THE PREDICTIVE EDGE: STACKING ENSEMBLE FOR SMARTER HR DECISION-MAKING

- BY UDOCHI OGBONNA

PROBLEM OVERVIEW

- Employee attrition refers to the natural process of employees voluntarily or involuntarily leaving a company or organization. According to Gallup, an average of 22% of employees depart annually, requiring replacement over a 12-month period. Furthermore, SHRM reports that approximately 45% of employees leave within the first two years. This trend can be concerning for organizations, as it may indicate issues with:- Recruitment and selection processes- Onboarding and integration- Employee engagement and satisfaction- Career development and growth opportunities- Managerial support and leadership.

In this case study, we aim to:

- 1. Identify the key factors contributing to employee attrition.
- 2. Develop a predictive model capable of classifying employee attrition.

By exploring these factors and building a predictive model, we hope to provide insights for organizations to address employee attrition and improve retention strategies.

METHODOLOGY

1.DATA COLLECTION :

- collect relevant data on employee attributes, performance, and exit information

2. EXPLORATORY DATA ANALYSIS:

- analyze data distribution, summary statistics, and correlations
- visualize data using plots and charts to understand trends and patterns

3. FEATURE ENGINEERING:

- extract relevant features from data, such as employee tenure, performance ratings, and job title
- create new features through feature transformations and combinations

4. DATA PREPROCESSING AND FEATURE IMPORTANCE/SELECTION:

- Clean and preprocess data by encoding categorical variables, and normalizing/scale data
- Features scaling of datasets to evaluate feature importance
- Select most relevant features for model development

METHODOLOGY CONTINUED

5. MODEL DEVELOPMENT:

- train machine learning models using selected features and algorithms.

6. HYPERPARAMETER OPTIMIZATION:

- tune models using hyperparameter optimization techniques

- use techniques like grid search to find optimal hyperparameters.

7. EVALUATION OF METRICS, COMPARISON, AND SELECTION:

- evaluate model performance using metrics like accuracy, precision, recall, f1-score, and ROC-AUC

- compare models and select the best-performing one

8 APPLYING STACKING ENSEMBLE TECHNIQUE

- choose base learners and a meta-learner that complement each other in terms of strengths and weaknesses.

METHODOLOGY CONTINUED

9. PRODUCTIONIZATION:

- **metrics evaluation**

- deploy the selected model in a production-ready environment

- **monitor model performance and retrain as needed**

10. POTENTIAL BENEFITS AND RECOMMENDATION:

- identify potential benefits of using the predictive model, such as reduced staff turnover rates and improved employee retention

- provide recommendations for HR and management to utilize the model and improve employee satisfaction and engagement.

DATA COLLECTION AND PREPROCESSING:

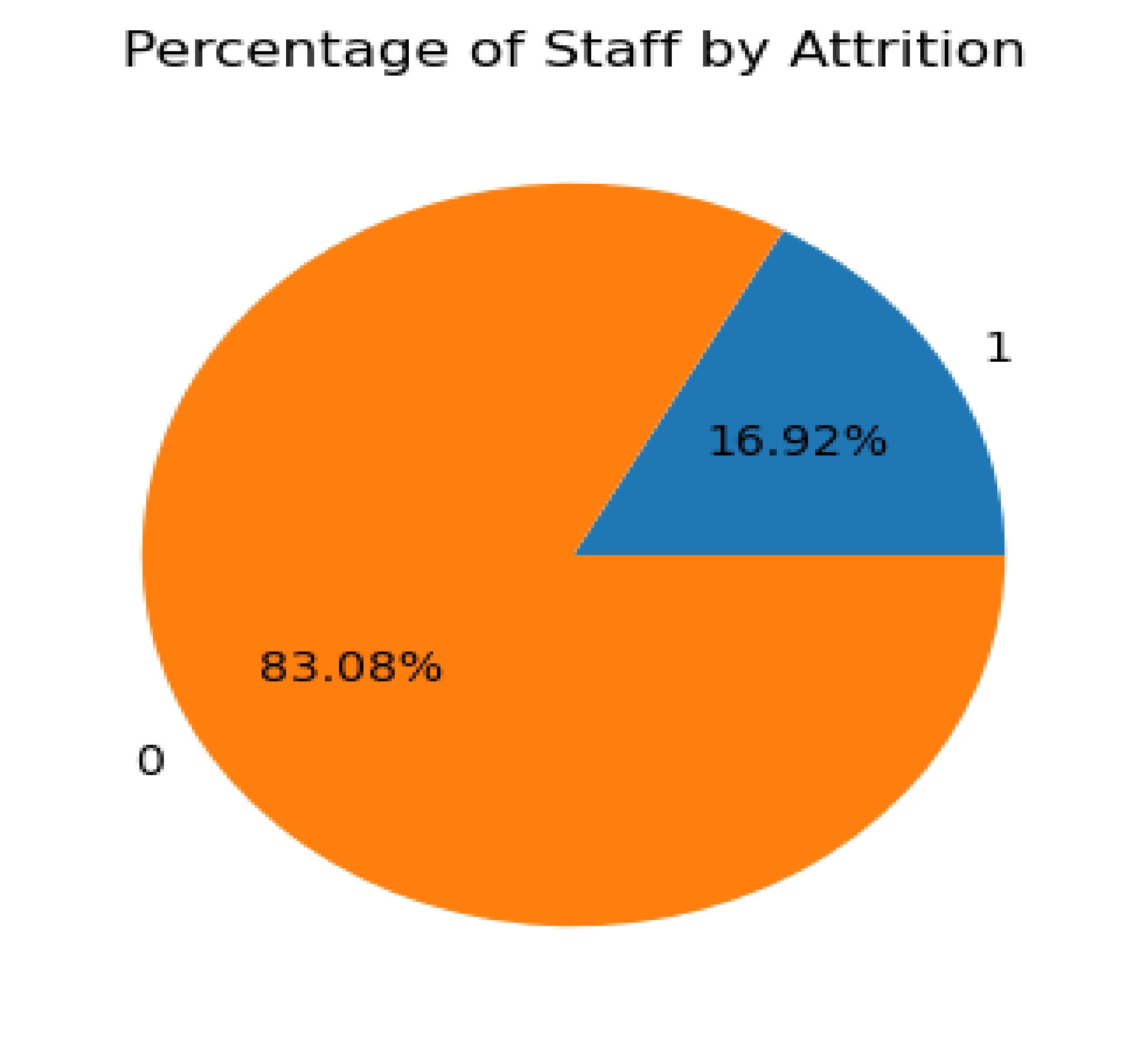
- The dataset contains 35 columns with 1,058 entries. It has int64(27) and object (8).
- The Oldest person is 60 years while the youngest is 18 years old. Attrition is classified as 0 and 1. The highest monthly income earner is approximately $20,000, the average monthly income earner is $4,900 while the least earns $1,009.
- It shows 3 department, 6 educational field and job role is 9.
- There are no missing and no duplicated values.

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

STAFF ATTRITION

This shows a notable class imbalance, with 17% of employees having attributed. This significant disparity between the minority class (attributed) and majority class (not attributed) requires careful consideration in modeling and prediction to avoid bias and ensure accurate results.

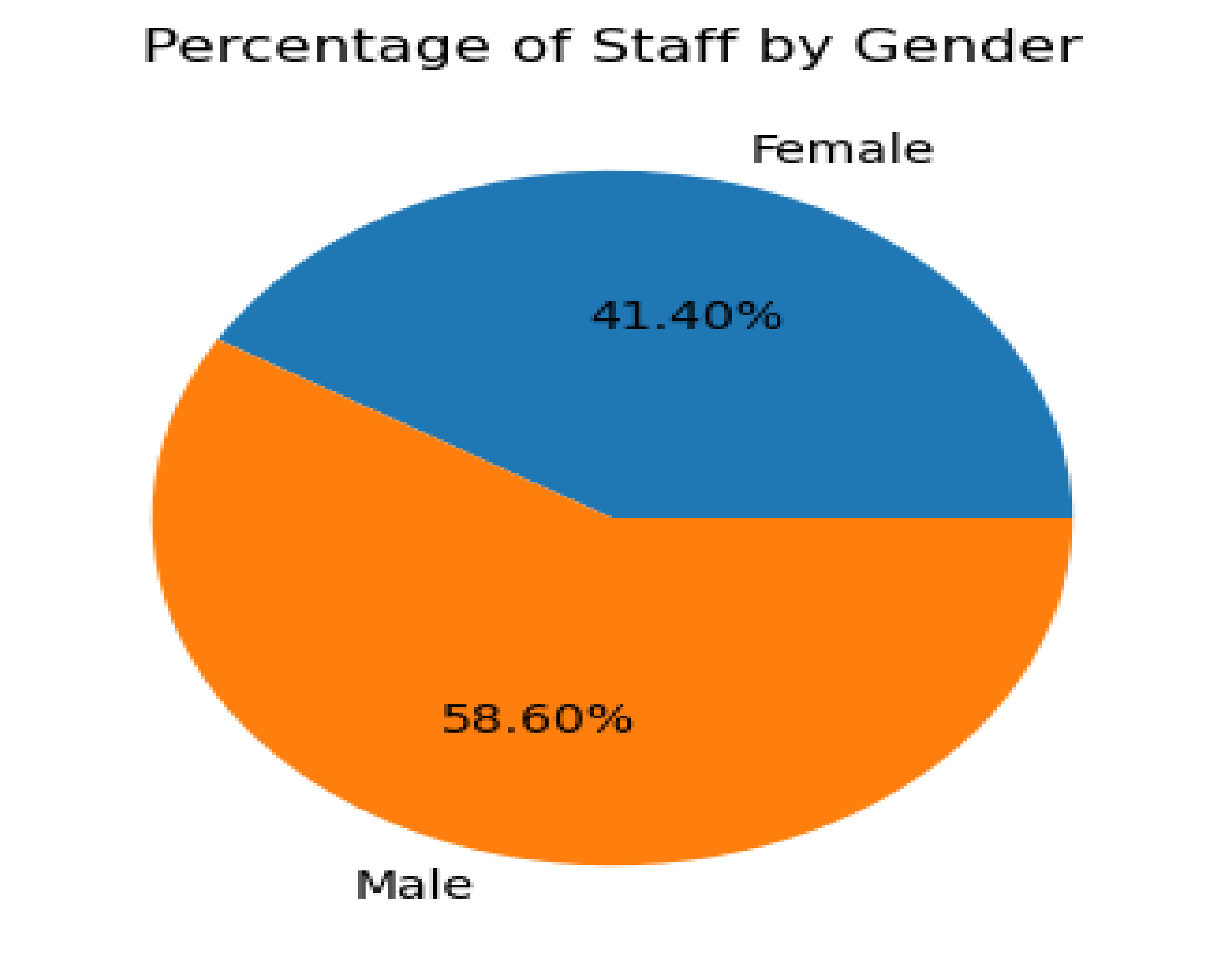


EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

GENDER

Gender ratio is male (59%) while female (41%).



