Employee Attrition Prediction Analysis Report

1. Introduction:

Employee attrition is a critical concern for organizations as it can lead to decreased productivity, loss of institutional knowledge, and increased recruitment costs. Predicting employee attrition using machine learning techniques can help organizations identify at-risk employees and take proactive measures to retain them. In this report, we summarize the findings and insights gained from analyzing the Employee Attrition dataset.

2. Dataset Analysis and Preprocessing:

• Dataset Description: The dataset contains various attributes related to employee demographics, job roles, satisfaction levels, performance ratings, etc., along with a target variable indicating whether an employee has left the company (Yes or No).

• Data Exploration: We explored the dataset to understand its structure, features, and distribution. This involved examining descriptive statistics, checking for missing values, and visualizing relationships between variables.

• Preprocessing Steps: We handled missing values, encoded categorical variables using one-hot encoding, and split the dataset into training and testing sets for model development
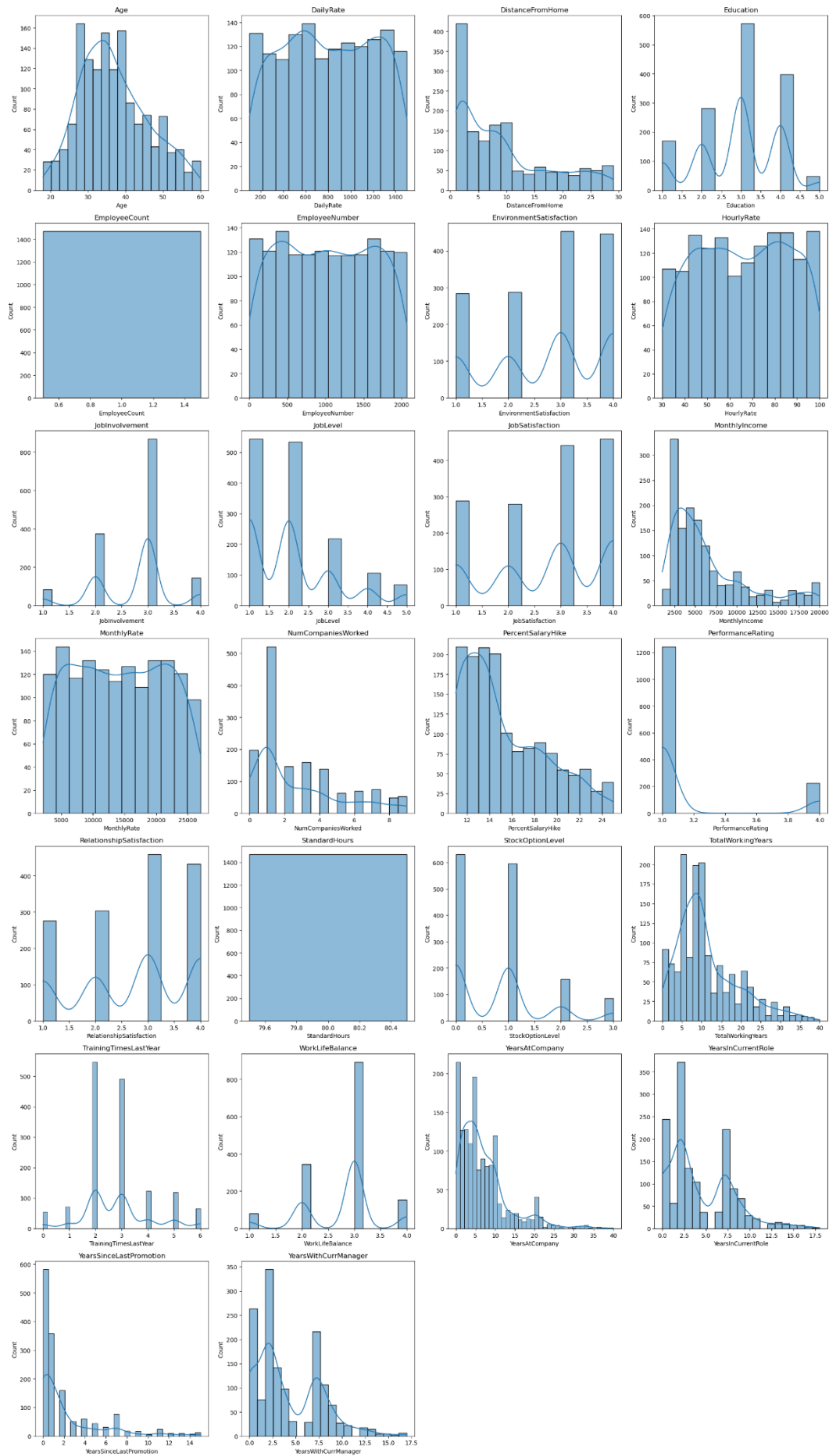
2.1 Data Analysis
INTERPRETATIONS:

