Project Report

On

"IBM HR Analytics for Employee Attrition and Performance Prediction"

Submitted
By

Shantanu Umrani
PRN: 1032212141

M. Tech. CSE (Data Science and Analytics**)**

SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY

Dr. Vishwanath Karad
MIT World Peace University, Pune

Academic Year 2021-202

ABSTRACT

In numerous software companies, there is a noticeable trend of employees resigning from their positions for various reasons. When skilled and valuable employees depart, it poses significant challenges for organizations in maintaining their operations. Consequently, it is crucial for companies to proactively identify and assess the causes of employee turnover and formulate suitable strategies and actions to address this issue. IBM's HR Analytics Employee Attrition and Performance datasets are being utilized for this purpose. Missing values were dropped to give better insights in data analysis. ANOVA and Chi-Square tests were carried out during statistical analysis. Machine Learning algorithms such as Logistic Regression (92%), Random Forest (89%), Support Vector Machine (93%), XGBoost (100%), CatBoost (98%), AdaBoost (90%) and LightGBM (100%) were applied to understand, manage, and mitigate employee attrition. Comparison of model performance was plotted on ROC Curve using True-Positive and False-Positive Rate.

INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

Companies, both in India and other countries, face a formidable challenge in recruiting and retaining top talent, all while dealing with talent loss through attrition, whether due to industry downturns or voluntary turnover. Losing employees not only results in performance setbacks but also has long-term negative impacts on companies. This includes potential disruptions to productivity, work team dynamics, and social goodwill.

The success and competitiveness of any organization are highly dependent on its workforce, with employees serving as the essential backbone of the company. This study aims to identify employee attitudes, pinpoint the factors that contribute to their dissatisfaction within the organization, and understand the reasons behind their decisions to seek alternative employment opportunities.

By identifying and assessing the levels of employee attitudes, management can gain valuable insights into areas that require improvement and take necessary action to reduce attrition rates. This proactive approach is essential for sustaining a productive and harmonious work environment while enhancing the organization's overall performance and long-term success.

Fig 1.1 Skills in Human Resource [HR]

1.2 OBJECTIVE

1. Assess the degree of employee satisfaction regarding their job and working environment.
2. Identify the elements that contribute to employee dissatisfaction with the company's policies and guidelines.
3. Pinpoint the areas where the company is falling short or facing shortcomings.
4. Understand the underlying causes of attrition within the company.
5. Develop strategies and methods to minimize attrition rates within the organization.

1.3 MOTIVATION

This project originates from the possibility of enhancing employee contentment, cutting down expenses, elevating organizational efficiency, and fostering a favorable workplace atmosphere. It represents a chance to leverage data and analytics to enact substantial improvements that are advantageous to both employees and the entire organization.

CHAPTER 2: SYSTEM ARCHITECTURE

2.1 PROPOSED ARCHITECTURE (BLOCK DIAGRAM)

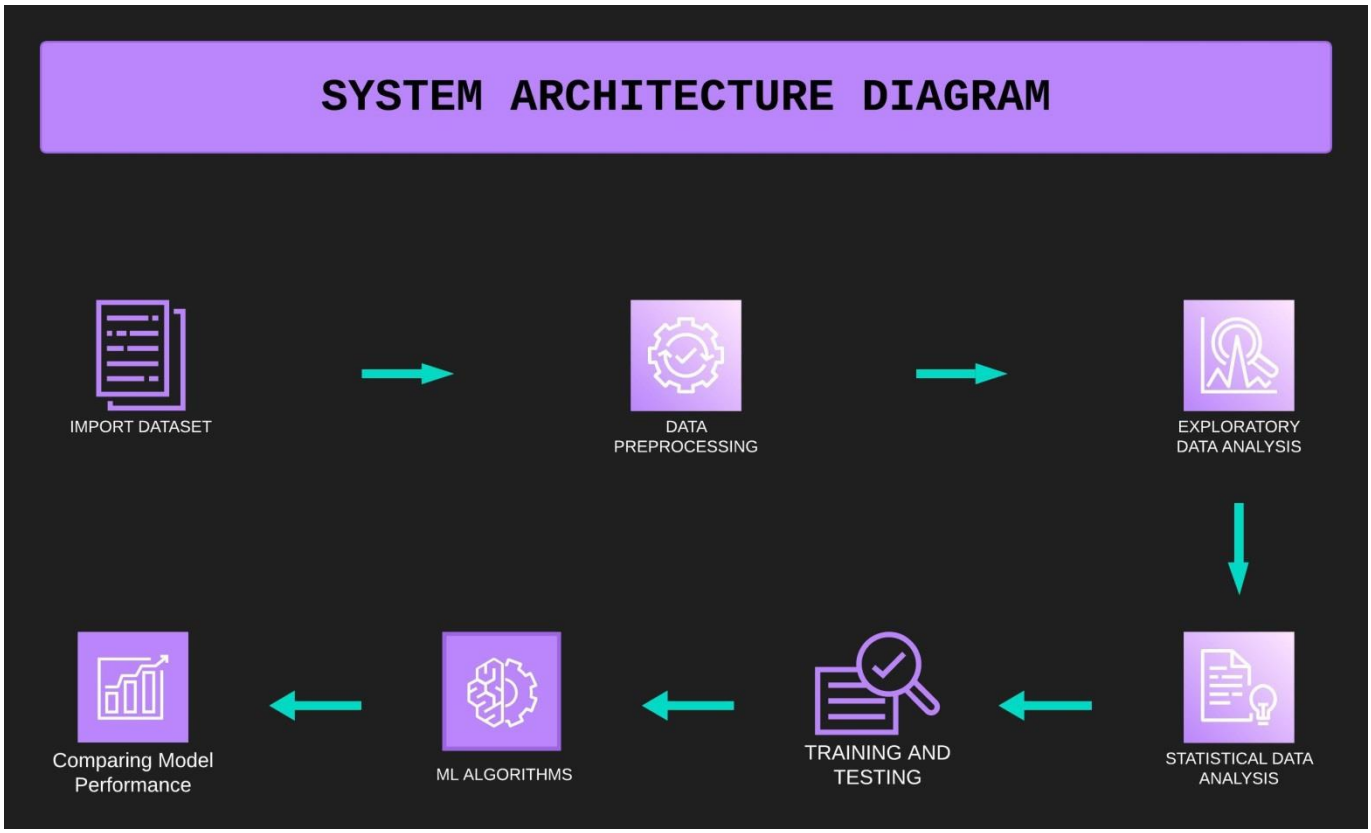

Fig.2.1 Proposed Architecture

The methodology for IBM HR Analytics Employee Attrition and Performance Prediction is as follows:-

- Input is taken by loading the ODIR dataset, which contains ocular
- Load the Dataset: The IBM HR Analytics Attrition Dataset is loaded using the pd.read_csv() function. The head() and info() methods are used to display the first few rows and get information about the dataset, respectively.
- Knowing the Dataset: Basic Information about the dataset is generated; numerical and categorical attributes are enlisted.

- Data Cleaning: Any missing values in the dataset are dropped using the dropna() method.
- Data Visualization: Matplotlib and Seaborn libraries are used to visualize the data.
- Statistical Analysis: The ANOVA Test is performed to analyze the Numerical Features' Importance in Employee Attrition, while the Chi-Square Test to Analyze the Categorical Feature Importance in Employee Attrition.
- Data Preprocessing: The target variable 'Attrition' is mapped to binary values (1 for 'Yes' and 0 for 'No'). Selected features are extracted from the dataset and one-hot encoded using the get_dummies() function.
- Splitting the Dataset: The dataset is split into training and testing sets using the train_test_split() method from scikit-learn.
- Implementing Machine Learning Algorithms: Logistic Regression, XGBoost, CatBoost, AdaBoost, LightGBM, Decision Tree, and Random Forest classifiers are initialized and trained using the training data.
- Model Evaluation: The accuracy score and confusion matrix are computed to evaluate the performance of each algorithm on the testing data.
- Results: The results, including the accuracy and confusion matrix, are printed for each algorithm.
- Model Performance Comparison: The hvPlot library is used to visualize the ROC curve diagram comparing the performance of all models used.

2.2 TECHNOLOGY USED

Technology that would be used in this project are:

Python Programming, Machine Learning, Data Analytics, Statistical Analytics.

2.3 DATASET

This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs. a random allocation of probability of attrition. IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. To use a matrix structure, we changed the model to reflect the following data.

Dataset Description:

During website session, browsing information about visited pages is collected and features are extracted as follows:

Table 1 – Numerical features used in the user attrition analysis model.

Feature Name	Feature Description	Min Value	Max Value	Std. Dev
Age	Age of employee	18	60	9.13
DailyRate	It is the billing cost for an individual's services for a single day	102	1499	403.50
DistanceFromHome	It is the distance between companyand home of the employee	1	29	8.10
Education	Education qualification of the employees of company	1	5	1.02
EmployeeCount	Count of employee	1	1	0.0

EmployeeNumber	It is a unique number that has been assigned to each current and former employee	1	2068	602.02
EnvironmentSatisfaction	It is all about an individual's feelings about the work environmentand organization culture.	1	4	1.09
HourlyRate	The amount of money that is paid to an employee for every hour worked	30	100	20.32
JobInvolvement	Job involvement refers to the degreeto which a job is central to a person'sidentity.	1	4	0.71
JobLevel	Job levels are categories of authority in an organization.	1	5	1.10
JobSatisfaction	Job satisfaction happens when an employee feels he or she is having job stability.	1	4	1.10
MonthlyIncome	Gross monthly income is the amountof income an employee earn in one month.	1009	19999	4707.95
MonthlyRate	If a monthly rate is set, employees should be paid in exchange for normal hours of work of a full-time worker.	2094	26999	7117.78
NumCompaniesWorked	Number of other companies the employee previously worked for	0	9	2.49
PercentSalaryHike	The amount a salary is increased of an employee in percentage	11	25	3.65
PerformanceRating	Rating means gauging and comparing the performance.	3	4	0.36
RelationshipSatisfaction	It is the rate of satisfaction between Employer employee relationship.	1	4	1.08

Table 1: Shows the numerical features along with their statistical parameters.

