and recall, AdaBoost is good choices, as they both have high F1 scores and AUC values. As for high accuracy, I'll also observe the feature importance of Random Forest.

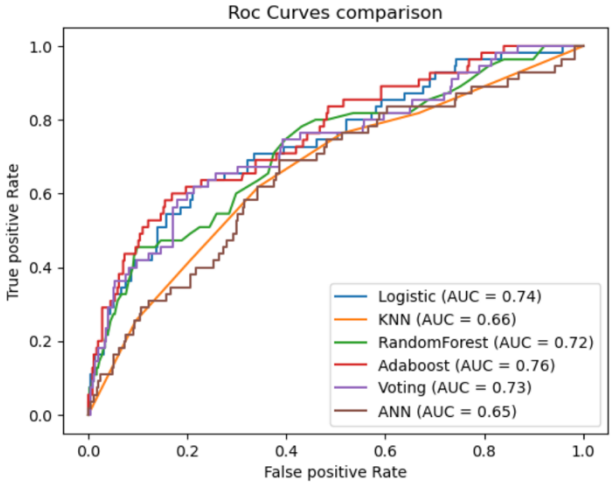


Figure 3. ROC Curves for all models

Upon a detailed examination of our top-performing AdaBoost model and our highest accuracy model, the Random Forest, we can utilize feature importance analysis to gain insights into the significance of various features within our dataset.

In the AdaBoost model (Figure 4), it is evident that the job satisfaction is the most critical factor. Additionally, environment satisdaction, monthly income, education, and major in life science provide substantial insights when addressing classification tasks. For the Random Forest model (Figure 5), years at company, age, marital status and distance from home emerge as additional important features.

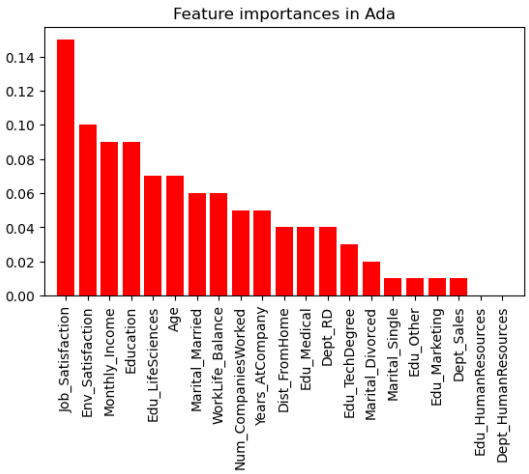


Figure 4. Feature Importance of Adaboost

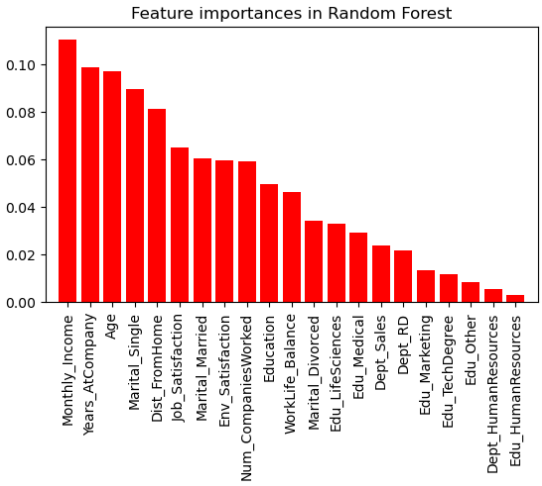
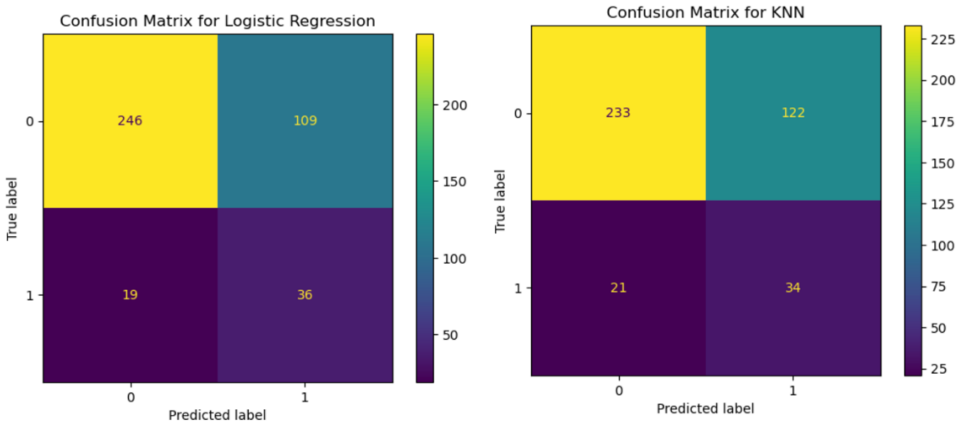