Boxplots of Selected Numerical Features
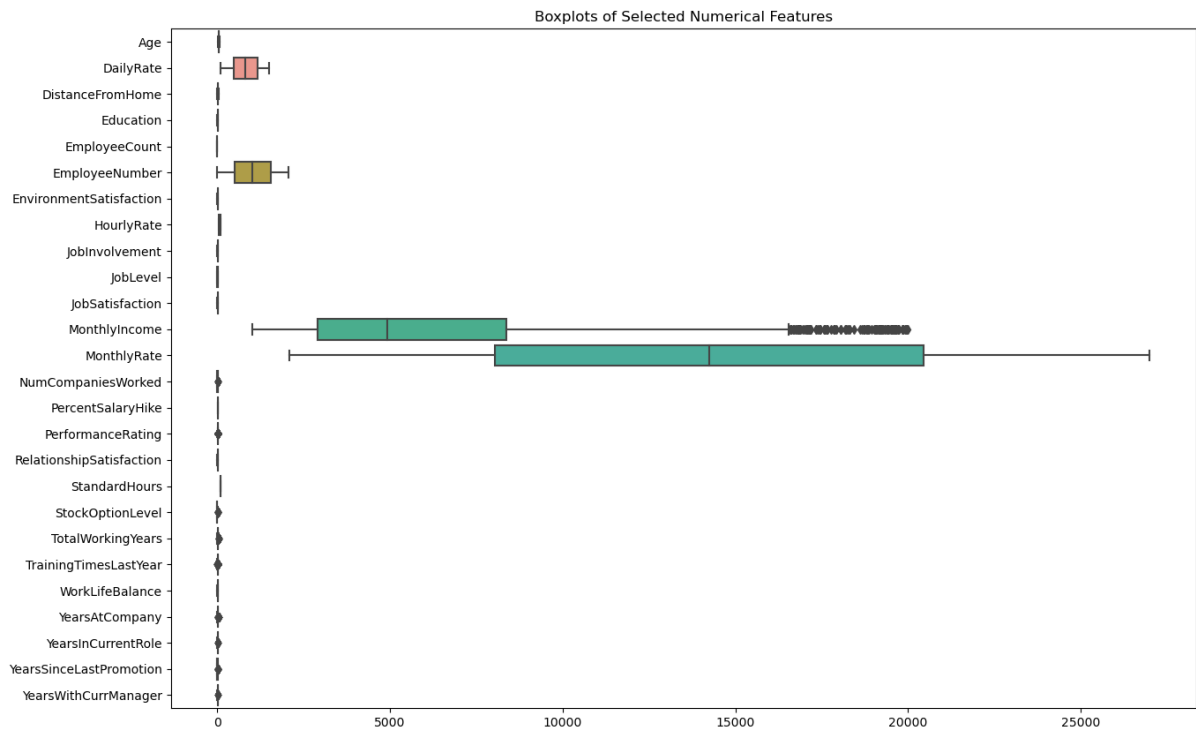
The dark points are known as Outliers. Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

Boxplot Insights: -

- Sometimes outliers may be an error in the data and should be removed. In this case these points are correct readings yet they are different from the other points that they appear to be incorrect.

- The best way to decide wether to remove them or not is to train models with and without these data points and compare their validation accuracy.

- So we will keep it unchanged as it won't affect our model.

A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases.

Correlation can have a value:

- 1 is a perfect positive correlation

- 0 is no correlation (the values don't seem linked at all)