Table 2 – Categorical Features used in the User Attrition Analysis Model.

Feature Name	Feature Description	Number of Categorical Values
Attrition	Attrition in business describes a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign, [NOTE: Target Variable] (0=no, 1=yes)	2
BusinessTravel	Business travel is travel undertaken for work or business purposes, as opposed to other types of travel (1=No Travel, 2=Travel Frequently, 3=Travel Rarely)	3
Department	Consists three departments that contribute to the company's overall mission. (1=HR, 2=R&D, 3=Sales)	3
EducationField	Education field of the employees(1=HR, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=others, 6= Technical)	6
Gender	Gender of the employee (1=Female, 2=Male)	2
JobRole	These refer to the specific activities or work that the employee will perform. (1=HC Rep, 2=HR, 3=Lab Technician, 4=manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=sales Executive, 9= Sales Representative)	9
MaritalStatus	Marital Status of the employee (1=divorced, 2=married, 3=single)	3
Over18	(1=Yes, 2=No)	2
Overtime	(1=No, 2=Yes)	2

Table 2: Shows the categorical features along with their number of categories.

CHAPTER 3: IMPLEMENTATION

3.1 DATA EXPLORATION AND PROCESSING

Compute Size:

In first step, we try to understand the dataset's size and structure at a glance by computing it's size.

1] COMPUTING SIZE OF DATASET

In [5]:
```
# Print the shape of the DataFrame
print("The shape of data frame:", employee_data.shape)
# Print the length (number of rows) of the DataFrame
print("Number of Rows in the dataframe:", len(employee_data))
# Print the number of columns in the DataFrame
print("Number of Columns in the dataframe:", len(employee_data.columns))
```

```
The shape of data frame: (1470, 35)
Number of Rows in the dataframe: 1470
Number of Columns in the dataframe: 35
```

Fig.3.1.1 Compute Size of Dataset.

The code reveals that the "employee_data" DataFrame contains 1,470 rows and 35 columns, providing a quick overview of its size and structure.

Drop Columns:

In this step, we notice that 'EmployeeCount', 'Over18', 'StandardHours' have only one unique values and 'EmployeeNumber' has 1470 unique values. These features aren't useful for us, so we are going to drop those columns.

In [20]:
```
employee_data.drop(['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours'], axis="columns", inplace=True)
```

In [22]:
```
# Print the shape of the DataFrame
print("The shape of data frame:", employee_data.shape)
# Print the length (number of rows) of the DataFrame
print("Number of Rows in the dataframe:", len(employee_data))
# Print the number of columns in the DataFrame
print("Number of Columns in the dataframe:", len(employee_data.columns))
```

```
The shape of data frame: (1470, 31)
Number of Rows in the dataframe: 1470
Number of Columns in the dataframe: 31
```

Fig.3.1.2 Compute Size of Dataset after dropping columns.

The code reveals that the "employee_data" DataFrame now contains 1,470 rows and 31 columns, providing a quick overview of its size and structure after dropping few columns.

3.2 DATA VISUALIZATION

By analyzing employee data, we can identify factors that contribute to employee attrition, such as job satisfaction, compensation, and work-life balance. This information can be used to develop strategies to retain top talent and reduce turnover rates. HR analytics can help identify high-performing employees by analyzing data related to performance metrics, such as productivity, quality, and customer satisfaction. This information can be used to develop strategies to retain top talent and improve overall organizational performance.

1] VISUALIZING THE EMPLOYEE ATTRITION RATE.

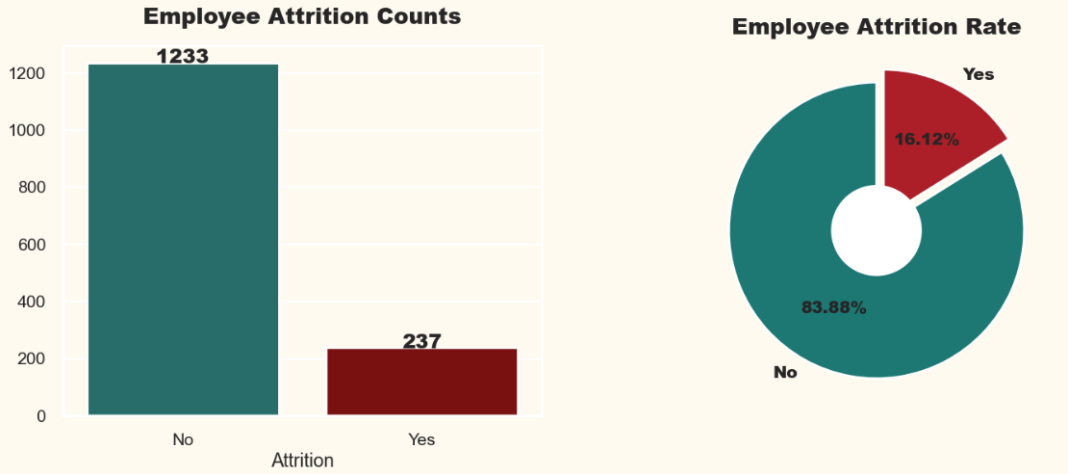

Fig.3.2.1: Visualizing Employee Attrition Rate.

Inference:

1. The employee attrition rate of this organization is 16.12%.
2. According to experts in the field of Human Resources, says that the attrition rate 4% to 6% is normal in organization.
3. So we can say the attrition rate of the organization is at a dangerous level.
4. Therefore the organization should take measures to reduce the attrition rate.

2] ANALYZING EMPLOYEE ATTRITION BY GENDER.

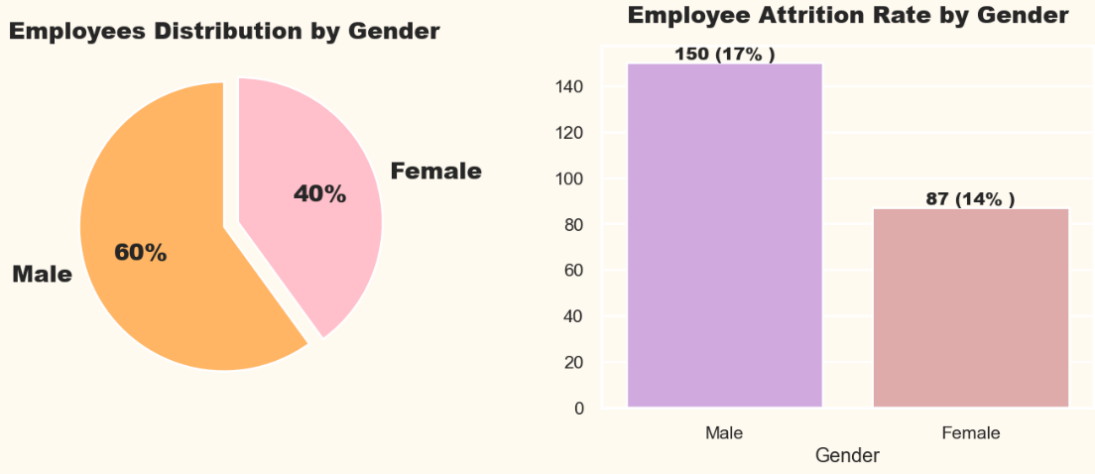

Fig.3.2.2: Analyzing Employee Attrition by Gender.

Inference:

1. The number of male employees in the organization accounts for a higher proportion than female employees by more than 20%.
2. Male employees are leaving more from the organization compared to female employees.

3] ANALYZING EMPLOYEE ATTRITION BY AGE.

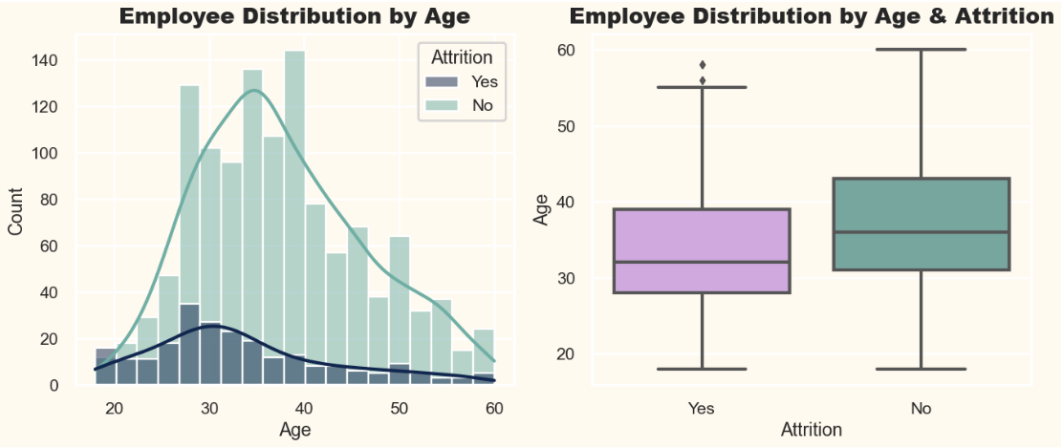

Fig.3.2.3: Analyzing Employee Attrition by Age.

Inference:

1. Most of the employees are between ages 30 to 40.
2. We can clearly observe a trend that as the age is increasing the attrition is decreasing.
3. From the boxplot we can also observe that the median age of employee who left the organization is less than the employees who are working in the organization.
4. Employees with young age leaves the company more compared to elder employees.

