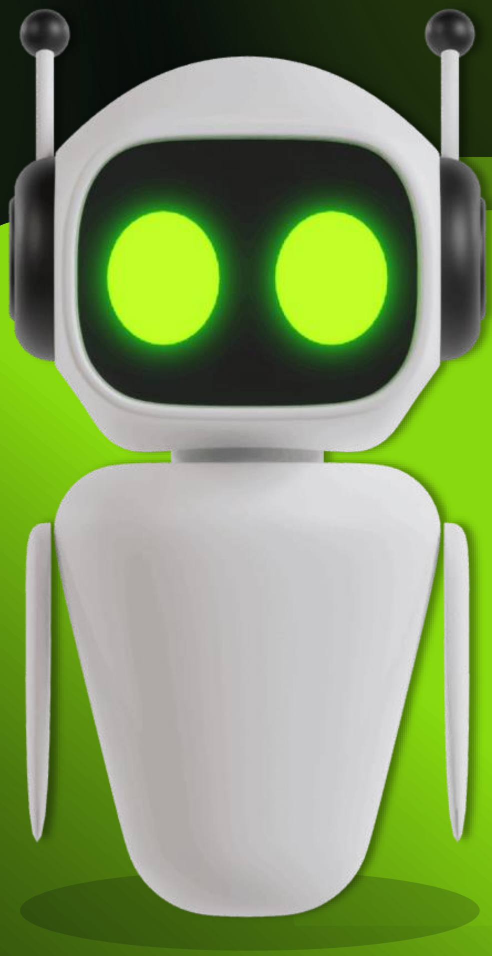
European Business Institute
LUXEMBOURG

EMPLOYEE ATTRITION PREDICTION

WITH MACHINE LEARNING & ARTIFICIAL INTELLIGENCE



REPORT BY:
IDRIS AGANG KEDISANG

Introduction

Employee attrition means the gradual reduction of the number of employees within an organization. It is in simplest of forms, the natural process through which employees leave the workforce.

The report is on research carried out to explore reasons employees leave organizations and use these insights to predict when and why an employee might leave the organizations so that resources can be directed to such employees and reduce employee attrition rate hence improving employee turnover.

Problem Statement

In a world where employee retention is one of the key performance indicators that Mobile Network Operators (MNOs) in Botswana keep track of, it is also equally important for them to monitor employee attrition. It is important for organizations to answer the below questions in order to anticipate and predict employee attrition:

01 Will an employee leave the company or not?

02 When will it occur?

03 Why it may happen?

The task at hand is to now develop and deploy Machine learning models that have the capacity to answer the above questions and predict employee attrition

EBU
European Business Institute
LUXEMBOURG

Dataset Details

The dataset used was acquired from Kaggle and it is from IBM HR Analytics Attrition datasets. The dataset consists of 1 470 rows being employee survey records as well 35 columns being features. The reason behind using this dataset is the lack of HR performance survey datasets within the local business environment.

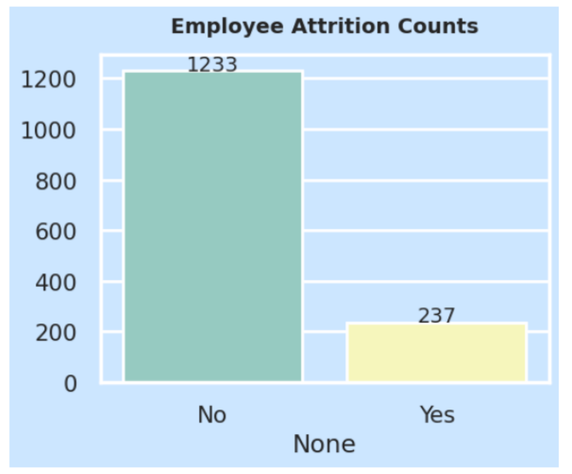
FEATURE	DATA TYPE
Age	int64
Attrition	object
Business Travel	object
DailyRate	int64
Department	object
DistanceFromHome	int64
Education	int64
EducationField	object
EmployeeCount	int64
EmployeeNumber	int64
EnvironmentSatisfaction	int64
Gender	object
HourlyRate	int64
JobInvolvement	int64
JobLevel	int64
JobRole	object
JobSatisfaction	int64
MaritalStatus	object
MonthlyIncome	int64
MonthlyRate	int64
NumCompaniesWorked	int64
Over18	object
OverTime	object
PercentSalaryHike	int64
PerformanceRating	int64
RelationshipSatisfaction	int64
StandardHours	int64
StockOptionLevel	int64
TotalWorkingYears	int64
TrainingTimesLastYear	int64
WorkLifeBalance	int64
YearsAtCompany	int64
YearsInCurrentRole	int64
YearsSinceLastPromotion	int64
YearsWithCurrManager	int64