- -1 is a perfect negative correlation
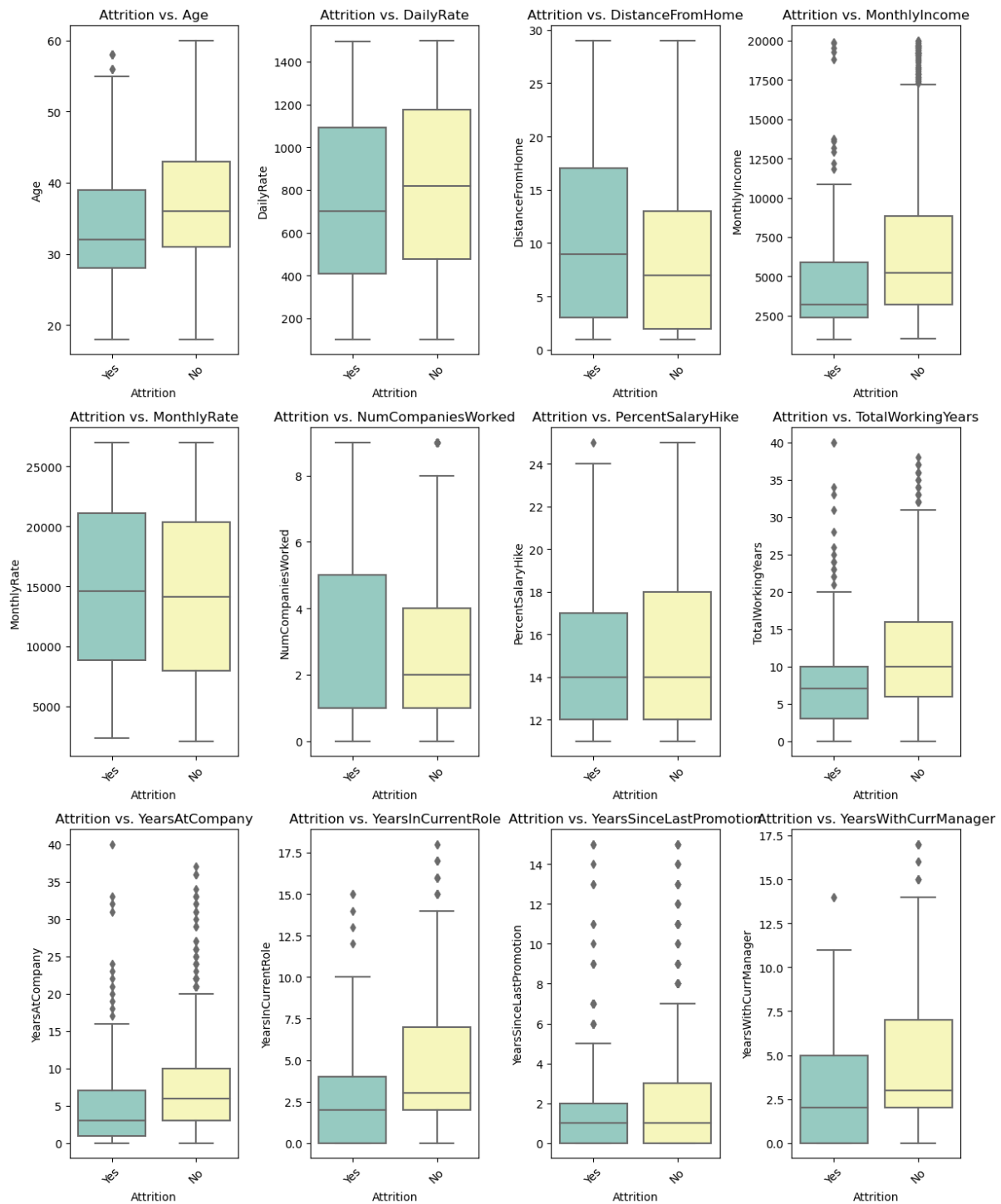
3. Model Development: We experimented with various machine learning algorithms for binary classification, including Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, Decision Tree, K-Nearest Neighbors (KNN) and Naive Bayes.

1. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

2. Random Forest Classifier

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest." It can be used for both classification and regression problems in R and Python.

Random Forest and Decision Tree Algorithm are considered best for the data that has outliers.

4. Model Evaluation and Optimization:

• We evaluated each model's performance using metrics such as accuracy, precision, recall, and F1-score on the test data.

• Insights were gained into the significance of various features and their impact on predicting employee attrition.

6. **Insights Gained:**

Factors such as job satisfaction, work-life balance, and job role were identified as significant predictors of employee attrition.

Random Forest Model with 0.99 accuracy is the best model

So we gonna save our model with Random Forest Algorithm.

7. **Recommendations for Reducing Employee Attrition:**

• Implement proactive measures based on predictive models to identify at-risk employees and intervene early.

• Focus on improving job satisfaction, work-life balance, and career development opportunities to increase employee retention.

• Regularly retrain and update predictive models with new data to ensure effectiveness in capturing evolving trends in attrition.

8. Conclusion: Predicting employee attrition is a complex but important task for organizations to manage their workforce effectively. By leveraging machine learning techniques and analyzing relevant factors, organizations can gain valuable insights into attrition patterns and take proactive steps to reduce attrition rates, improve employee satisfaction, and enhance overall organizational performance

Submitted by: ASHMI DATTA