4] ANALYZING EMPLOYEE ATTRITION BY BUSINESS TRAVEL.

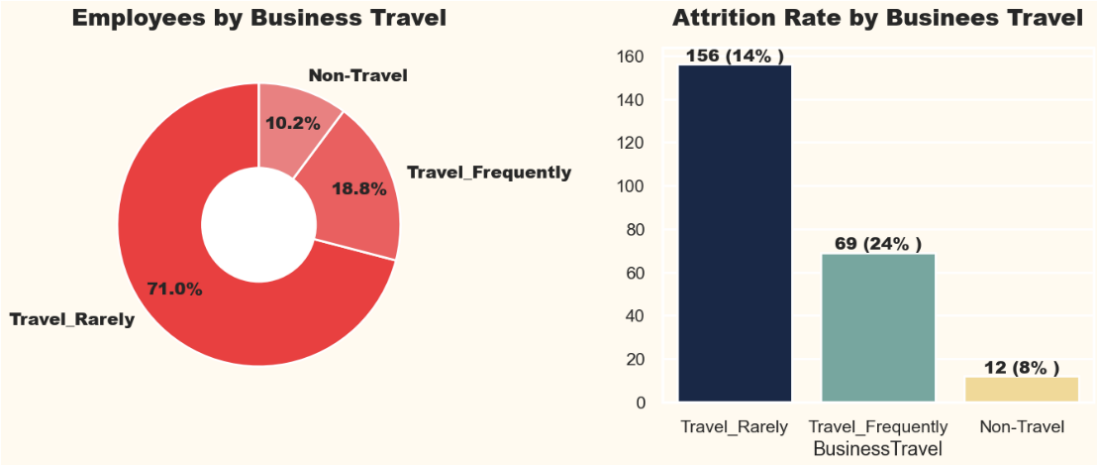

Fig.3.2.4: Analyzing Employee Attrition by Business Travel.

Inference:

1. Most of the employees in the organization Travel Rarely.
2. Highest employee attrition can be observed by those employees who Travels Frequently.
3. Lowest employee attrition can be observed by those employees who are Non-Travel.

5] ANALYZING EMPLOYEE ATTRITION BY DEPARTMENT.

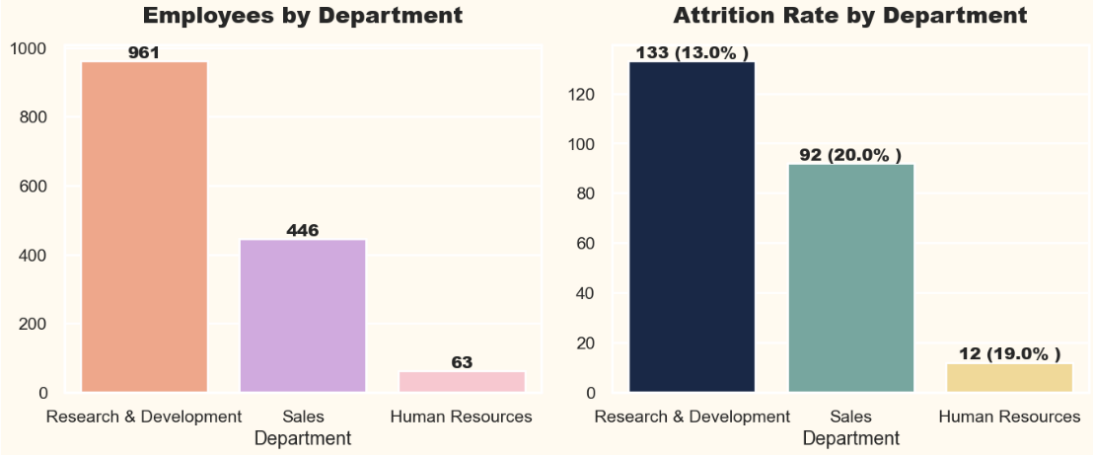

Fig.3.2.5: Analyzing Employee Attrition by Department.

Inference:

1. Most of the employees are from Research & Development Department.
2. Highest Attrition is in the Sales Department.
3. Human Resources Department Attrition rate is also very high.
4. Though of highest employees in Research & Development department there is least attrition compared to other departments.

6] ANALYZING EMPLOYEE ATTRITION BY DAILY RATE.

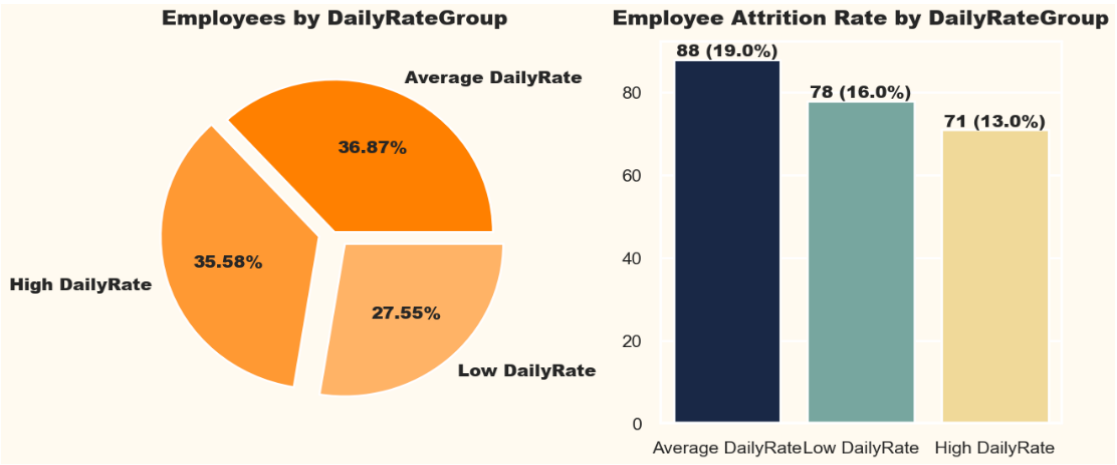

Fig.3.2.6: Analyzing Employee Attrition by Daily Rate.

Inference:

1. Employees with Average DailyRate & High Daily Rate are approximately equal.
2. But the attrition rate is very high of employees with average Daily Rate compared to the employees with High DailyRate.
3. The attrition rate is also high of employees with low DailyRate.
4. Employees who are not getting High Daily Rate are mostly leaving the organization.

7] ANALYZING EMPLOYEE ATTRITION BY DISTANCE FROM HOME.

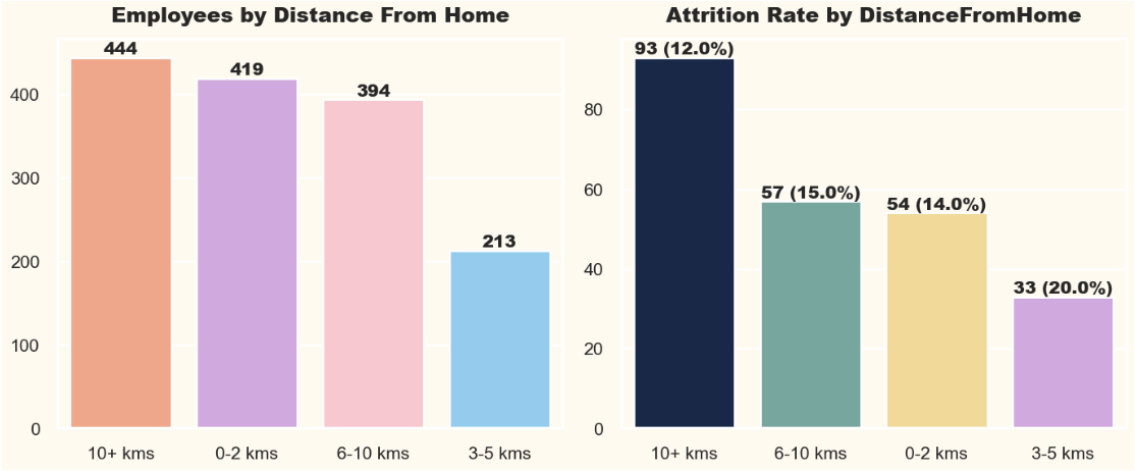

Fig.3.2.7: Analyzing Employee Attrition by Distance from Home.

Inference:

1. In the organization there is all kind of employees staying close or staying far from the office.
2. The feature Distance from Home doesn't follow any trend in attrition rate.
3. Employees staying close to the organization are mostly leaving compared to employees staying far from the organization.

8] ANALYZING EMPLOYEE ATTRITION BY EDUCATION.

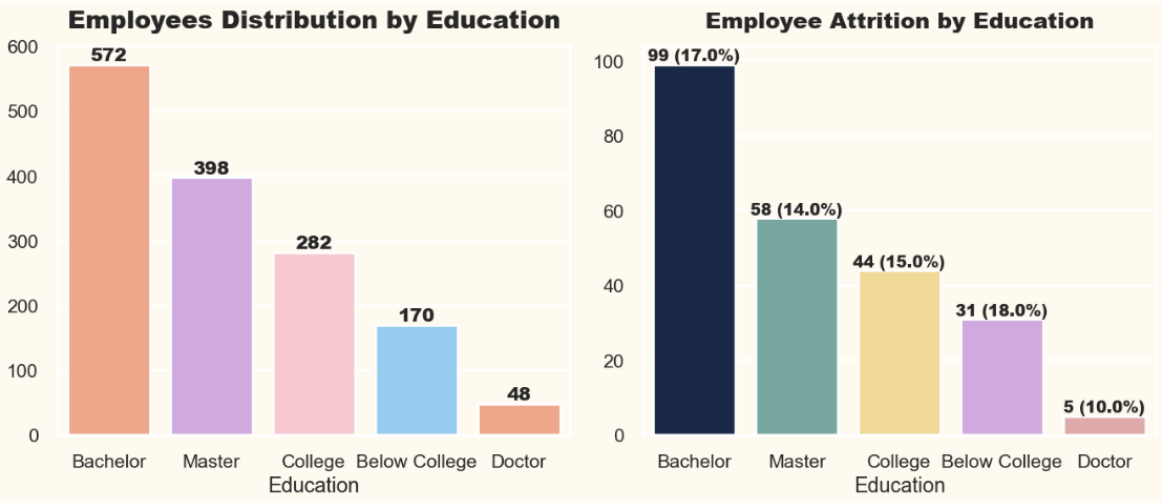

Fig.3.2.8: Analyzing Employee Attrition by Education.