Figure 5. Feature Importance of Random Forest

Findings in modeling

However, we did not apply SMOTE to the testing set. The rationale behind this decision was to preserve the original distribution of the data in the testing set, ensuring that our model evaluation reflects real-world performance. The impact of data imbalance and the effectiveness of our methods can be observed in the confusion matrix. Specifically, without addressing data imbalance, the confusion matrix (Figure 6) often shows a high number of true negatives (majority class correctly identified) and low numbers of true positives (minority class correctly identified). This skewed distribution leads to **low precision** and **low recall**, as a significant portion of the predicted positives are actually false positives.



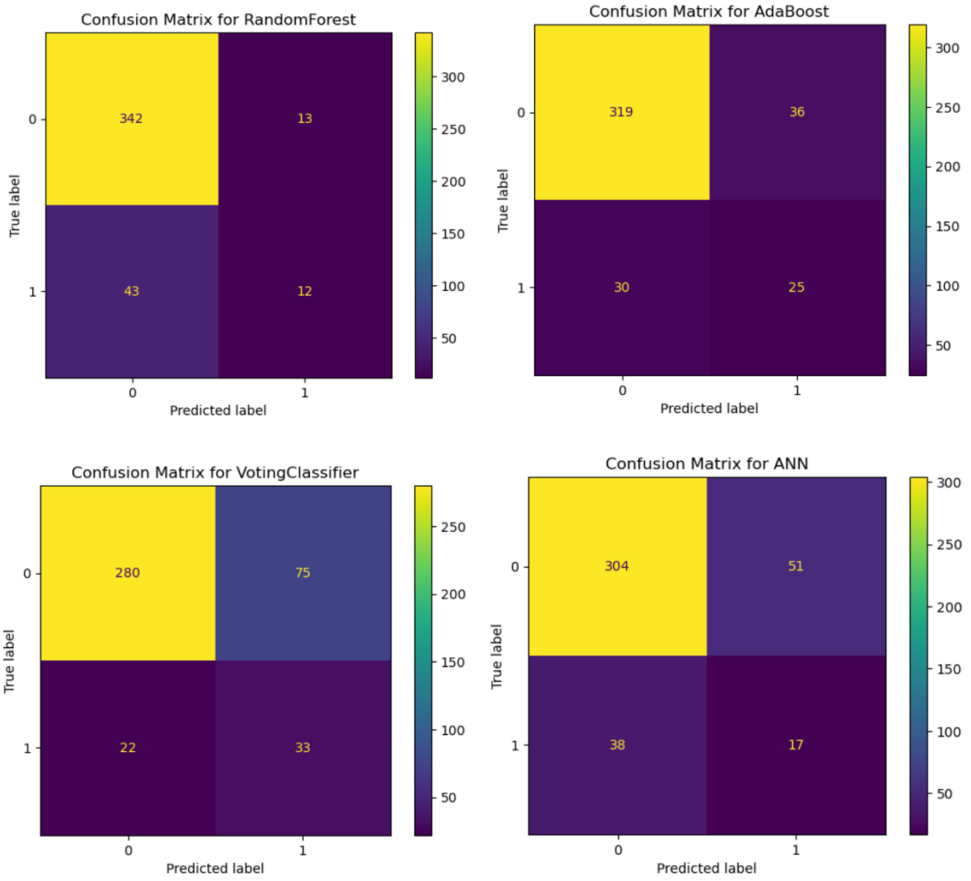


Figure 6. Confusion matrix for all models

V. Conclusion and Business Insights

Employee Attrition Profile

By identifying top 10 feature importance feature in Adaboost and random forest, these key features are indicators which influence employee to quit jobs. IBM's human resources specialists can implement targeted interventions to retain valuable employees through setting these standards in their database. In table3 and figure 7-15 outlines a detailed employee portfolio based on various alarm and alert levels for key features that may signal an employee's intent to leave the company.