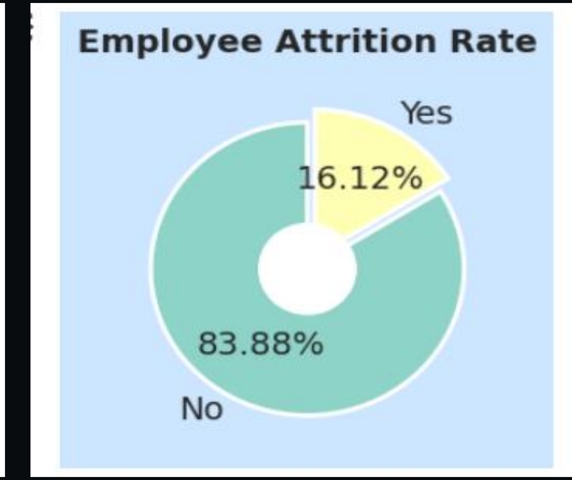
Data Exploration

- All employees are adults over the age of 18, evidenced through the minimum age of the Age attribute
- Standard deviation for EmployeeCount and StandardHours is 0.0, implying that the values for these attributes are the same.
- EmployeeNumber is unique to each employee.
- The above attributes did not have any meaningful impact on our analysis so we dropped them from the dataset.