Inference:

1. Most of the employees in the organization have completed Bachelors or Masters as their education qualification.
2. Very few employees in the organization have completed Doctorate degree as their education qualification.
3. We can observe a trend of decreasing in attrition rate as the education qualification increases.

9] ANALYZING EMPLOYEE ATTRITION BY EDUCATION FIELD.

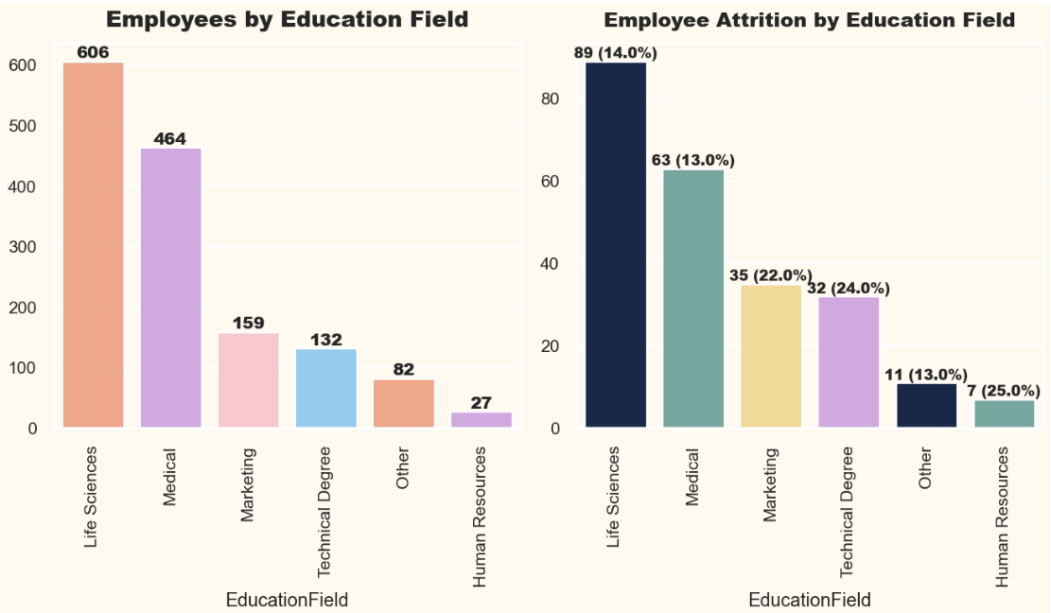

Fig.3.2.9: Analyzing Employee Attrition by Education Field.

Inference:

1. Most of the employees are either from Life Science or Medical Education Field.
2. Very few employees are from Human Resources Education Field.
3. Education Fields like Human Resources, Marketing, and Technical is having very high attrition rate.
4. This may be because of work load because there are very few employees in these education fields compared to education field with less attrition rate.

10] ANALYZING EMPLOYEE ATTRITION BY ENVIRONMENT SATISFACTION.

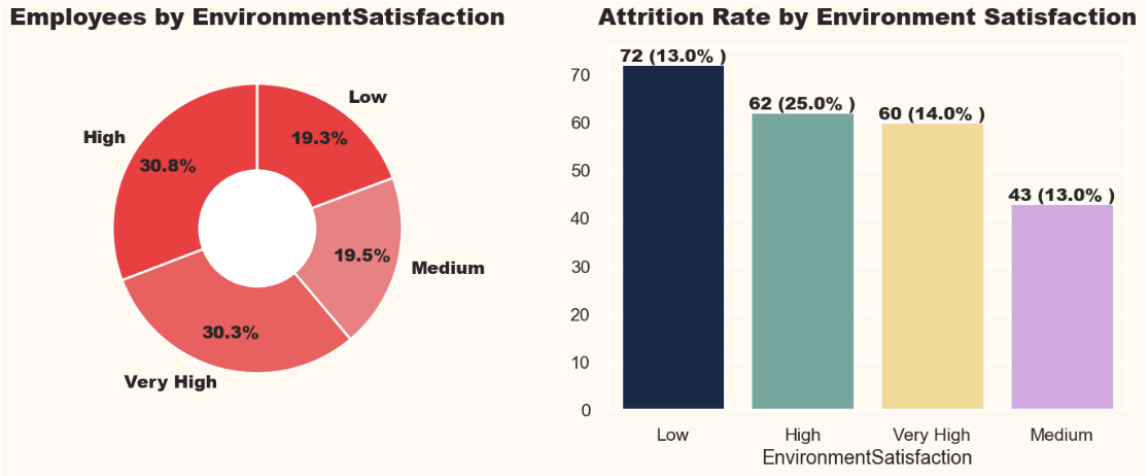

Fig.3.2.10: Analyzing Employee Attrition by Environment Satisfaction.

Inference:

1. Most of the employees have rated the organization environment satisfaction High & Very High.
2. Though the organization environment satisfaction is high still there's very high attrition in this environment.
3. Attrition Rate increases with increase in level of environment satisfaction.

11] ANALYZING EMPLOYEE ATTRITION BY JOB ROLES.

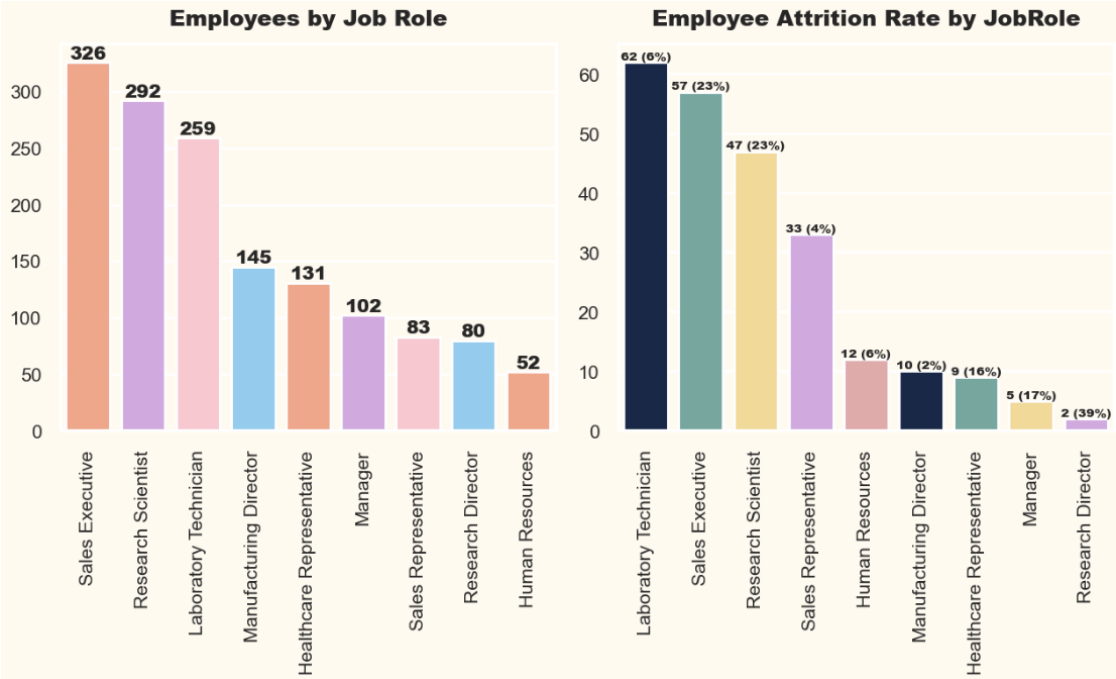

Fig.3.2.11: Analyzing Employee Attrition by Job Roles.

Inference:

1. Most employees are working as Sales executive, Research Scientist or Laboratory Technician in this organization.
2. Highest attrition rates are in sector of Research Director, Sales Executive, and Research Scientist.

12] ANALYZING EMPLOYEE ATTRITION BY JOB LEVEL.

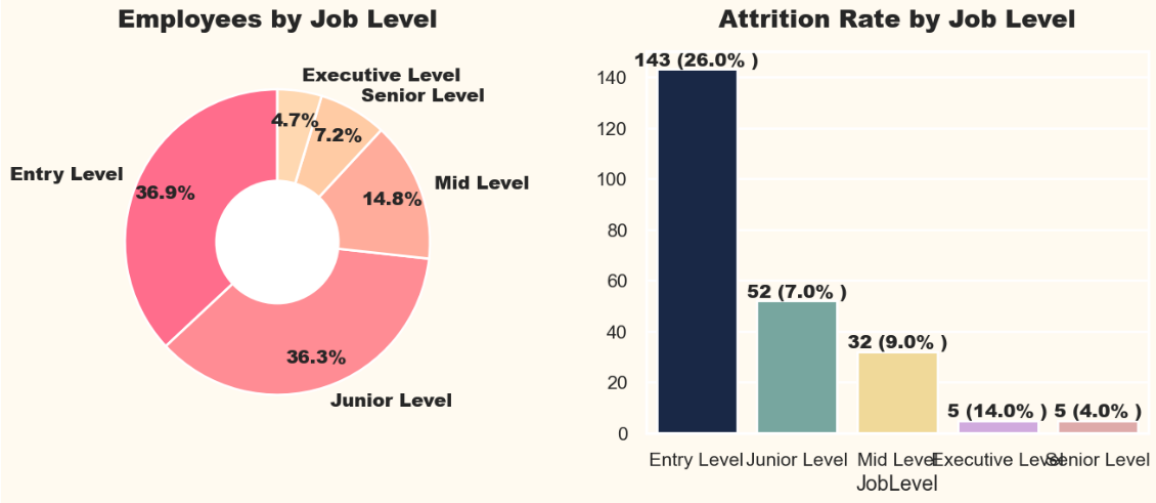

Fig.3.2.12: Analyzing Employee Attrition by Job Level.