EXPLORATORY DATA ANALYSIS

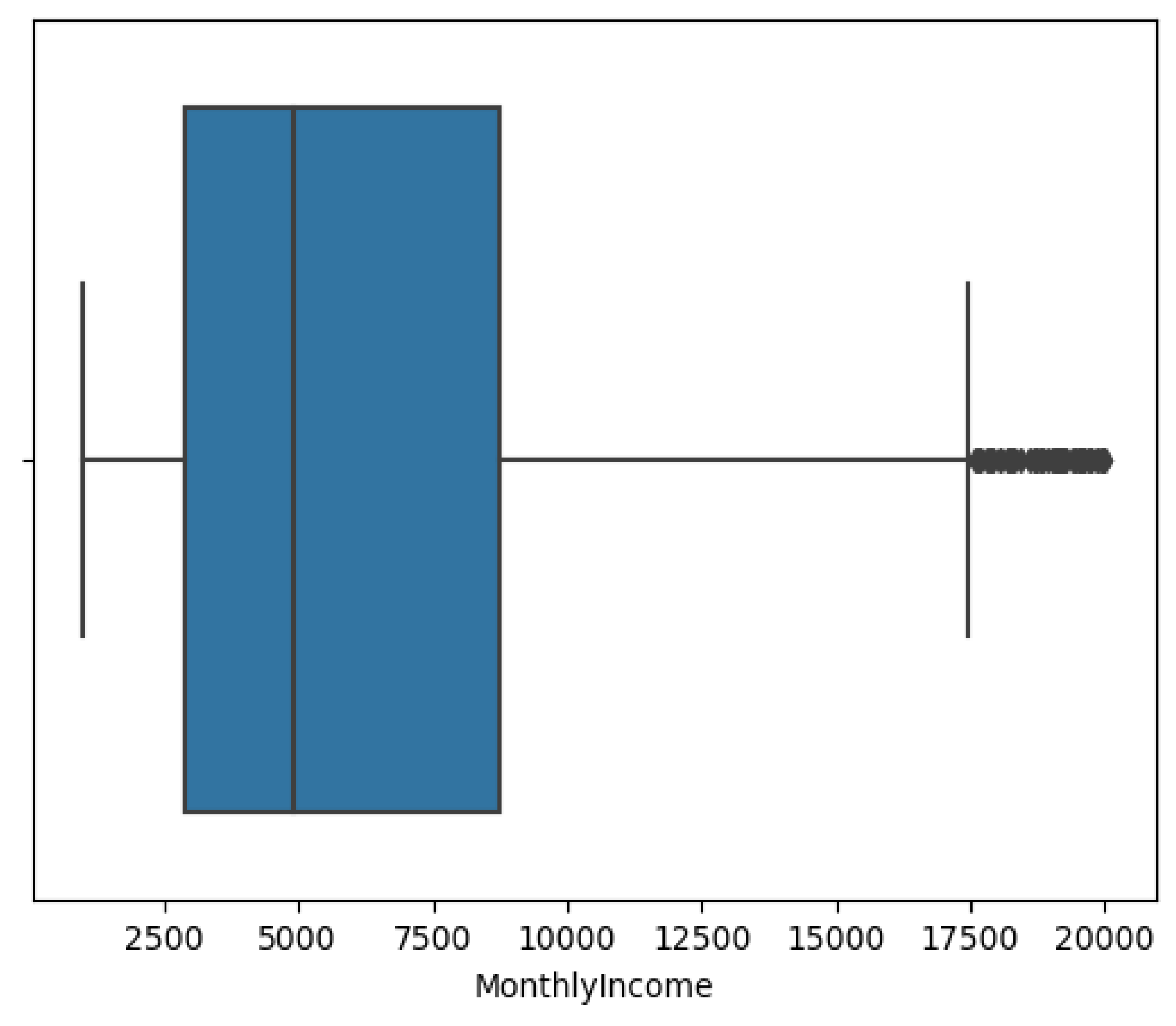
UNIVARIATE ANALYSIS

MONTLY INCOME DISTRIBUTION

The presence of outliers from (17,500 –19,990) in the monthly income data suggests that there are individuals who earn significantly more than the average income of $5000. These outliers could be due to various factors such as:

1. High-paying jobs or professions
2. Errors in data entry or collection
3. Other factors not accounted for in the data

- These outliers can significantly impact the distribution of the data and may affect the accuracy of models trained on this data.

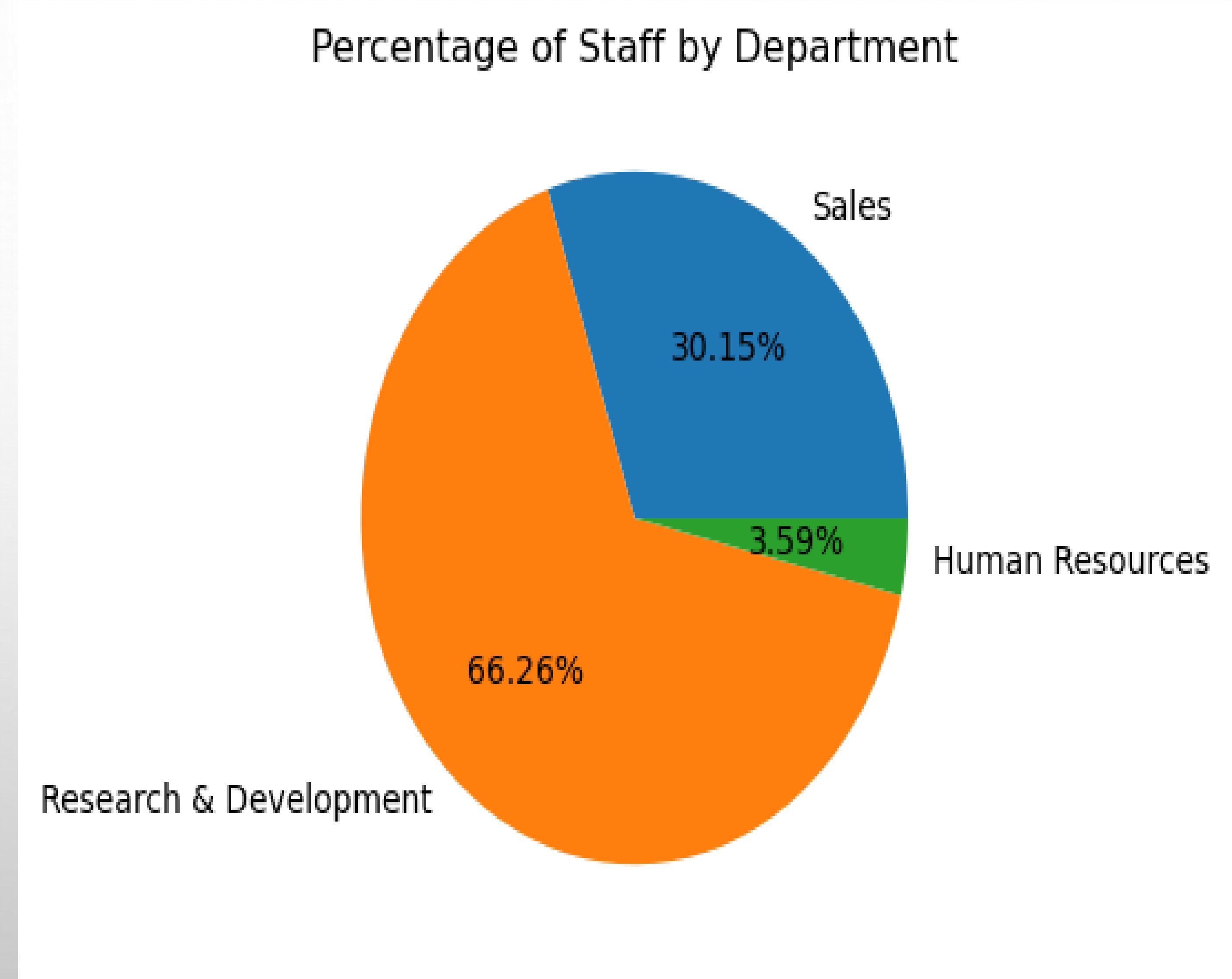


EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

PERCENTAGE OF STAFF BY DEPARTMENT

About 66% of the staff are in R&D, 30% in sales while 4% are in HR.