EBU
European Business Institute
LUXEMBOURG

Data Exploration

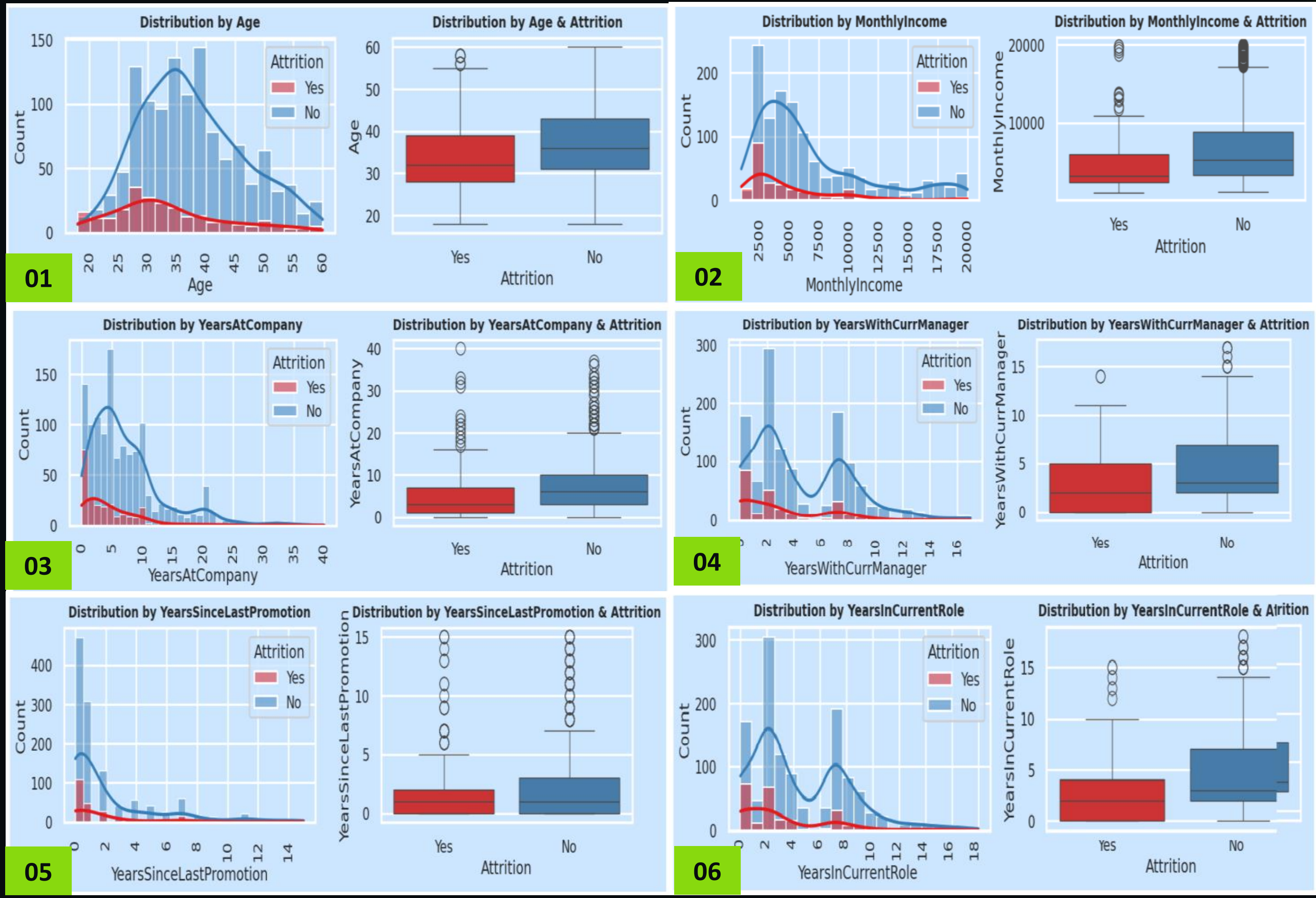




The dataset has an Attrition Rate of 16.12%

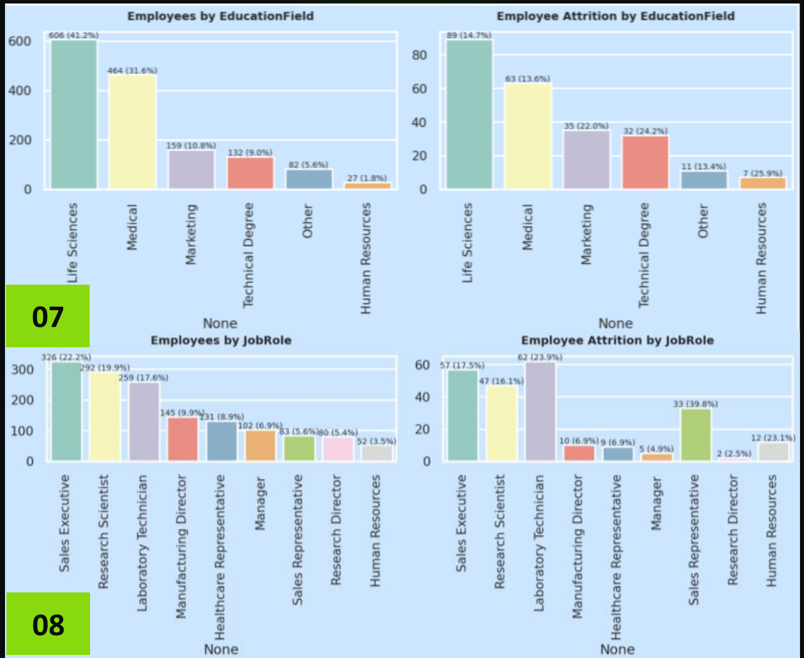
Data Exploration on Numerical Data Types

The below visualizations depicts the distribution of employees based on each numerical variables as well as the distribution by each variable and attrition.



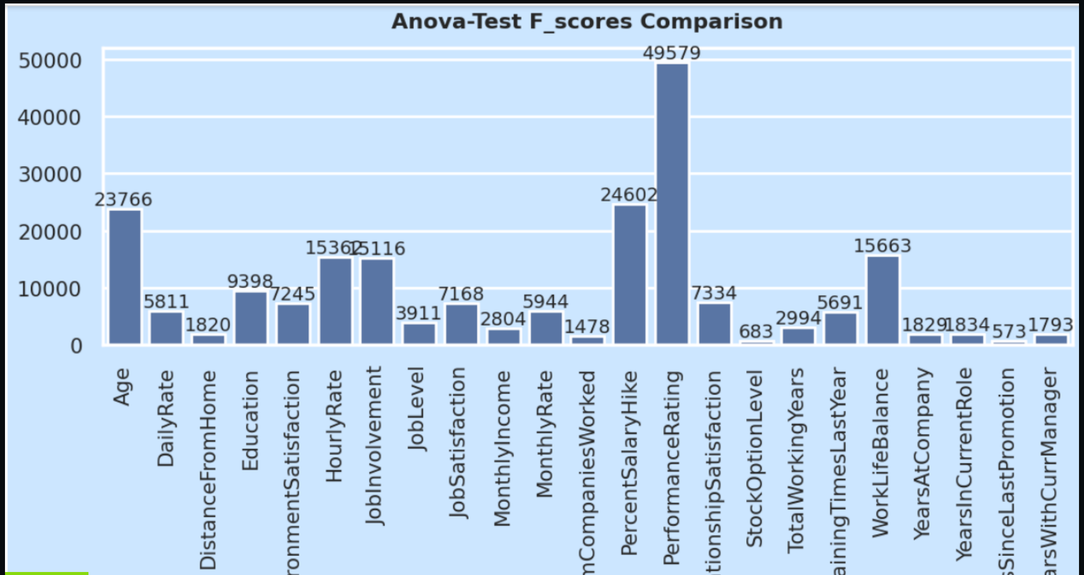
Data Exploration on Categorical Data Types