Inference:

1. Most of the employees in the organization are at Entry Level or Junior Level.
2. Highest Attrition is at the Entry Level.
3. As the level increases the attrition rate decreases.

13] ANALYZING EMPLOYEE ATTRITION BY JOB SATISFACTION.

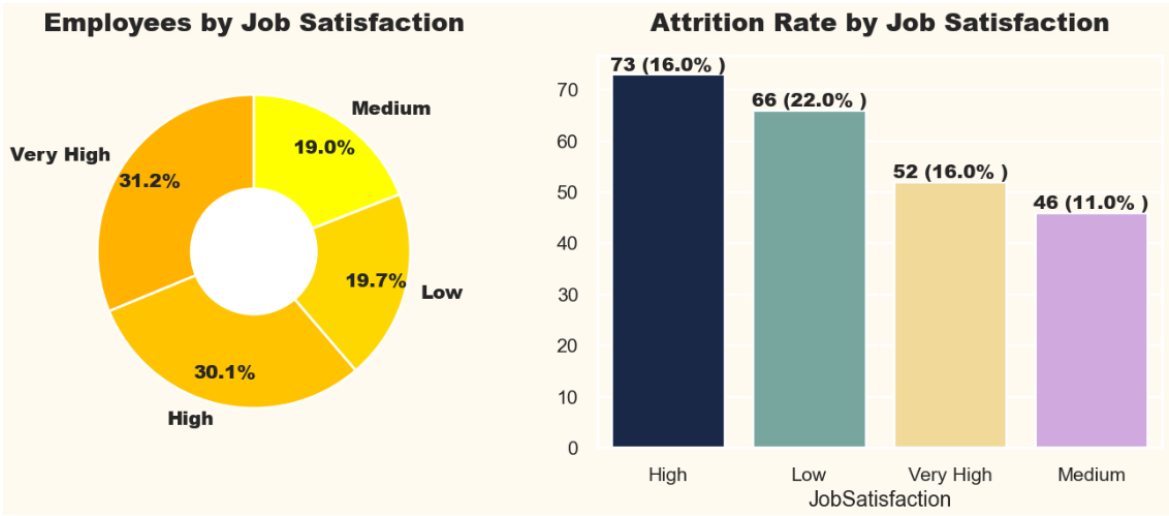

Fig.3.2.13: Analyzing Employee Attrition by Job Satisfaction.

Inference:

1. Most of the employees have rated their job satisfaction as high or very high.
2. Employees who rated their job satisfaction low are mostly leaving the organization.
3. All the categories in job satisfaction is having high attrition rate.

14] ANALYZING EMPLOYEE ATTRITION BY MARTIAL STATUS.

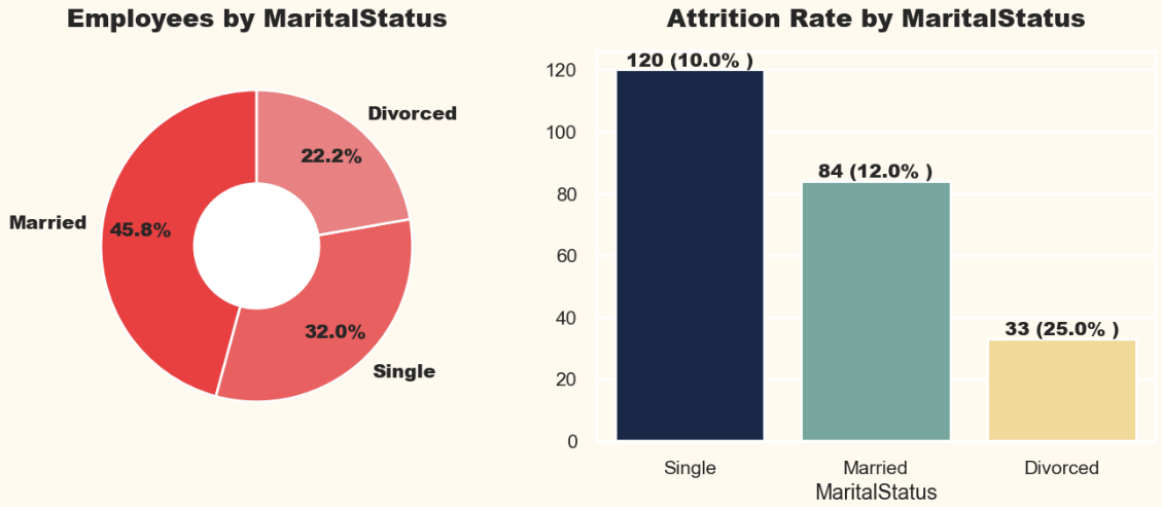

Fig.3.2.14: Analyzing Employee Attrition by Martial Status.

Inference:

1. Most of the employees are married in the organization.
2. The attrition rate is very high of employees who are divorced.
3. The attrition rate is low for employees who are single.

15] ANALYZING EMPLOYEE ATTRITION BY MONTHLY INCOME.

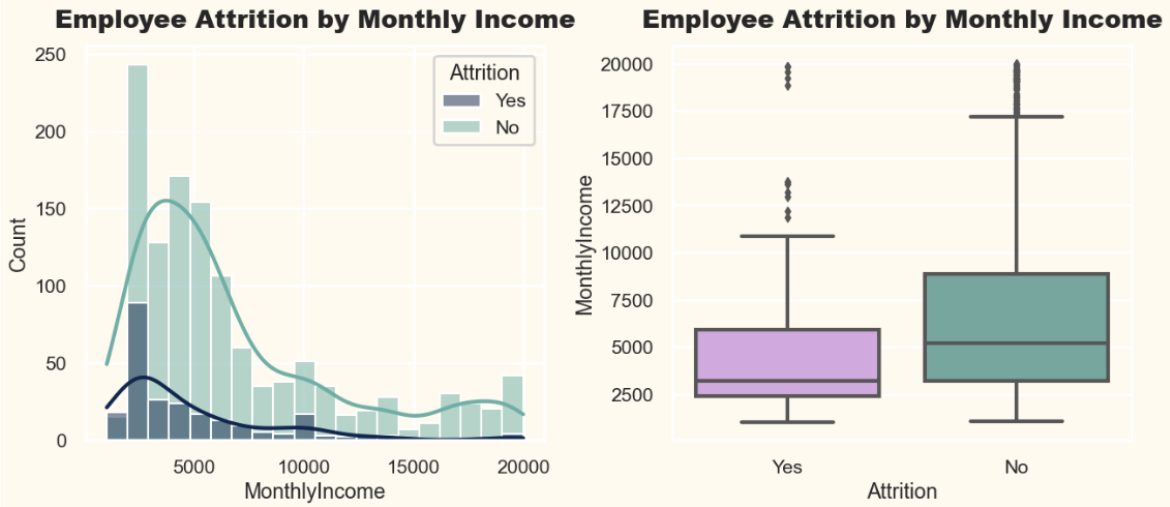

Fig.3.2.15: Analyzing Employee Attrition by Monthly Income.

Inference:

1. Most of the employees are getting paid less than 10000 in the organization.
2. The average monthly income of employee who has left is comparatively low with employee who is still working.
3. As the Monthly Income increases the attrition decreases.

16] ANALYZING EMPLOYEE ATTRITION BY WORK EXPERIENCE.

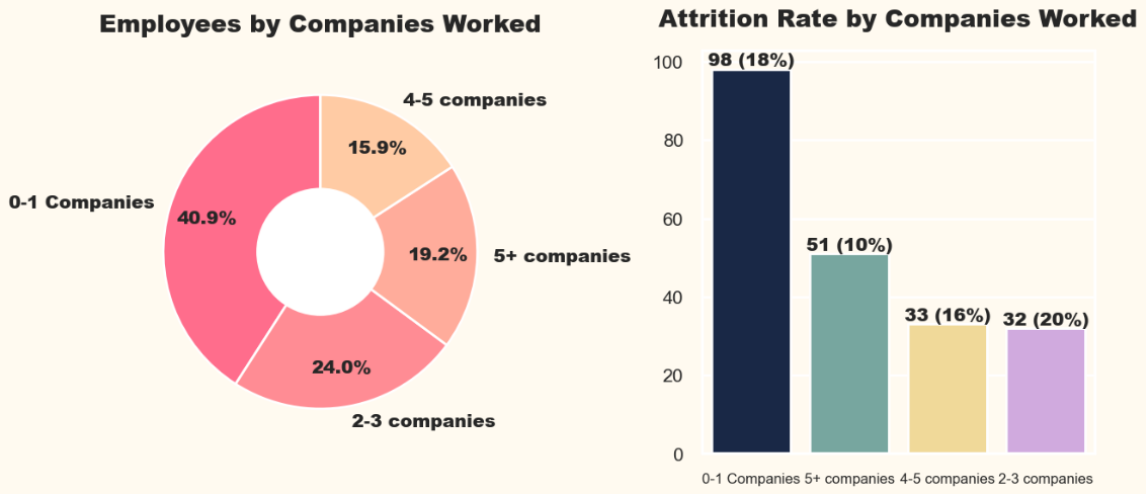

Fig.3.2.16: Analyzing Employee Attrition by Work Experience.