Table 3. Employee Profile for Attrition Risk made by Feature Importance

Feature Name	**Attrition level**
Job Satisfaction	Alarm level: 1

	Alert level: 2&3
Environment Satisfaction	Alarm level: 1 Alert level: 2
Monthly Income	Alarm level: below median Alert level: above median
Education	Alarm level:1 Alert level: 3
Major In Life Science	Not majoring life science
Years At Company	Alarm level: below median Alert level: above median
Age	Alarm level: Freshman (Age<30)
Marital Status	Alarm level: Single Alert level: Married
Distance From Home	Alarm level: Distance over 7km Alert level: Distance over 14km

Business Recommendations

To address and mitigate the risk of employee attrition identified in the above portfolio, the following strategic recommendations are proposed:

1. **Conduct In-depth Interviews with High-Risk Employees**
   - Proactively engage with employees identified through alarm and alert levels. These interviews should aim to uncover underlying issues, provide support, and collaboratively develop personalized action plans to enhance their job satisfaction. For instance, consider scheduling interviews with new hires who are single and did not major in life sciences.
2. **Implement Customized Welfare Programs**
   - Develop and offer tailored welfare initiatives that address specific needs of at-risk employees. This could include flexible working hours, enhanced health benefits, shuttle bus to work, shorten transportation time, career development opportunities for new hires, and other personalized incentives that align with their individual circumstances and preferences.
3. **Develop a Comprehensive HR Monitoring Dashboard**
   - Create an advanced HR dashboard that continuously tracks key attrition indicators showed in table 3. This tool should provide real-

time analytics and alerts, enabling HR to swiftly identify real-time news about employee who may want to quit. Such a dashboard would facilitate data-driven decision-making and proactive management of employee retention efforts.

Conclusion

By leveraging detailed attrition risk profiles and implementing targeted interventions, companies can significantly reduce employee turnover and foster a more engaged and loyal workforce. The proposed strategies not only address immediate concerns but also contribute to long-term organizational stability and success.