The below visualizations depicts the distribution of employees based on each categorical variables as well as the attrition rate for each categorical variable.

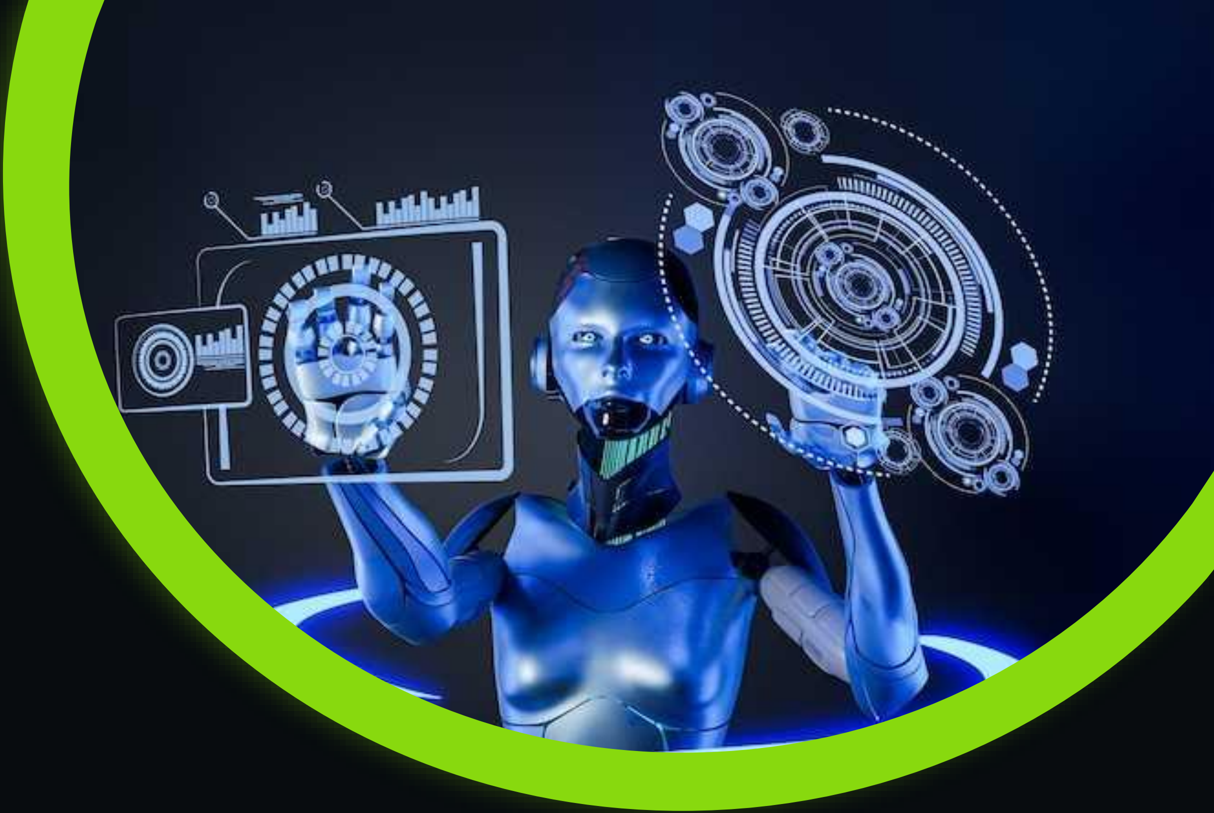


Feature Importance

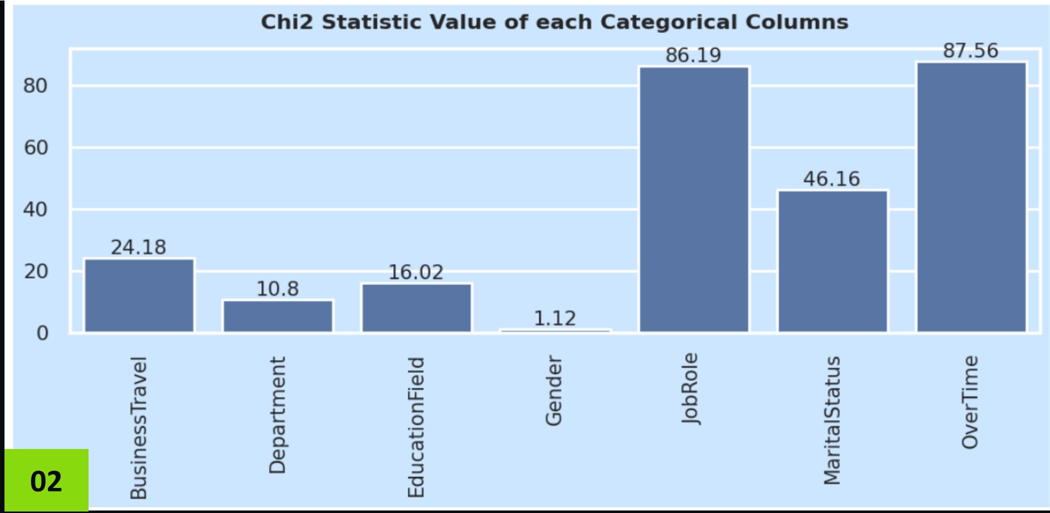
The below visualizations shows the test scores of an ANOVA test which analyses the importance of each feature



01



The below visualizations shows the test scores of a Chi2 Statistic test which analyses the importance of each categorical feature



02

Selection of Machine Learning Model

The idea is to test and compare the performances of the following machine learning techniques and choose the best fitting and performing classification model:

- Decision tree classifier
- Logistics Regression
- Naïve Bayes
- Random Forest
- XG_Boost

Model Performance Results and Analysis

The table below presents the performance metrics of various classification models tested for the employee attrition use case. Here is an interpretation and discussions based on the performance evaluation results:

Algorithm	Training Score	Testing Score	Precision	Recall	ROC_AUC	F1-Score	Kappa Score	G_Mean
Logistic Regression	86.1	84.8	0.839	0.856	0.929	0.847	0.696	0.848
Naïve Bayes	75.2	75.7	0.697	0.897	0.876	0.784	0.515	0.758
Decision Tree	100.0	85.0	0.819	0.893	0.850	0.854	0.700	0.850
Random Forest	100.0	93.3	0.977	0.885	0.973	0.929	0.866	0.932
XG_Boost	100.0	92.1	0.947	0.889	0.967	0.917	0.841	0.920

INTERPRETATION

01

- **Training Score**: Represents the model's performance on the training data used to fit the model.
- **Testing Score**: Represents the model's performance on unseen data, indicating its generalizability.