Inference:

1. Most of the employees have worked for less than 2 companies.
2. There's a high attrition rate of employees who have for less than 5 companies.

17] ANALYZING EMPLOYEE ATTRITION BY OVERTIME.

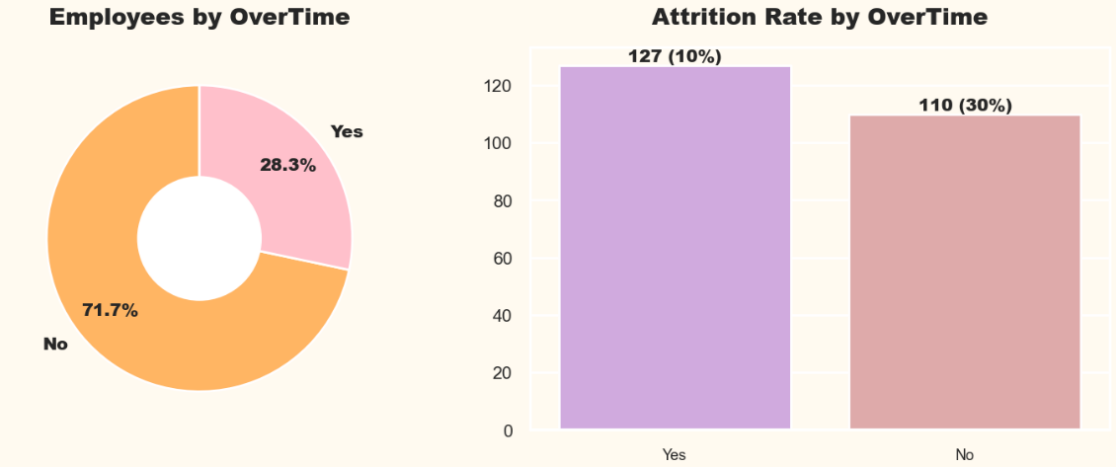

Fig.3.2.17: Analyzing Employee Attrition by Overtime.

Inference:

1. Most of the employees don’t work for OverTime.
2. The feature OverTime is having a very high class imbalance due to which we can't make any meaningful insights.

18] ANALYZING EMPLOYEE ATTRITION BY SALARY HIKE.

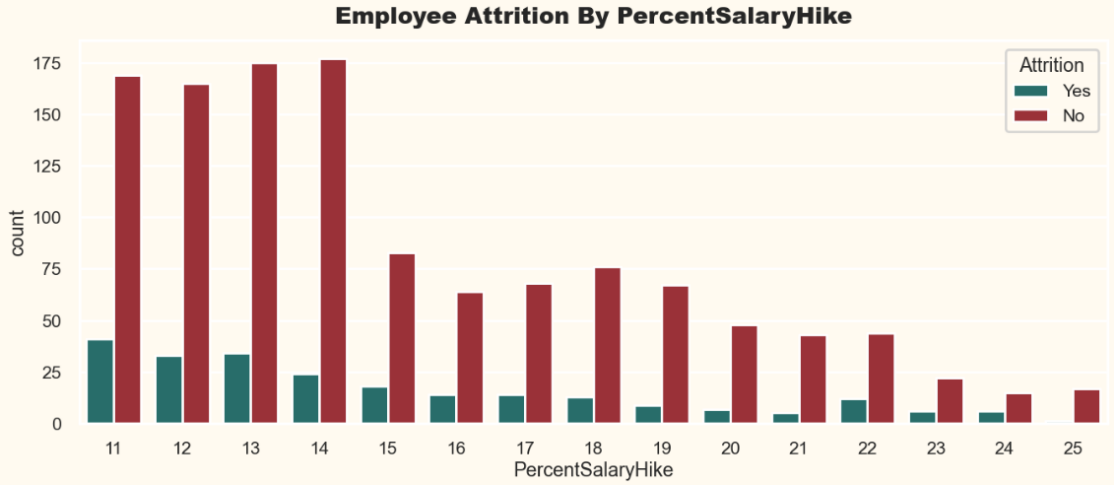

Fig.3.2.18: Analyzing Employee Attrition by Salary Hike.

Inference:

1. Very Few employees are getting a high percent salary hike.
2. As the amount of percent salary increases the attrition rate decreases.

19] ANALYZING EMPLOYEE ATTRITION BY PERFORMANCE RATING.

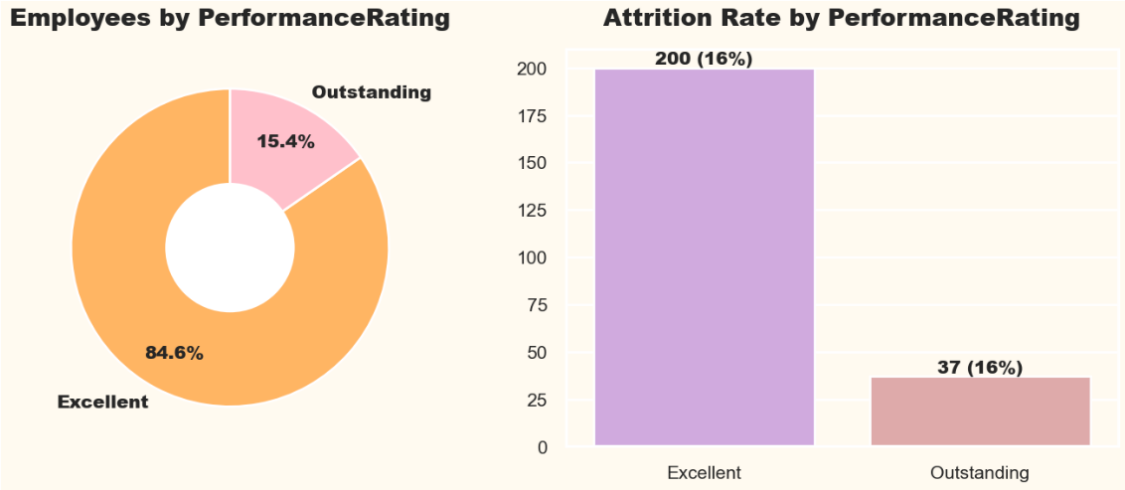

Fig.3.2.19: Analyzing Employee Attrition by Performance Rating.

Inference:

1. Most of the employees are having excellent performance rating.
2. Both the categories in this field is having same attrition rate.
3. That's why we can't generate any meaningful insights.

20] ANALYZING EMPLOYEE ATTRITION BY RELATIONSHIP SATISFACTION.

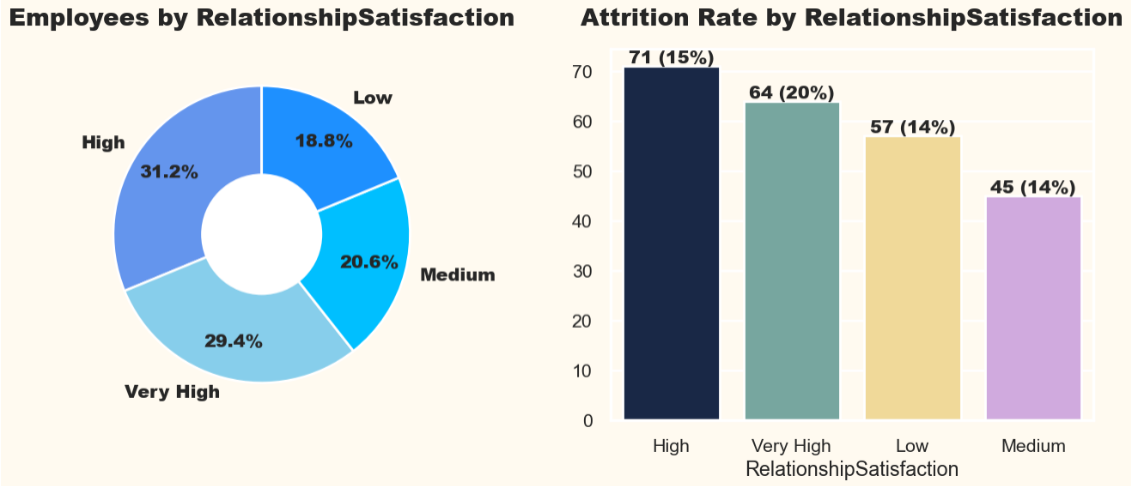

Fig.3.2.20: Analyzing Employee Attrition by Relationship Satisfaction.

Inference:

1. Most of the employees are having high or very high relationship satisfaction.
2. Though the relationship satisfaction is high there's a high attrition rate.
3. All the categories in this feature are having a high attrition rate.

21] ANALYZING EMPLOYEE ATTRITION BY WORK LIFE BALANCE.