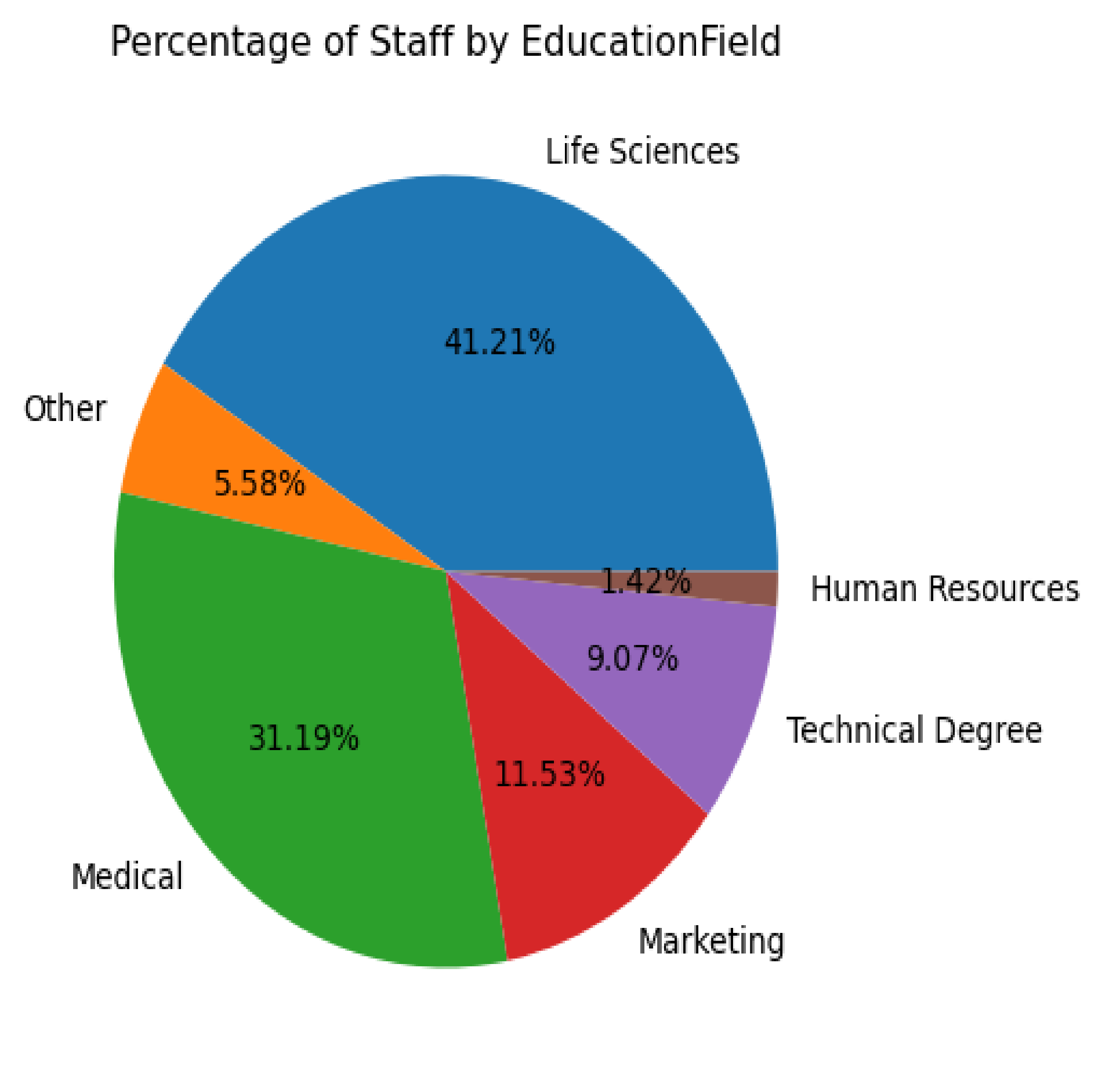
EXPLORATORY DATA ANALYSIS

- **UNIVARIATE ANALYSIS**

PERCENTAGE OF STAFF BY EDUCATIONAL FIELD

The data reveals

- Life sciences: 41%
- Medical: 31%
- Sales: 11%
- Others: 17% (which includes HR, technical and other support functions)
- This distribution suggests that the company is likely focused on developing and commercializing life sciences and medical products or services. The high percentage of staff with life sciences and medical backgrounds indicates a strong technical expertise in these areas the outliers as seen on the monthly income chart,
- The relatively smaller proportion of staff with sales backgrounds (11%) may indicate that the company is more focused on research and development, and may be relying on external partners or channels for sales and marketing.

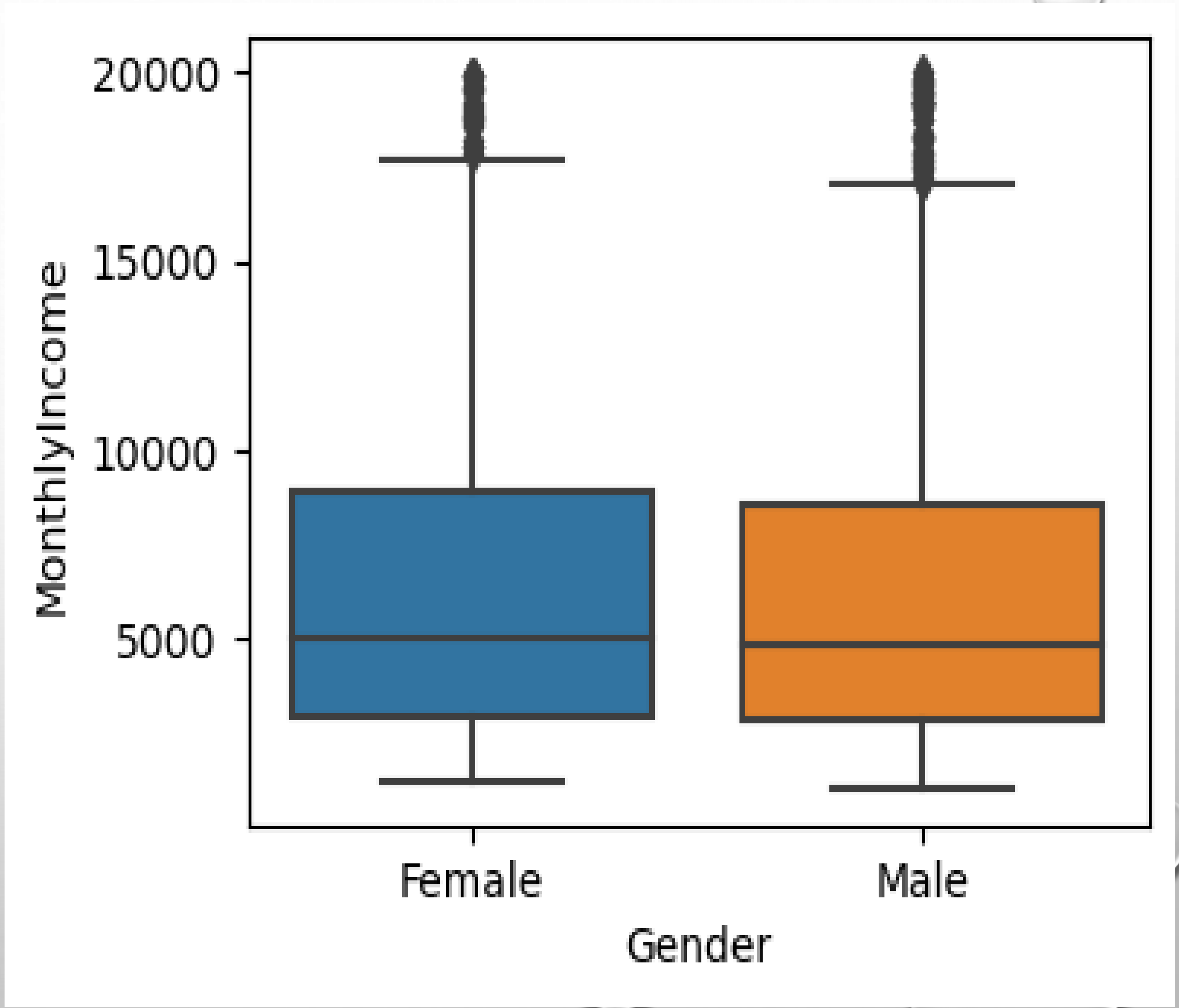


EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS

RELATIONSHIP BETWEEN GENDER VS MONTHLY INCOME

Gender does not determine the earning capacity of the staff



EXPLORATORY DATA ANALYSIS

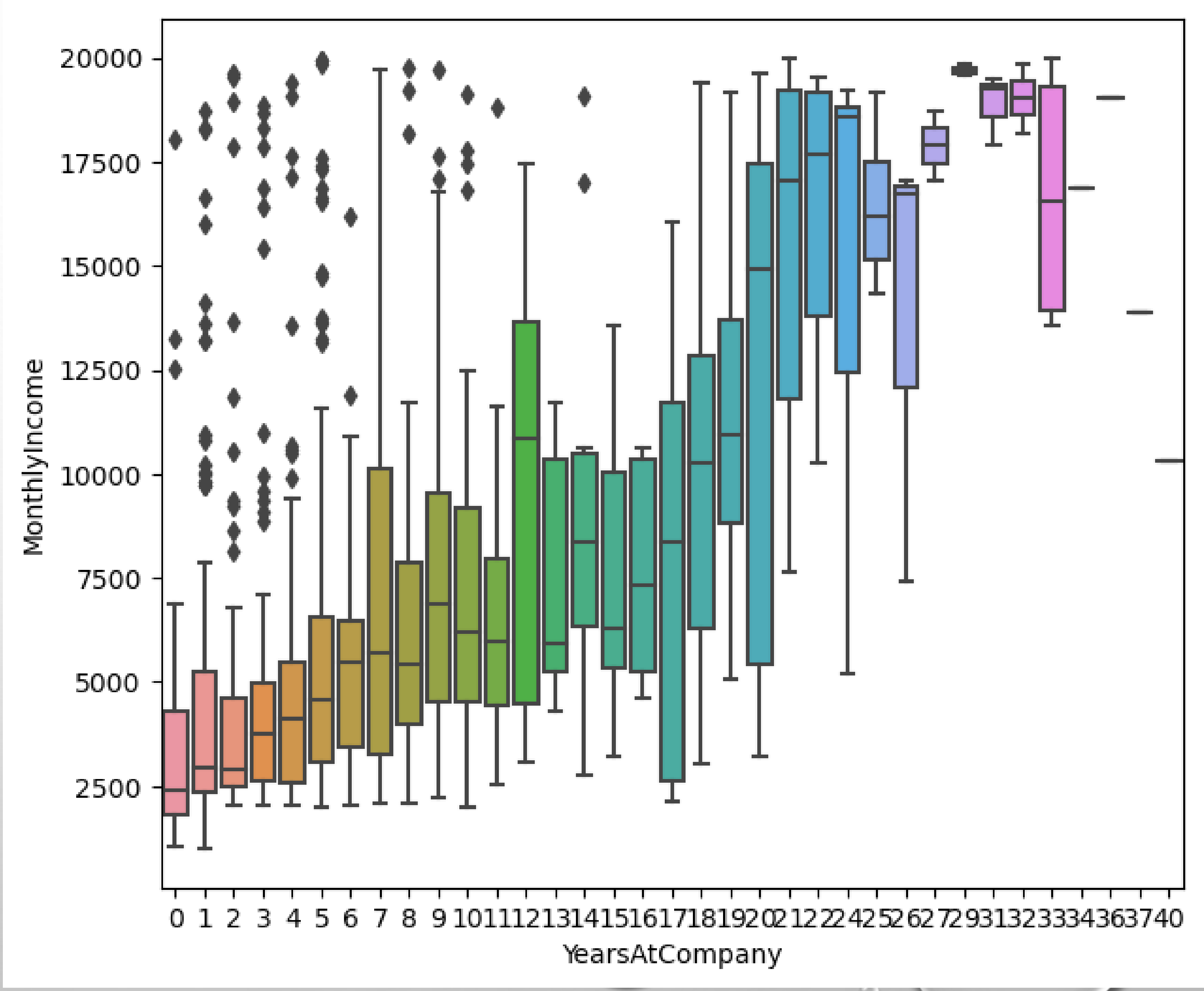
BIVARIATE ANALYSIS

- **RELATIONSHIP BETWEEN YEARS AT THE COMPANY VS MONTHLY INCOME**

The data shows that employees who stay with the company longer tend to earn more, but there are some exceptions that could indicate:

- Fast-tracked promotions for high performers
- Specialized skills leading to higher pay
- Successful salary negotiations
- - Data errors