VI. Appendix: Plot For Employee Attrition Profile

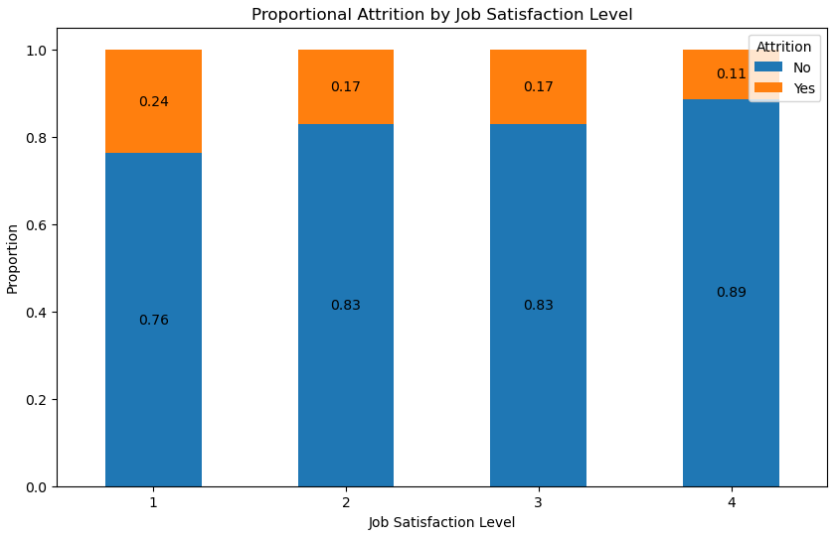


Figure 7. Proportional Attrition by Satisfaction Level

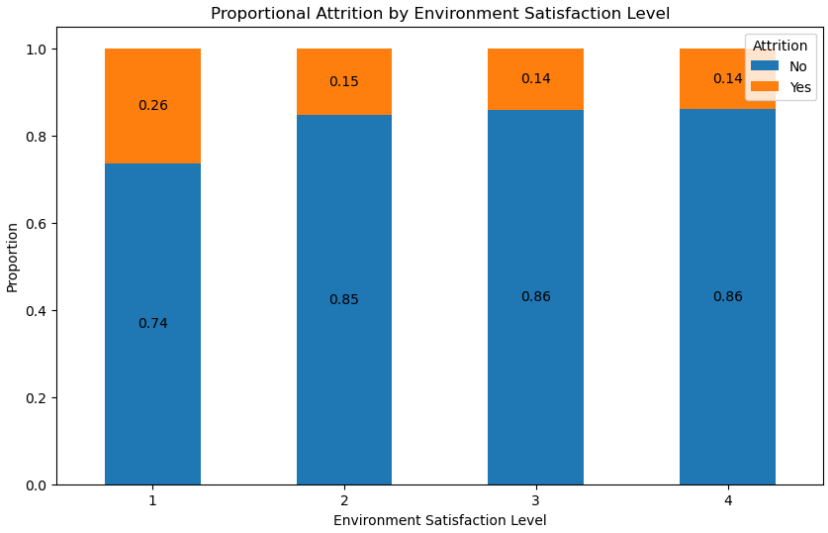


Figure 8. Proportional Attrition by Environment Satisfaction Level

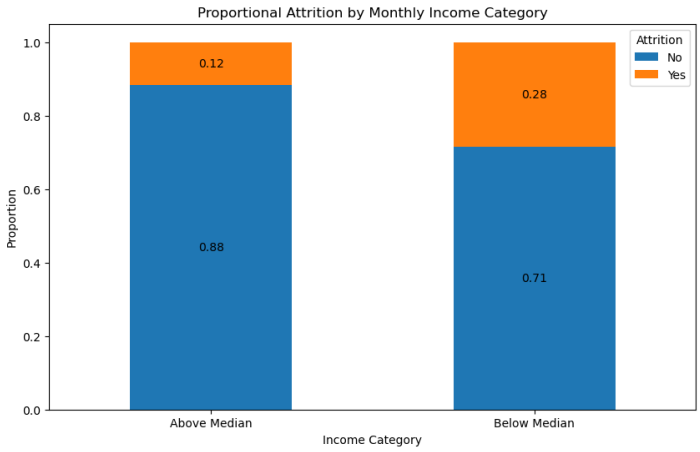


Figure 9. Proportional Attrition by Monthly Income Level

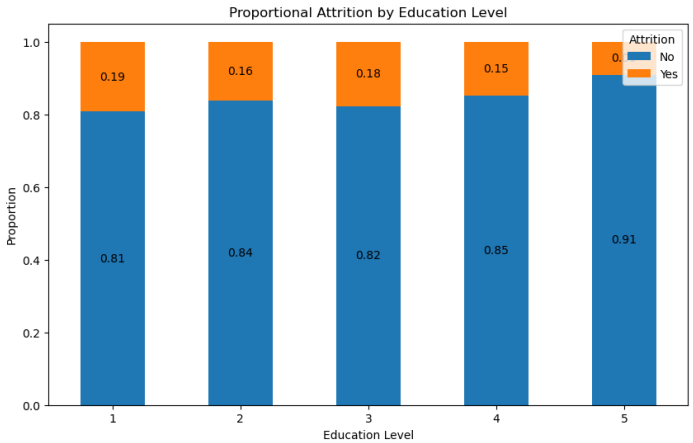


Figure 10. Proportional Attrition by Education Level

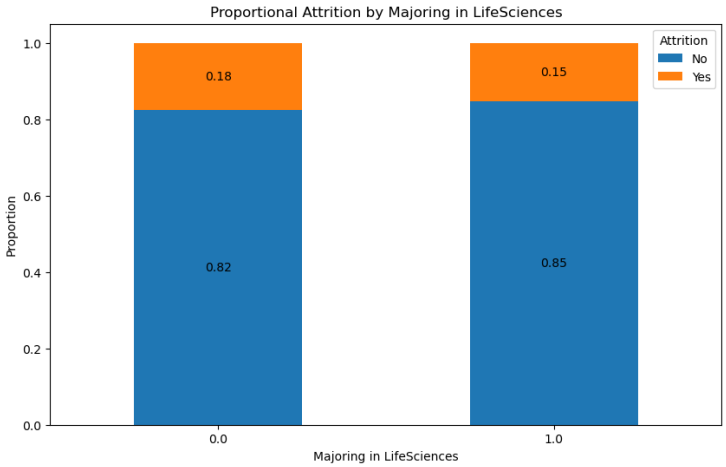


Figure 11. Proportional Attrition by Majoring in Life Science