- **Precision**: is the measurement of the proportion of predicted positives that are actually true positives.
- **Recall**: Measures the proportion of actual positives that are correctly identified by the model.
- **f1_Score**: Harmonic mean of precision and recall, balancing both aspects.
- **Kappa_Score**: Measures inter-rater agreement between the true labels and the model's predictions.
- **G_Mean**: Geometric mean of sensitivity and specificity, considering both true positives and negatives.

DISCUSSIONS

02

- **Overfitting**: Decision Tree and Random Forest achieve perfect training scores (100%), suggesting potential overfitting. Their testing scores are lower, indicating they might not generalize well to unseen data.
- **Balance**: While Random Forest has the highest testing score (93.32%), XGBoost closely follows with a slightly lower score (92.11%). However, XGBoost has a higher precision (0.9476) compared to Random Forest (0.9774), indicating it might be better at avoiding false positives.
- **Overall Performance**: Based on the combined metrics, XGBoost appears to be the best performing algorithm. It achieves a good balance between precision, recall, and other evaluation metrics, suggesting strong performance and generalizability.

Conclusion

Based on the model performance results data, XGBoost demonstrates the most balanced and generalizable performance among the presented algorithms. However, the choice of the optimal algorithm should be based on the specific requirements and priorities of the use case at hand. It's also crucial to consider potential limitations like overfitting and explore further evaluation techniques for a more comprehensive understanding of the models' capabilities