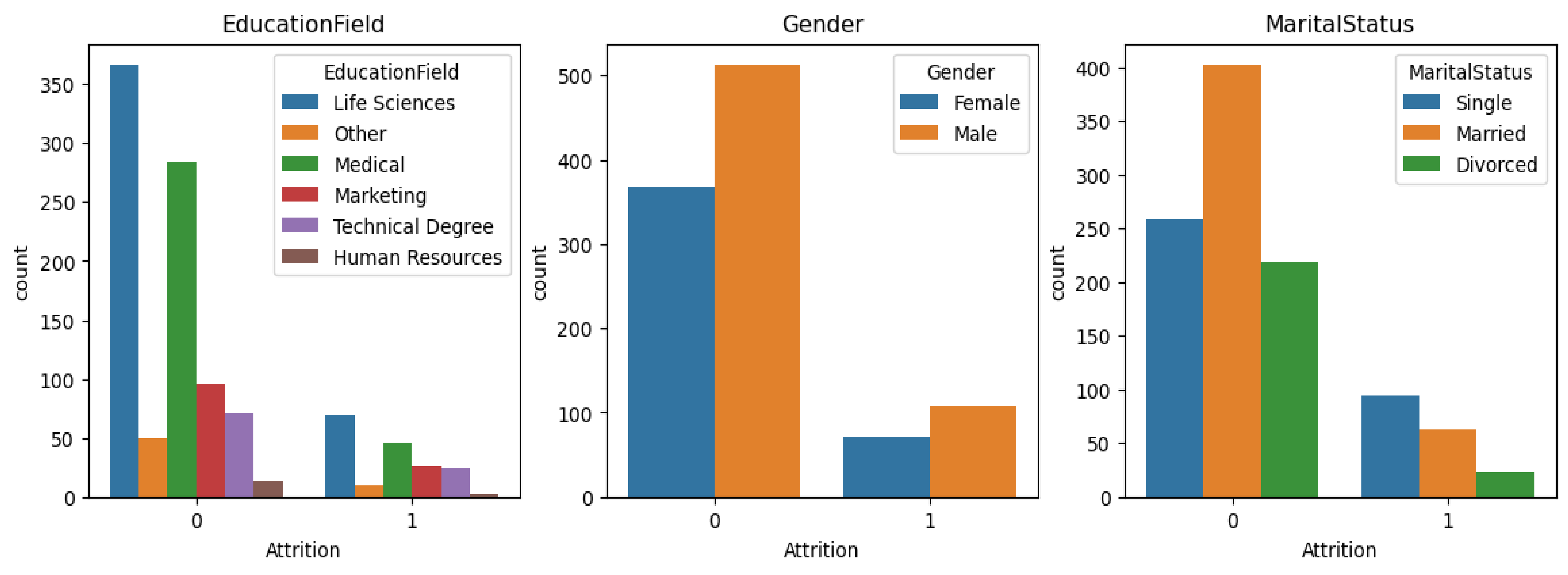
These exceptions can help the company identify areas to improve its compensation and retention strategies.



EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS

- **RELATIONSHIP BETWEEN ATTRITION AND CATEGORICAL VARIABLES.**

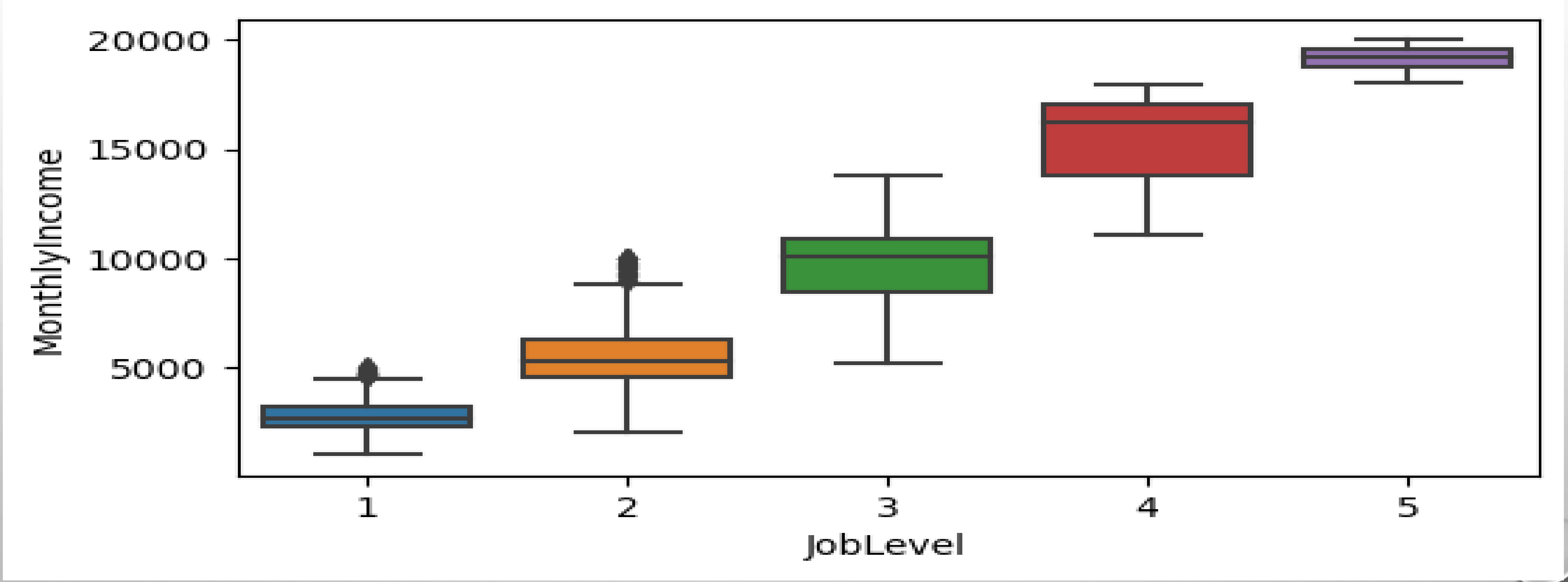


Singles, in Life Sciences and Medical have a high attrition rate.

EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS

- **RELATIONSHIP BETWEEN THE JOB LEVEL VS THE MONTHLY INCOME.**



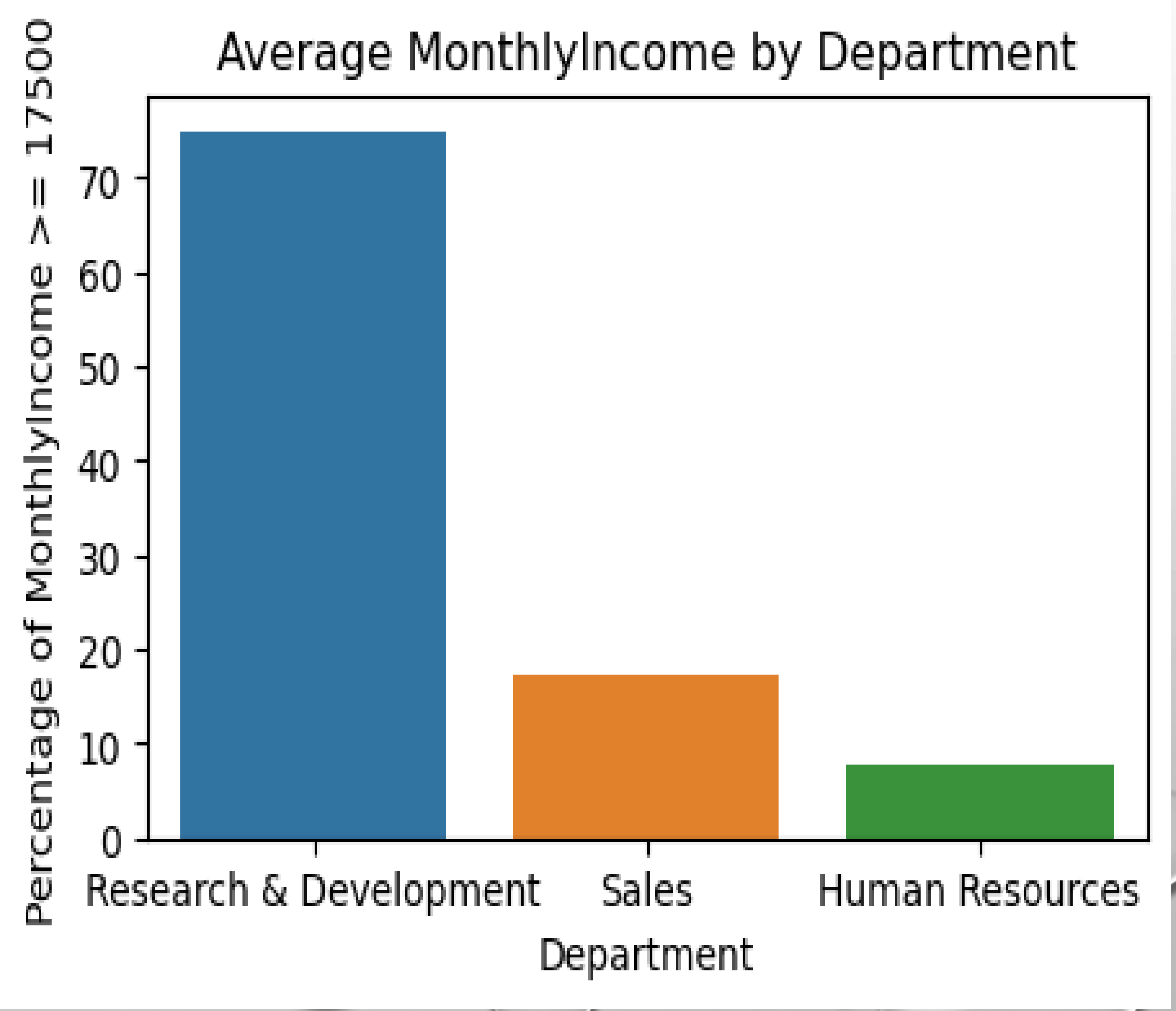
The data shows that the job levels are in 5 category and has a positive correlation with the monthly income. The chart depicts an average earning as low as 2,600 and others earn as high as 19,000 USD

EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS

RELATIONSHIP BETWEEN THE AVERAGE MONTHLYINCOME >= $17,500 BY DEPARTMENT

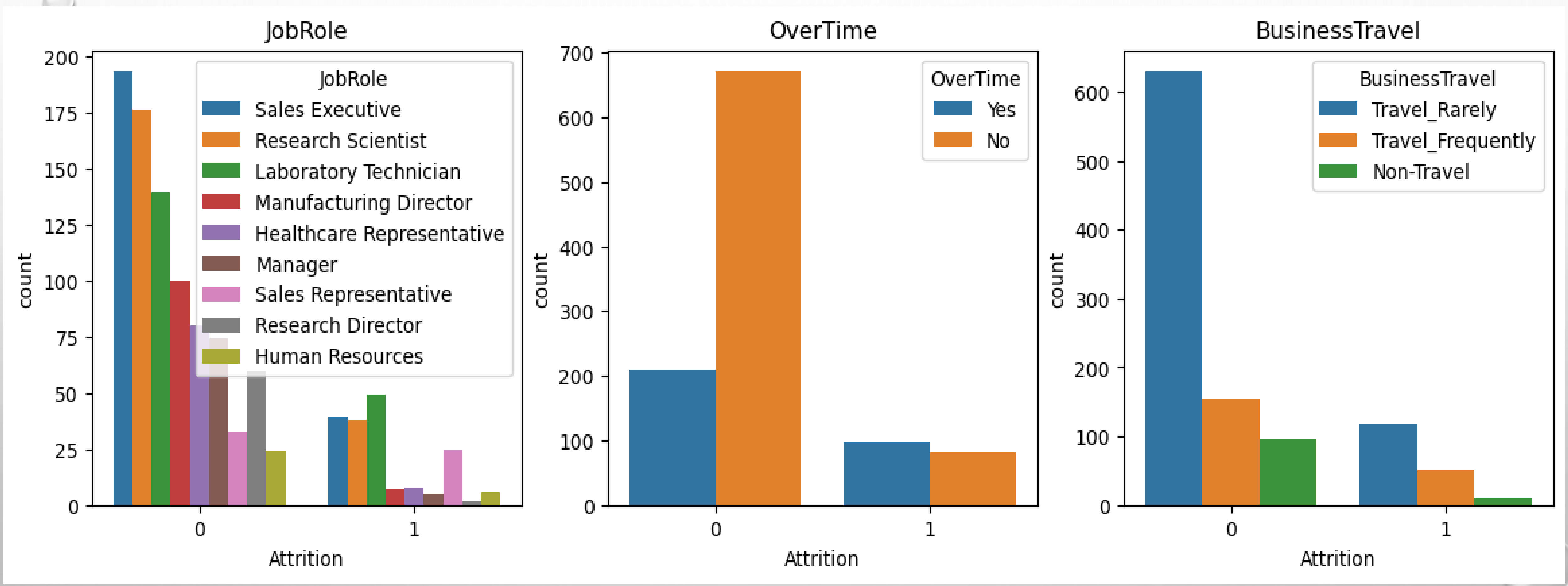
- **Based on insight from the outliers, it appears that:**
- **75% of Research & Development (R&D) department employees earn above $17,500, which is a significant proportion.**
- **17% of Sales department employees earn above $17,500, which is a relatively smaller proportion compared to R&D.**
- **7% of Human Resources department employees earn above $17,500, which is the smallest**



EXPLORATORY DATA ANALYSIS

BIVARIATE ANALYSIS

RELATIONSHIP BETWEEN ATTRITION AND CATEGORICAL VARIABLES.



Attrition risk is higher among Laboratory Technicians, Sales Executives, Research Scientists, and Sales Representatives, as well as employees with minimal travel demands. Overtime has no significant impact on attrition.

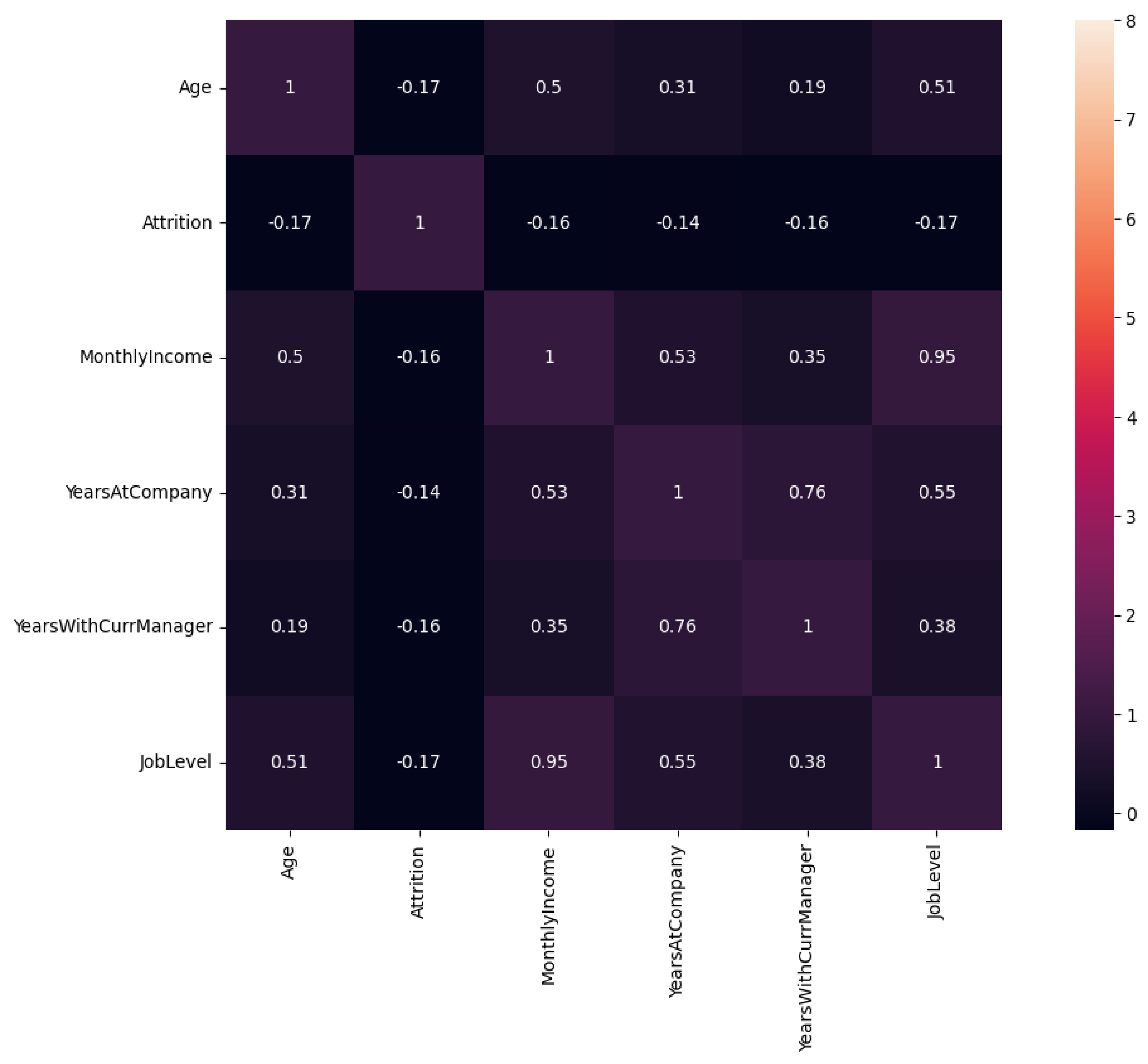
EXPLORATORY DATA ANALYSIS

MULTIVARIATE ANALYSIS (CORRELATION)

The analysis reveals that attrition is negatively correlated with the following factors:

- **Age**
- **Monthly income**
- **Years at the company**
- **Years with the current manager**
- **Job level**

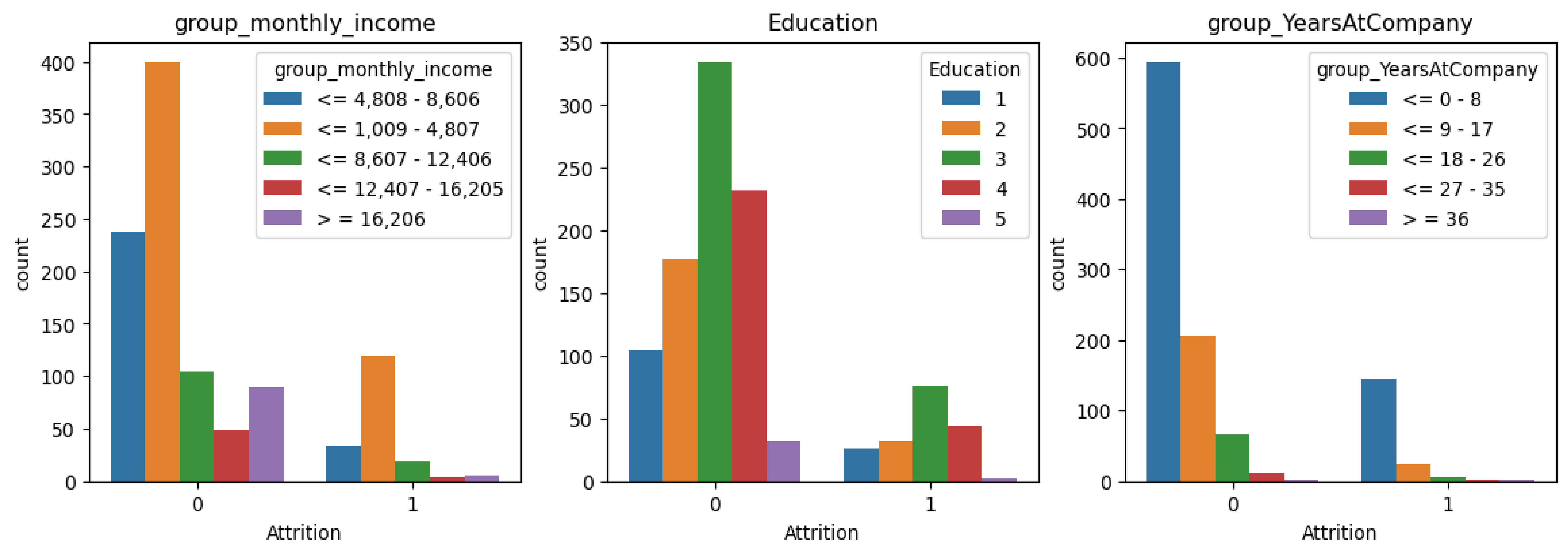
This means that as these factors increase, the likelihood of attrition decreases. In other words, older employees, those with higher monthly incomes, longer tenure, longer time with their current manager, and higher job levels tend to have lower rates of attrition.