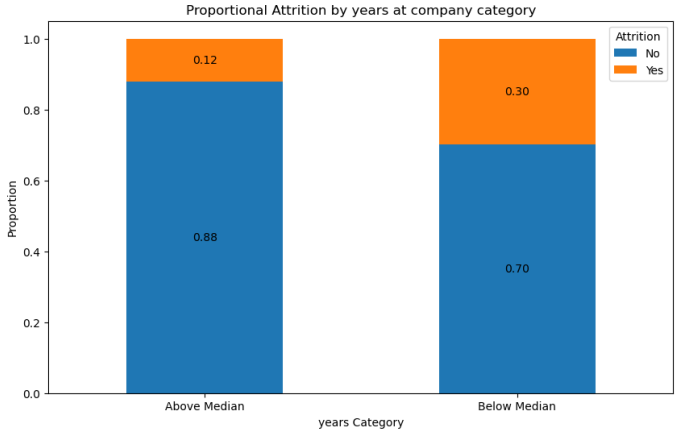


Figure 12. Proportional Attrition by years at company

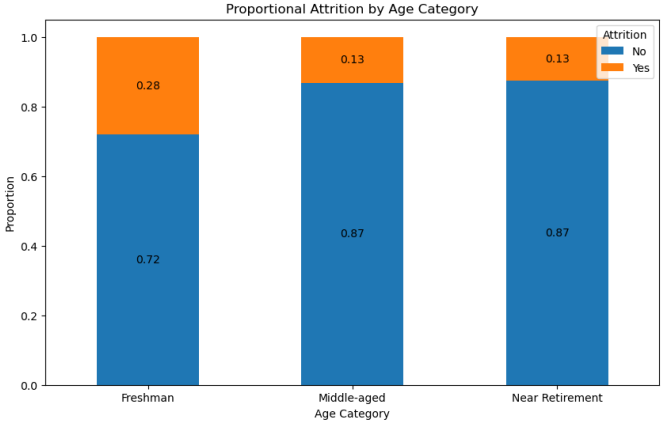


Figure 13. Proportional Attrition by Age Category

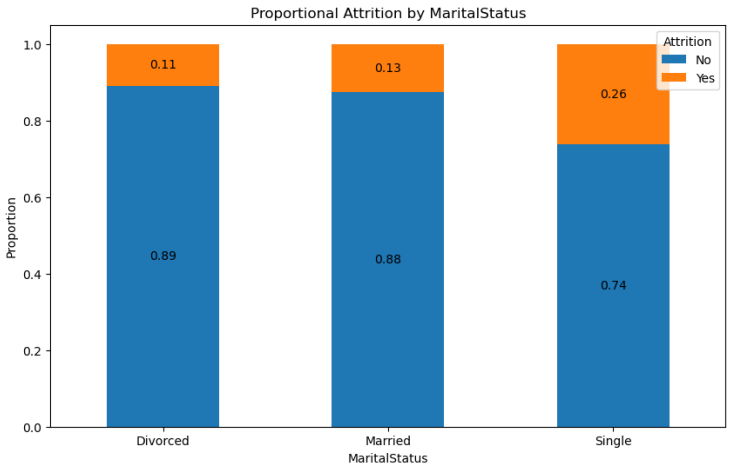


Figure 14. Proportional Attrition by Marital Status

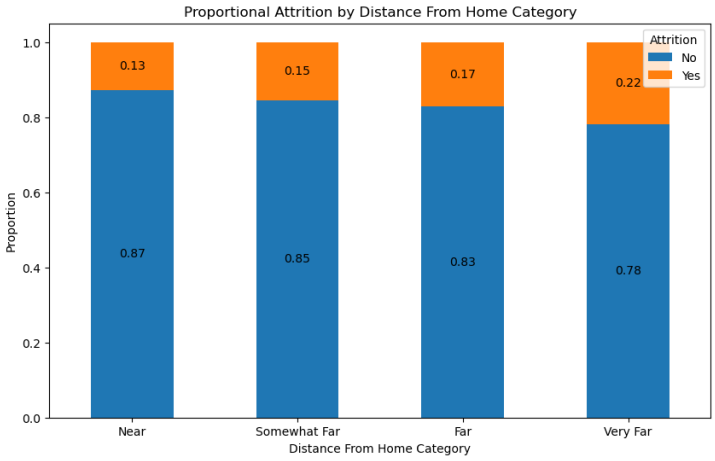


Figure 15. Proportional Attrition by Distance From Home

VII. Reference

1. Hossain, Y. (2021). *IBM Attrition Dataset* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/yasserh/ibm-attrition-dataset/data