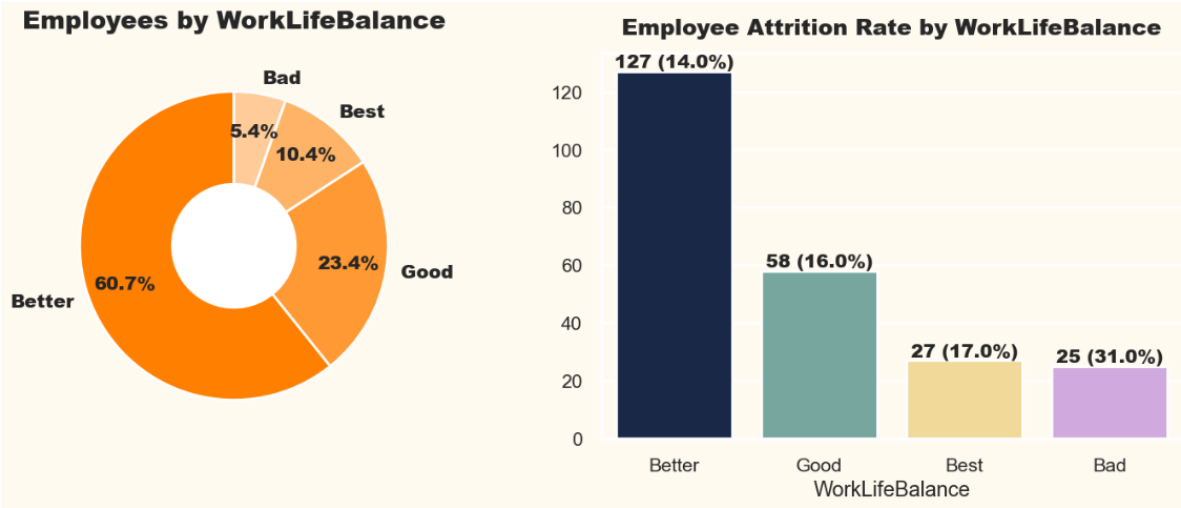

Fig.3.2.21: Analyzing Employee Attrition by Work Life Balance.

Inference:

1. More than 60% of employees are having a better work life balance.
2. Employees with Bad Work Life Balance are having Very High Attrition Rate.
3. Other Categories is also having High attrition Rate.

22] ANALYZING EMPLOYEE ATTRITION BY TOTAL WORKING EXPERIENCE.

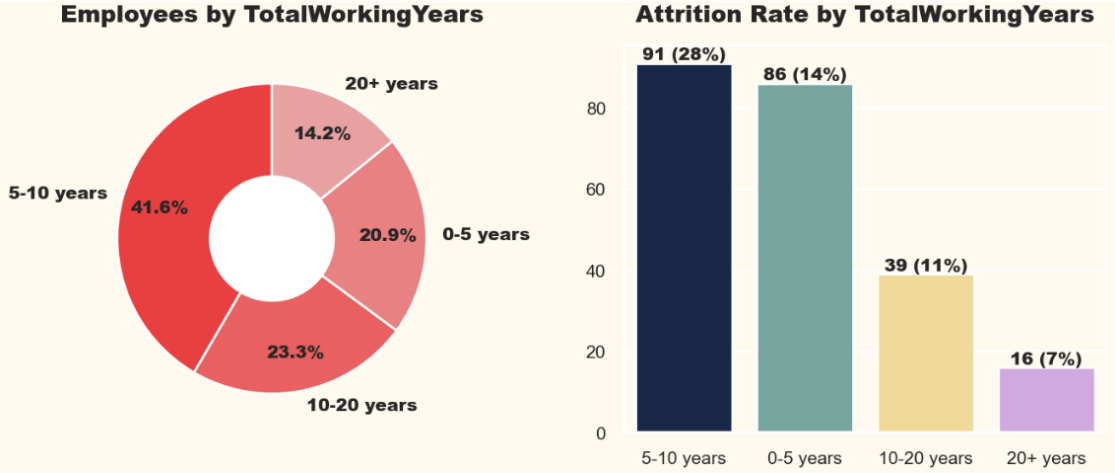

Fig.3.2.22: Analyzing Employee Attrition by Total Working Experience.

Inference:

1. Most of the employees are having a total of 5 to 10 years of working experience. But their Attrition Rate is also very high.
2. Employees with working experience of less than 10 years are having High Attrition Rate.
3. Employees with working experience of more than 10 years are having Less Attrition Rate.

23] ANALYZING EMPLOYEE ATTRITION BY YEARS AT COMPANY.

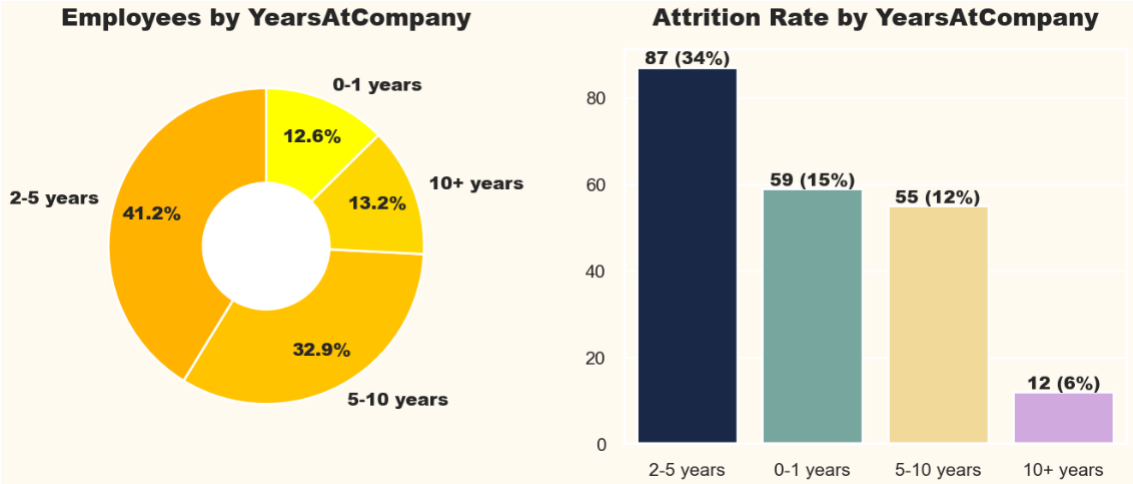

Fig.3.2.23: Analyzing Employee Attrition by Years at Company.

Inference:

1. Most employees have worked for 2 to 10 years in the organization.
2. Very few employees have working for less than 1 year or more than 10 years.
3. Employee who have worked for 2-5 years are having very high attrition rate.
4. Employee who have worked for 10+ years are having low attrition rate.

24] ANALYZING EMPLOYEE ATTRITION BY YEARS IN CURRENT ROLE.

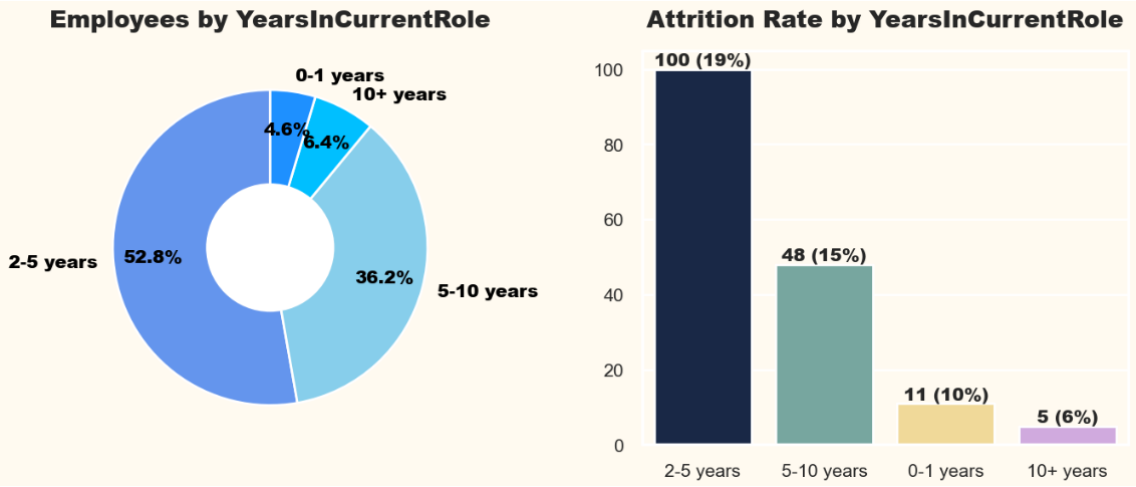

Fig.3.2.24: Analyzing Employee Attrition by Years in Current Role.

Inference:

1. Most employees have worked for 2 to 10 years for the same role in the organization.
2. Very few employees have worked for less than 1 year or more than 10 years in the same role.
3. Employee who has worked till 2 years in the same role are having very high attrition rate.
4. Employee who has worked for 10+ years in the same role are having low attrition rate.

25] ANALYZING EMPLOYEE ATTRITION BY YEARS SINCE LAST PROMOTION.

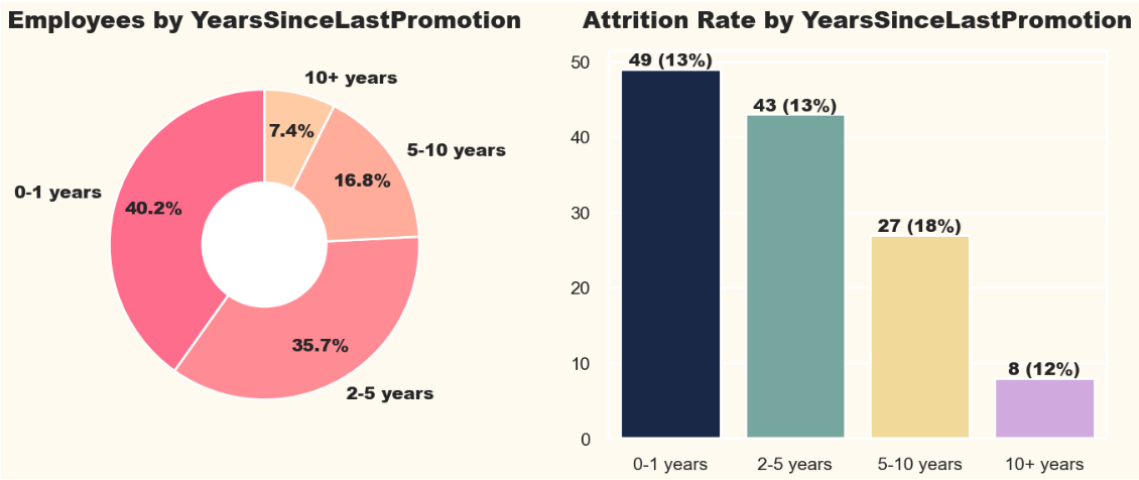

Fig.3.2.25: Analyzing Employee Attrition by Years Since Last Promotion.