	Age	Attrition	MonthlyIncome	YearsAtCompany	YearsWithCurrManager	JobLevel
Age	1	-0.17	0.5	0.31	0.19	0.51
Attrition	-0.17	1	-0.16	-0.14	-0.16	-0.17
MonthlyIncome	0.5	-0.16	1	0.53	0.35	0.95
YearsAtCompany	0.31	-0.14	0.53	1	0.76	0.55
YearsWithCurrManager	0.19	-0.16	0.35	0.76	1	0.38
JobLevel	0.51	-0.17	0.95	0.55	0.38	1

FEATURE ENGINEERING

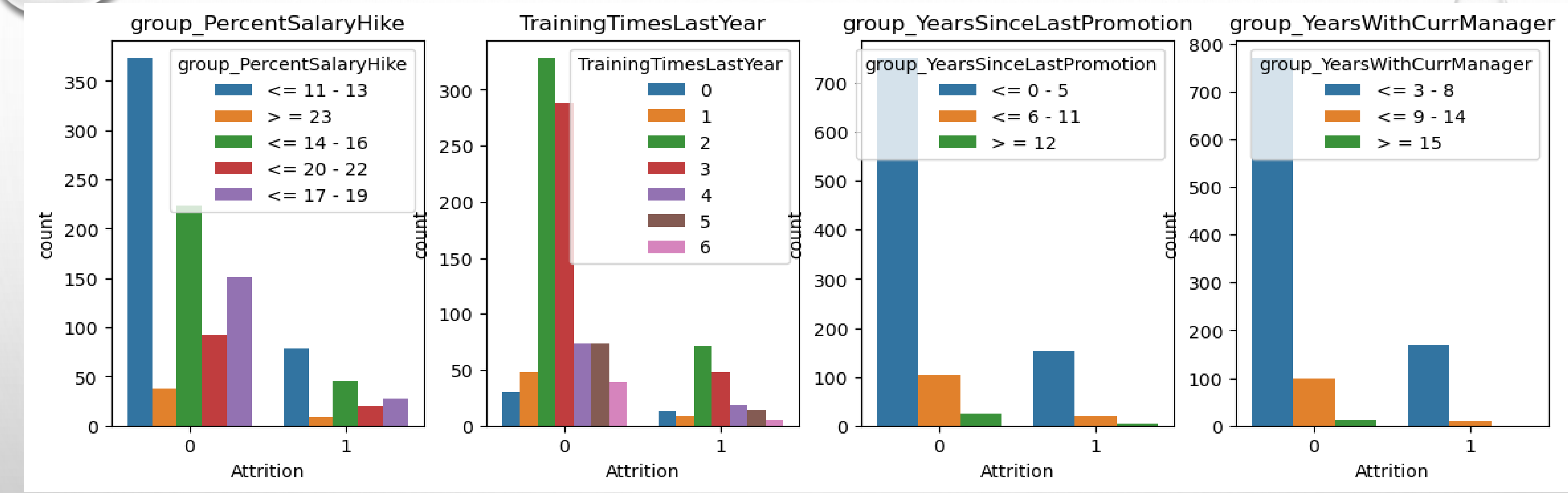
RELATIONSHIP BETWEEN ATTRITION AND FEATURED ENGINEERED VARIABLES.



The data indicates a significant correlation between employee turnover and the following factors: lower monthly earnings (USD 1000-4800), shorter tenure (0-8 years), and skill levels 3. this suggests that the company is experiencing a disproportionate loss of employees within these demographics.

FEATURE ENGINEERING

RELATIONSHIP BETWEEN ATTRITION AND FEATURED ENGINEERED VARIABLES.

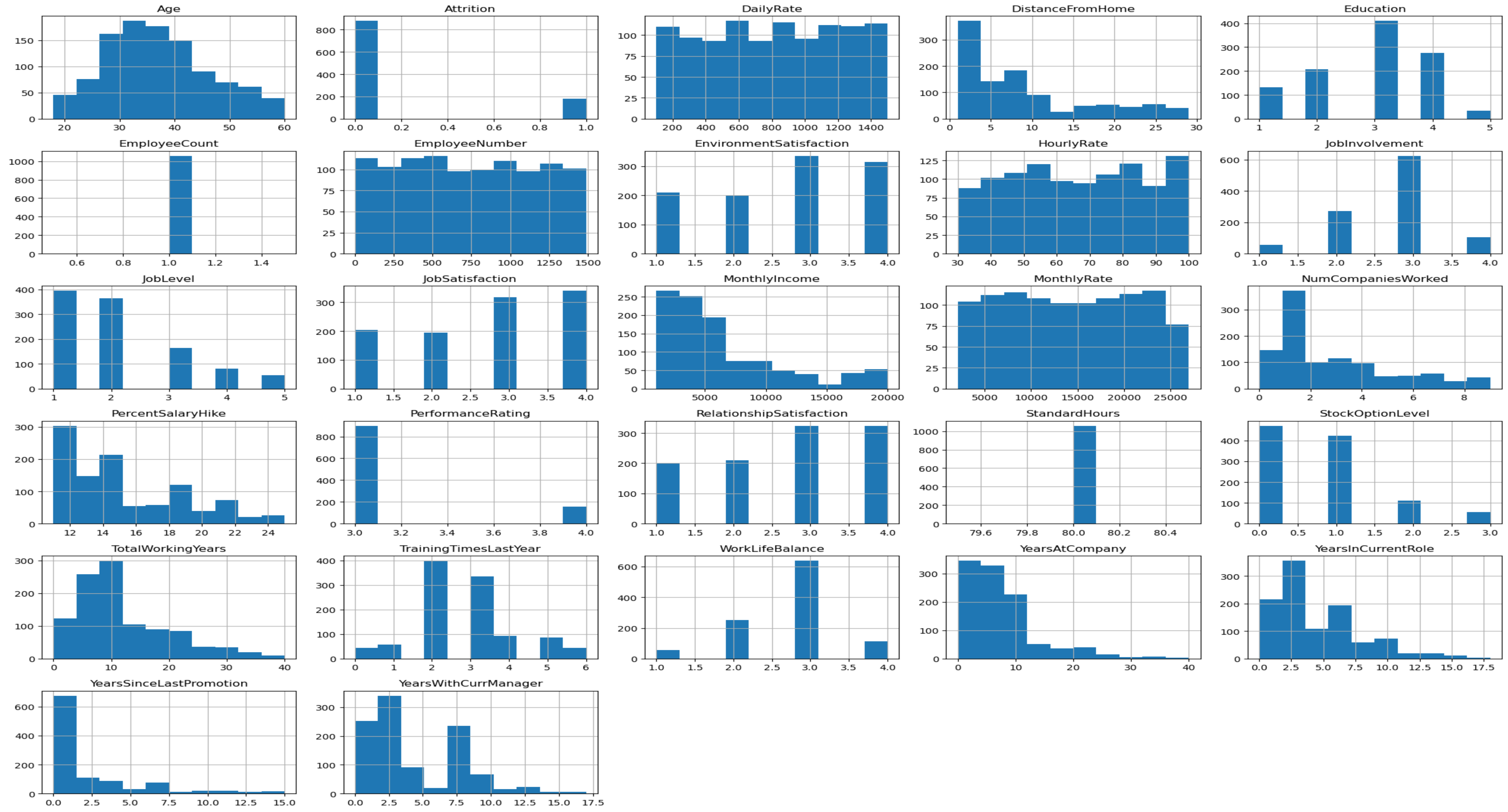


The dataset reveals that the employees who left the company typically had:

- **Received a salary increase of 11-13% or 14-16% in their last hike**
- **Participated in 2, 3, or 1 training sessions last year**
- **Received their last promotion 0-5 years ago**
- **Been under the same manager for 3-8 years**

This suggests that the company is losing employees who have recently received moderate salary increases, have had some training and development opportunities, and have been in their current role or under the same manager for a few years but may be seeking further opportunities or challenges.

RELATIONSHIP BETWEEN ATTRITION AND NUMERICAL VARIABLES.



DATA PREPROCESSING AND FEATURE IMPORTANCE/SELECTION

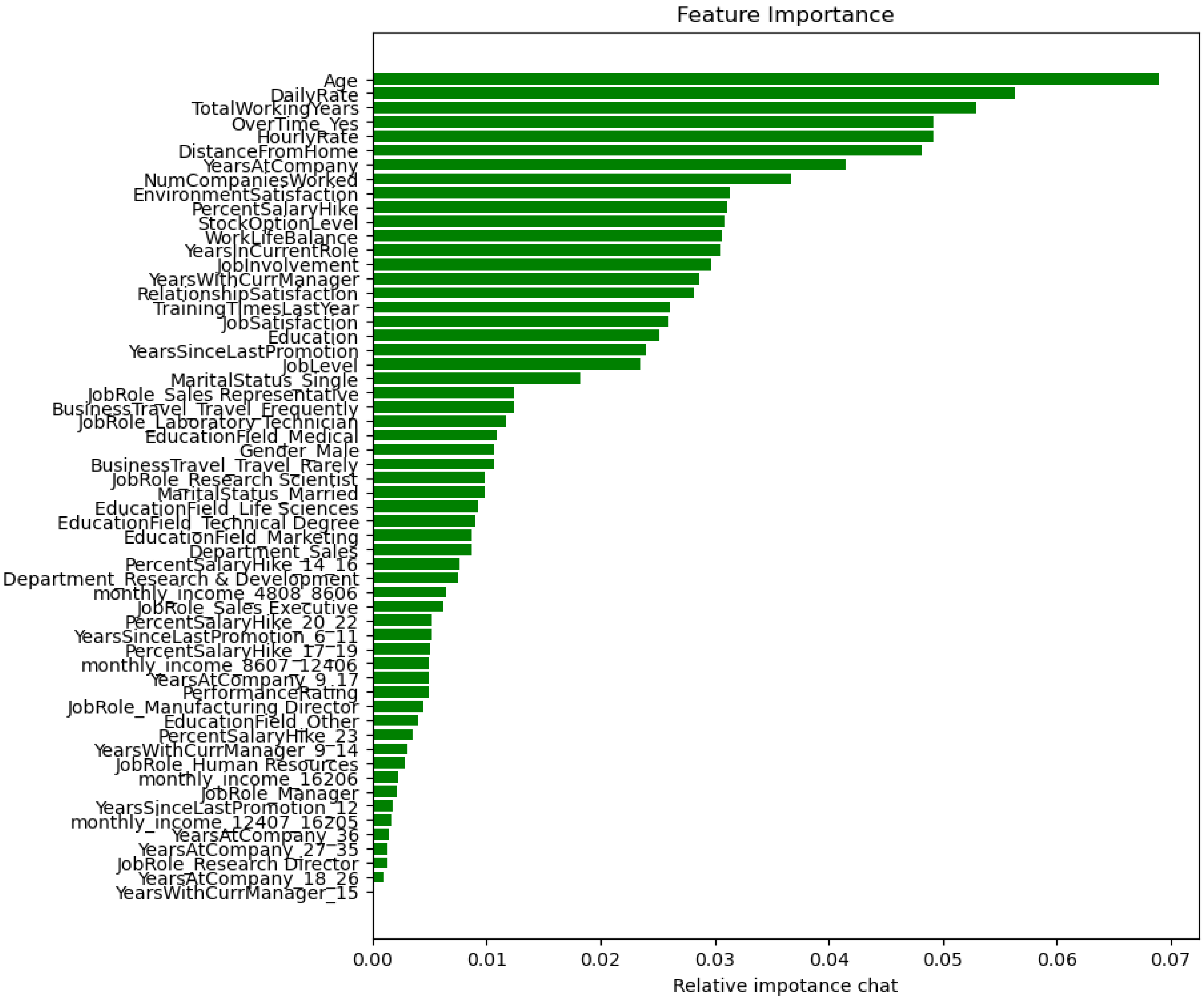
- DROPPING OF REDUNDANT FEATURES
- ENCODE THE CATEGORICAL FEATURES IN THE DATA
- SEGMENT DATASET INTO DATA AND TARGET LABEL
- SCALE DATASET FEATURES
- INSTANTIATING THE SCALER OBJECT
- IDENTIFYING KEY FEATURES FROM THE DATA SET
- PLOTTING A FEATURE IMPORTANCE CHART
- FEATURE SELECTION

PLOTTING A FEATURE IMPORTANCE CHART

The most significant predictors of attrition include:

1.Age
2.Daily rate
3.Total working years
4.Overtime
5.Hourly Rate
6.Distance from home
7.Years at the Company
8.Number of Companies worked
9.Environmental Satisfaction
10.Percentage Salary hike
11.Etc.

These features are crucial in predicting employee turnover, indicating that a combination of demographic, financial, work-related, and environmental factors contribute to an employee's likelihood of leaving the company.



MODEL DEVELOPMENT

- SPLITTING DATA INTO TRAINING AND EVALUATION DATASETS
- OVERSAMPLING BECAUSE THE DATASET IS IMBALANCE
- IMPORTING CLASSIFIER LIBRARIES
- IMPORTING NECESSARY MODULES
- PERFORM CROSS-VALIDATION MAJORLY TO PREVENT OVERFITTING
- MODEL BUILDING USING 7 MACHINE LEARNING ALGORITHMS AND EVALUATION.
- HYPER PARAMETER OPTIMIZATION TUNNING OF THE 7 MODELS.
- METRICS EVALUATION

METRICS EVALUATION

- **WITHOUT TUNNING**:
- **ANALYSIS**
- Random forest is the best classifier for class 0, with a high accuracy rate and low error rate.
- Logistic regression is the best classifier for class 1, with a high accuracy rate and low error rate.
- Naive bayes has a high error rate for both class 0 and class 1.
- Xgb classifier has a high error rate for class 1.
- **Overall, the models were poor in discriminating attrition and non attrition.**

This requires further investigation using hyper parameter tunning.

HYPER PARAMETER OPTIMIZATION

```
Confusion Matrix for RFC:
[[74.52830189      7.0754717 ]
 [13.20754717       5.18867925]]

Classification Report for RFC:
              precision    recall  f1-score   support

           0       0.85      0.91      0.88       173
           1       0.42      0.28      0.34        39

    accuracy                           0.80

AUC/ROC for RFC: 0.5976730398695717
```

```
Confusion Matrix for SVC:
[[69.33962264      12.26415094]
 [12.26415094      6.13207547]]

Classification Report for SVC:
              precision    recall  f1-score   support

           0       0.85      0.85      0.85       173
           1       0.33      0.33      0.33        39

    accuracy                           0.75

AUC/ROC for SVC: 0.5915221579961464
```

HYPER PARAMETER OPTIMIZATION

```
Confusion Matrix for LR:
[[61.79245283       19.81132075]
 [ 4.24528302       14.1509434 ]]

Classification Report for LR:
              precision    recall  f1-score   support

           0       0.94      0.76      0.84       173
           1       0.42      0.77      0.54        39

    accuracy                           0.76

AUC/ROC for LR: 0.7632281013783904
```

```
Confusion Matrix for XGBC:
[[73.58490566       8.01886792]
 [12.26415094       6.13207547]]

Classification Report for XGBC:
              precision    recall  f1-score   support

           0       0.86      0.90      0.88       173
           1       0.43      0.33      0.38        39

    accuracy                           0.80

AUC/ROC for XGBC: 0.617533718689788
```

HYPER PARAMETER OPTIMIZATION

```
Confusion Matrix for SGDC:
[[77.83018868   3.77358491]
 [13.20754717   5.18867925]]

Classification Report for SGDC:
              precision    recall  f1-score   support

           0       0.85      0.95      0.90       173
           1       0.58      0.28      0.38        39

    accuracy                           0.83

AUC/ROC for SGDC: 0.617904253742404
```

```
Confusion Matrix for GNB:
[[45.75471698   35.8490566 ]
 [ 3.30188679   15.09433962]]

Classification Report for GNB:
              precision    recall  f1-score   support

           0       0.93      0.56      0.70       173
           1       0.30      0.82      0.44        39

    accuracy                           0.61

AUC/ROC for GNB: 0.6906032310656588
```

HYPER PARAMETER OPTIMIZATION

```
Confusion Matrix for DTC:
[[69.33962264      12.26415094]
 [11.32075472       7.0754717 ]]

Classification Report for DTC:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85       173
           1       0.37      0.38      0.37        39

    accuracy                           0.76

AUC/ROC for DTC: 0.617163183637172
```

METRICS EVALUATION

Model Suitability Analysis for Imbalanced Dataset:

Logistic Regression (LR) stands out as the best model for predicting employee attrition. It has the highest recall (0.77) for the minority class, which is crucial in imbalanced classification problems. This high recall ensures that most employees likely to leave are correctly identified, which is critical in attrition prediction. Moreover, the AUC-ROC (0.76) indicates that it can effectively distinguish between those who stay and those who leave.

Random Forest Classifier (RFC) and XGBoost (XGBC) are good for the majority class but underperform for the minority class. Their precision and recall for class 1 are lower than LR, making them less suitable for detecting employee attrition cases in imbalanced datasets.

Gaussian Naive Bayes (GNB) offers the best recall (0.82) for the minority class, but its very low precision (0.30) means that it has a high false-positive rate. This could lead to unnecessary interventions for employees who are not likely to leave.

Support Vector Classifier (SVC), SGDC, and DTC offer moderate performance but lack the sensitivity needed for the minority class in terms of recall and precision.

METRICS EVALUATION

Conclusion:

For predicting employee attrition in an imbalanced dataset:

Logistic Regression (LR) is the best choice, balancing recall for identifying those likely to leave while maintaining a good AUC-ROC. **This is not robust for non –attrition cases**
Gaussian Naive Bayes (GNB) could be considered if the primary concern is minimizing missed attrition cases, but its high false positives may be problematic.
Ensemble models like Random Forest and XGBoost are less suited due to their lower recall for the minority class.
The Confusion Matric indicates a relatively high FN which might be too costly for the Organization.

To enhance the rigor of our modeling, we must employ a Stacking Ensemble approach that leverages the synergy of strengths and weaknesses among multiple models, resulting in a robust and reliable prediction.

STACKING ENSEMBLE METHOD (CONFIGURATION)

For stacking ensemble models in the context of imbalanced datasets and employee attrition prediction, we need to choose base learners and a meta-learner that complement each other in terms of strengths and weaknesses. Here's an analysis of which models from the previous (tuned) results are suitable for stacking:

Base Learners: Base learners should be diverse and complement each other. A good ensemble should combine models that have different biases or approaches, improving the overall predictive power.

Logistic Regression (LR):

Strengths: High recall (0.77) for the minority class, making it excellent at catching attrition cases.

Weaknesses: Moderate precision.

Role: Ideal as a base learner because its linear nature and high recall can provide strong identification of class 1 (attrition) cases. It can capture the overall trend.

Random Forest Classifier (RFC):

Strengths: Strong performance for the majority class (class 0), with balanced performance overall.

Weaknesses: Weak minority class recall (0.28).

Role: RFC can be a strong base learner as its tree-based method captures non-linear relationships well, and its strength in class 0 compensates for weaknesses of models that focus more on the minority class.

STACKING ENSEMBLE METHOD (CONFIGURATION)

XGBoost Classifier (XGBC):

Strengths: Generally robust, handles non-linear patterns, and has a moderate recall (0.33) for the minority class.

Weaknesses: Slightly lower minority class recall than LR.

Role: XGBoost is another solid base learner, bringing in robust feature handling, especially for complex interactions between features.

Support Vector Classifier (SVC):

Strengths: Good balance of precision and recall for both classes.

Weaknesses: Lower overall recall for the minority class.

Role: SVC can be added to the stack, as its kernel-based approach complements the decision-based models (trees) and linear models (LR), helping to capture more nuanced decision boundaries.

Gaussian Naive Bayes (GNB):

Strengths: Excellent recall (0.82) for the minority class.

Weaknesses: Low precision for class 1 (attrition), meaning many false positives.

Role: GNB is useful in cases where sensitivity to the minority class is key. It can be a good base learner because its probabilistic approach may complement the deterministic methods of the other models.

STACKING ENSEMBLE METHOD (CONFIGURATION)

Meta-Learner:
The meta-learner will take the predictions from the base models and make the final prediction. It should be able to combine the strengths of the base learners effectively.

Logistic Regression (LR): Frequently used as a meta-learner because it performs well in linearly combining the outputs of base models. It is simple, interpretable, and tends to perform well when used as a second-level model in stacking ensembles.
XGBoost (XGBC) or Random Forest (RFC): Can also be strong choices for a meta-learner, as they can capture complex relationships between the base model predictions. However, they can overfit if the base learners are too similar.

Recommended Stacking Ensemble Configuration:
Base Learners:
Logistic Regression (LR): High recall and overall balanced performance.
Random Forest (RFC): Strong performance for class 0 and good handling of non-linear features.
XGBoost (XGBC): A powerful model for non-linear relationships, with moderate recall for class 1.
Gaussian Naive Bayes (GNB): High recall for the minority class, which may help reduce missed attrition cases.
Support Vector Classifier (SVC): Good balance between precision and recall, with kernel-based decision boundaries.
Meta-Learner:
Logistic Regression (LR) or XGBoost (XGBC): LR is preferred for its simplicity and generalization ability. XGBoost could be used if you suspect non-linear relationships between the base learners’ outputs.
This setup allows the ensemble to balance between catching the minority class (attrition cases) and not sacrificing too much precision, ultimately leading to a well-rounded predictive model suitable for the imbalanced dataset.

PRODUCTIONIZING OUR MODEL USING STACKING ENSEMBLE

- Importing, detailed information on the new dataset.
- Apply transforms to the new data similar to the training dataset by dropping off some redundant features and encode the categorical features to numerical ones
- Import joblib
- Load the saved base models and train with the best parameters for each algorithm
- Generate meta-features for training the meta-learner.
- Save the trained meta-learner to a file using joblib.dump
- Make predictions with the base learners (Logistic Regression and XGBoost)
- Create meta-features for the testing data and Load the meta-learner model
- Make final predictions using the meta-learner
- Apply cross-validation with stacking

METRIC EVALUATION OF STACKING ENSEMLE (CONFIGURATION)

METRIC EVALUATION FOR CLASS 0 AND CLASS 1

ACCURACY: 0.9745

- **PRECISION (CLASS 0): 0.9666,** **PRECISION (CLASS 1): 0.9827**
- **RECALL (CLASS 0): 0.9830,** **RECALL (CLASS 1): 0.9660**
- **F1 SCORE (CLASS 0): 0.9747,** **F1 SCORE (CLASS 1): 0.9743**

AUC-ROC: 0.9925

- **CONFUSION MATRIX:**

```
[[694  12]
 [ 24  682]]
```

Analysis and Conclusion:

Overall Performance: The stacking ensemble has provided an exceptional level of performance, with high precision, recall, and F1 scores for both classes, as well as an almost perfect AUC-ROC score. This means that the model is very reliable for predicting employee attrition, which is challenging in imbalanced datasets.

Class Imbalance Handling: Both class 0 (stay) and class 1 (leave) are handled well. The model has a good balance of predicting those who will leave (class 1) with high recall (96.60%) and precision (98.27%), which is crucial in this context since identifying employees who are likely to leave is the main goal.

Trade-off Between Precision and Recall: The model achieves an excellent balance between precision and recall for both classes, particularly for the minority class (class 1), which is usually harder to predict. This indicates that the ensemble method has successfully learned from the base models and meta-learner to optimize performance across the classes. In summary, the stacking ensemble method is highly effective for this imbalanced dataset and is well-suited for predicting employee attrition. It strikes a good balance between minimizing false positives (predicting someone will leave when they won’t) and false negatives (failing to predict someone will leave), which is critical in employee retention strategies.

IMPACT OF STACKING ENSEMBLE ON THE 7 ALGORITHMS

```
Confusion Matrix for Logistic Regression:
[[49.2917847       0.7082153]
 [ 1.9121813      48.0878187]]

Classification Report for Logistic Regression:
             precision    recall  f1-score   support

           0       0.96      0.99      0.97       706
           1       0.99      0.96      0.97       706

    accuracy                           0.97      1412
   macro avg       0.97      0.97      0.97      1412
weighted avg       0.97      0.97      0.97      1412

AUC/ROC for Logistic Regression:
0.9737960339943342
```

```
Confusion Matrix for Random Forest:
[[49.07932011      0.92067989]
 [ 1.62889518     48.37110482]]

Classification Report for Random Forest:
             precision    recall  f1-score   support

           0       0.97      0.98      0.97       706
           1       0.98      0.97      0.97       706

    accuracy                           0.97      1412
   macro avg       0.97      0.97      0.97      1412
weighted avg       0.97      0.97      0.97      1412

AUC/ROC for Random Forest: 0.9745042492917846
```

IMPACT OF STACKING ENSEMBLE ON THE 7 ALGORITHMS

```
Confusion Matrix for XGBoost:
[[48.0878187    1.9121813 ]
 [ 2.05382436   47.94617564]]

Classification Report for XGBoost:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       706
           1       0.96      0.96      0.96       706

    accuracy                           0.96      1412
   macro avg       0.96      0.96      0.96      1412
weighted avg       0.96      0.96      0.96      1412

AUC/ROC for XGBoost: 0.9603399433427763
```

```
Confusion Matrix for Support Vector Classifier:
[[49.15014164    0.84985836]
 [ 1.98300283   48.01699717]]

Classification Report for Support Vector Classifier:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       706
           1       0.98      0.96      0.97       706

    accuracy                           0.97      1412
   macro avg       0.97      0.97      0.97      1412
weighted avg       0.97      0.97      0.97      1412

AUC/ROC for Support Vector Classifier: 0.9716713881019830
```

IMPACT OF STACKING ENSEMBLE ON THE 7 ALGORITHMS

```
Confusion Matrix for Stochastic Gradient Descent Classifier:
[[48.58356941        1.41643059]
 [ 2.26628895        47.73371105]]

Classification Report for Stochastic Gradient Descent Classifier:
              precision    recall  f1-score   support

           0       0.96      0.97      0.96       706
           1       0.97      0.95      0.96       706

    accuracy                           0.96      1412
   macro avg       0.96      0.96      0.96      1412
weighted avg       0.96      0.96      0.96      1412

AUC/ROC for Stochastic Gradient Descent Classifier:
0.9631728045325779
```

```
Confusion Matrix for Gaussian Naive Bayes:
[[48.79603399       1.20396601]
 [ 1.77053824       48.22946176]]

Classification Report for Gaussian Naive Bayes:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       706
           1       0.98      0.96      0.97       706

    accuracy                           0.97      1412
   macro avg       0.97      0.97      0.97      1412
weighted avg       0.97      0.97      0.97      1412

AUC/ROC for Gaussian Naive Bayes: 0.9702549575070821
```

IMPACT OF STACKING ENSEMBLE ON THE 7 ALGORITHMS

```
Confusion Matrix for Decision Tree Classifier:
[[49.15014164      0.84985836]
 [ 2.33711048        47.66288952]]

Classification Report for Decision Tree Classifier:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97       706
           1       0.98      0.95      0.97       706

    accuracy                           0.97      1412
   macro avg       0.97      0.97      0.97      1412
weighted avg       0.97      0.97      0.97      1412

AUC/ROC for Decision Tree Classifier: 0.9681303116614731
```

Analysis and Conclusion:

The output shows performance results for seven tuned models (Logistic Regression, Random Forest, XGBoost, Support Vector Classifier, Stochastic Gradient Descent Classifier, Gaussian Naive Bayes, and Decision Tree Classifier) used in a stacking ensemble method to predict a classification task.

Performance Overview

Across all models, the performance metrics are consistently strong, with accuracies around 96-97%, and high precision, recall, and F1 scores. This indicates that each model individually performs well, likely contributing positively to the stacking ensemble.

The confusion matrices reveal that:

Class 0 (Majority Class): All models consistently show low false positive rates (misclassifying class 1 as class 0), with errors generally ranging from 0.7% to 1.9%. Random Forest and Logistic Regression perform exceptionally well in this aspect.

Class 1 (Minority Class): The models show a similar trend, with false negatives (misclassifying class 0 as class 1) ranging from 1.6% to 2.3%. Logistic Regression, Random Forest, and Support Vector Classifier have fewer false negatives than XGBoost and SGD.

Comparative Evaluation Performance on stack ensemble technique class 1 vs individual tuned/optimized algorithms class 1 (Attrition)

This compares the performance of a stacked ensemble technique and individual tuned/optimized algorithms in predicting employee attrition. The ensemble technique outperforms individual tuned algorithms in:

1. Accuracy: Ensemble (96%-97%) vs. Individual algorithms (61%-83%)
2. Precision (Class 1 - Attrition): Ensemble (0.96-0.99) vs. Individual algorithms (0.30-0.42)
3. Recall (Class 1 - Attrition): Ensemble (0.95-0.98) vs. Individual algorithms (0.28-0.77)
4. F1-Score (Class 1 - Attrition): Ensemble (0.96-0.97) vs. Individual algorithms (0.33-0.54)
5. AUC/ROC: Ensemble (0.96-0.97) vs. Individual algorithms (0.59-0.76)

The ensemble technique provides more consistent and accurate predictions, making it a more reliable approach for HR decision-making, especially in predicting employee turnover or identifying high-potential candidates for promotions. Additionally, the ensemble technique handles class imbalance more effectively, which is critical in HR settings where false negatives (missed attrition cases) are costly.

Comparative Evaluation Performance on stack ensemble technique class 0 vs individual tuned/optimized algorithms class 0 (Non – Attrition)

This compares the performance of a stacked ensemble technique and optimized algorithms in predicting class 0 (non-attrition). The metrics used are precision, recall, F1-score, and AUC/ROC. The results show that the stacked ensemble consistently outperforms the optimized algorithms in all metrics, demonstrating better accuracy, reliability, and discrimination between non-attrition and attrition cases.

Key findings:

- Precision: Stacked ensemble (0.95-0.97) vs. Optimized algorithms (0.85-0.94)
- Recall: Stacked ensemble (0.97-0.99) vs. Optimized algorithms (0.56-0.95)
- F1-Score: Stacked ensemble (0.96-0.97) vs. Optimized algorithms (0.70-0.90)
- AUC/ROC: Stacked ensemble (0.96-0.97) vs. Optimized algorithms (0.61-0.76)

The stacked ensemble shows superior performance in predicting non-attrition, with higher precision, recall, F1-score, and AUC/ROC values. This indicates that the ensemble is better at identifying true non-attrition cases while minimizing errors, making it a more reliable approach for predicting class 0 (non-attrition).

Comparative Evaluation of the Confusion Metric Performances on stack ensemble technique vs individual tuned/optimized algorithms

- Stacked Ensemble: Across models, the number of misclassified instances (both false positives and false negatives) is consistently low. For instance, for Logistic Regression in the ensemble, there are approximately 2 false negatives and 1 false positive, which is highly efficient for HR decision-making.
- Tuned/Optimized Algorithms: For individual models, the number of misclassified instances is notably higher. For example, the Random Forest Classifier shows 13 false negatives and 7 false positives, and Gaussian Naive Bayes shows an even more imbalanced misclassification, with 35 false positives.

Analysis and Conclusion:

The stacking ensemble approach has achieved excellent results by combining the strengths of multiple algorithms, including Logistic Regression, XGBoost, and others. The ensemble has demonstrated robust predictive performance, generalizing well on test data. The success of the ensemble can be attributed to the complementary strengths of its base models, with Logistic Regression and XGBoost providing strong generalizations, and other models offering nuanced decision boundaries. With each base model performing well, the ensemble is likely to be robust and outperform individual models. Overall, the stacking ensemble approach has yielded a powerful and generalizable predictive model.

RECOMMENDATION

Recommendations for Reducing Employee Attrition:

Focus on retaining experienced employees
Offer competitive compensation
Promote work-life balance
Provide recognition and career growth opportunities
Enhance environmental satisfaction
Invest in training and development
Provide managerial support
Boost employee engagement
Leverage predictive modeling
Continuously monitor attrition trends

RECOMMENDATION

Recommendations for Predictive Modeling of Employee Attrition Using Stacking Ensemble:

Stacking ensemble method combine the strengths of multiple models to give a synergy effect on the metrics

Strategic Benefits for HR:

Proactive retention

Optimized resource allocation

Data-driven insights

Improved employee engagement

Handling imbalanced data which is common with the HR

Next Steps for Implementation:

Model deployment

Incorporate deep learning models into the ensemble to capture complex relationships in the data.

Apply natural language processing (NLP) on unstructured data such as employee feedback or performance reviews.

Continuous monitoring and tuning

Integrate employee engagement surveys

By implementing these recommendations, organizations can reduce employee attrition, improve employee satisfaction, and enhance overall performance.

THANK YOU