Inference:

1. Almost 36% of employee has not been promoted since 2 to 5 years.
2. Almost 8% of employees have not been promoted since 10+ years.
3. All the categories in this feature is having high attrition rate specially employee who has not been promoted since 5+ years.

26] ANALYZING EMPLOYEE ATTRITION BY YEARS WITH CURRENT MANAGER.

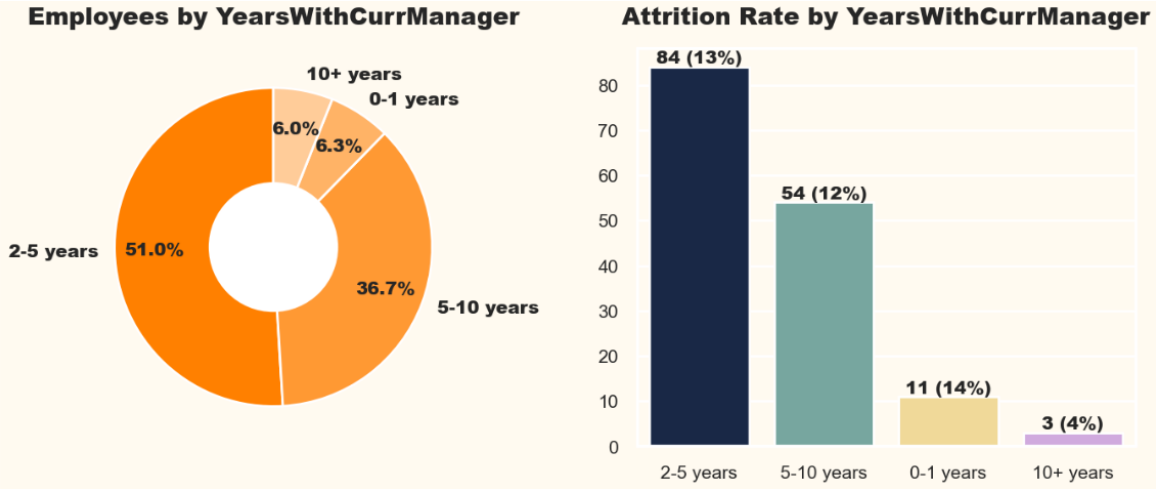

Fig.3.2.26: Analyzing Employee Attrition by Years with Current Manager.

Inference:

1. Almost 51% of employees have worked for 2-5 years with the same manager.
2. Almost 38% of employees have worked for 5-10 years with the same manager.
3. Employee who has worked for 10+ year with the same manager are having very low attrition rate.
4. Other Categories is having high attrition rate.

3.3 STATISTICAL ANALYSIS

Statistical analysis plays a crucial role in HR analytics by helping organizations make informed decisions about their human resources and workforce management. It enables evidence-based decision-making, enhances workforce planning strategies, and fosters a deeper understanding of the organization's human capital dynamics.

1] Perform ANOVA Test: ANOVA test is used to analysing the impact of different numerical features on a response categorical feature.

Inference:

The following features show a strong association with attrition, as indicated by their high F-scores and very low p-values.

1. Age
2. DailyRate
3. HourlyRate
4. MonthlyIncome
5. MonthlyRate
6. NumCompaniesWorked
7. PercentSalaryHike
8. TotalWorkingYears
9. TrainingTimesLastYear
10. YearsAtCompany
11. YearsWithCurrManager

The following features don’t shows significant relationship with attrition because of their moderate F-scores and extremely high p-values.

1. DistanceFromHome
2. StockOptionLevel
3. YearsInCurrentRole
4. YearsSinceLastPromotion

It is important for the organization to pay attention to the identified significant features and consider them when implementing strategies to reduce attrition rates.

2] Perform CHI-SQUARE Test: CHI-SQUARE test is used to analysing the impact of different categorical features.

Inference:

The following features showed statistically significant associations with employee attrition:

1. Department
2. EducationField
3. EnvironmentSatisfaction
4. JobInvolvement
5. JobLevel
6. JobRole
7. JobSatisfaction
8. MaritalStatus
9. OverTime
10. WorkLifeBalance

The following features did not show statistically significant associations with attrition.

1. Gender
2. Education
3. PerformanceRating
4. RelationshipSatisfaction

It is important for the organization to pay attention to the identified significant features and consider them when implementing strategies to reduce attrition rates.

3.4 DATA MODELING

Data modeling plays a significant role in HR analytics when integrating machine learning techniques. Machine learning algorithms leverage data models to make predictions, classifications, and recommendations based on patterns and relationships found in the HR data.

Data splitting to train and test:

The data set was split into 70% for training and 30% for testing and we have considered Attrition as target feature.

Fitting the different machine learning models:

1. Logistic Regression Model.

```
In [17]: lr_clf = LogisticRegression(solver='liblinear', penalty='l1')
         lr_clf.fit(X_train_std, y_train)

         evaluate(lr_clf, X_train_std, X_test_std, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[847  16]
 [ 59 107]]
ACCURACY SCORE:
0.9271
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.934879    0.869919  0.927114     0.902399      0.924399
recall       0.981460    0.644578  0.927114     0.813019      0.927114
f1-score     0.957603    0.740484  0.927114     0.849044      0.922577
support    863.000000  166.000000  0.927114  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[351  19]
 [ 41  30]]
ACCURACY SCORE:
0.8639
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.895408   0.612245  0.863946    0.753827      0.849820
recall       0.948649   0.422535  0.863946    0.685592      0.863946
f1-score     0.921260   0.500000  0.863946    0.710630      0.853438
support    370.000000  71.000000  0.863946  441.000000    441.000000
```

Fig.3.4.1: Training and Testing results by using Logistic Regression Model.

2. Random Forest Model.

In [21]:
```
param_grid = dict(
    n_estimators= [100, 500, 900],
    max_features= ['auto', 'sqrt'],
    max_depth= [2, 3, 5, 10, 15, None],
    min_samples_split= [2, 5, 10],
    min_samples_leaf= [1, 2, 4],
    bootstrap= [True, False]
)

rf_clf = RandomForestClassifier(random_state=42)
search = GridSearchCV(rf_clf, param_grid=param_grid, scoring='roc_auc', cv=5, verbose=1, n_jobs=-1)
search.fit(X_train, y_train)

rf_clf = RandomForestClassifier(**search.best_params_, random_state=42)
rf_clf.fit(X_train, y_train)
evaluate(rf_clf, X_train, X_test, y_train, y_test)
```

```
Fitting 5 folds for each of 648 candidates, totalling 3240 fits
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[863   0]
 [113  53]]
ACCURACY SCORE:
0.8902
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.884221    1.000000  0.890185     0.942111      0.902899
recall       1.000000    0.319277  0.890185     0.659639      0.890185
f1-score     0.938554    0.484018  0.890185     0.711286      0.865227
support    863.000000  166.000000  0.890185  1029.000000   1029.000000
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[365   5]
 [ 65   6]]
ACCURACY SCORE:
0.8413
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.848837   0.545455   0.84127    0.697146      0.799993
recall       0.986486   0.084507   0.84127    0.535497      0.841270
f1-score     0.912500   0.146341   0.84127    0.529421      0.789150
support    370.000000  71.000000   0.84127  441.000000    441.000000
```

Fig.3.4.2: Training and Testing results by using Random Forest Model.

3. Support Vector Machine Model.

```
In [26]:
svm_clf = SVC(random_state=42)

param_grid = [
    {'C': [1, 10, 100, 1000], 'kernel': ['linear']},
    {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001], 'kernel': ['rbf']}
]

search = GridSearchCV(svm_clf, param_grid=param_grid, scoring='roc_auc', cv=3, refit=True, verbose=1)
search.fit(X_train_std, y_train)
```

```
Fitting 3 folds for each of 12 candidates, totalling 36 fits
```

```
Out[26]: GridSearchCV(cv=3, estimator=SVC(random_state=42),
             param_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']},
                         {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001],
                          'kernel': ['rbf']}],
             scoring='roc_auc', verbose=1)
```

```
In [27]:
svm_clf = SVC(**search.best_params_)
svm_clf.fit(X_train_std, y_train)

evaluate(svm_clf, X_train_std, X_test_std, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[861   2]
 [ 65 101]]
ACCURACY SCORE:
0.9349
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.929806    0.980583  0.934888     0.955194      0.937997
recall       0.997683    0.608434  0.934888     0.803058      0.934888
f1-score     0.962549    0.750929  0.934888     0.856739      0.928410
support    863.000000  166.000000  0.934888  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[360  10]
 [ 49  22]]
ACCURACY SCORE:
0.8662
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.880196   0.687500  0.866213    0.783848      0.849172
recall       0.972973   0.309859  0.866213    0.641416      0.866213
f1-score     0.924262   0.427184  0.866213    0.675723      0.844234
support    370.000000  71.000000  0.866213  441.000000    441.000000
```

Fig.3.4.3: Training and Testing results by using Support Vector Machine Model.

4. XGBoost Model.

```
In [30]: xgb_clf = XGBClassifier()
         xgb_clf.fit(X_train, y_train)

         evaluate(xgb_clf, X_train, X_test, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision     1.0    1.0       1.0        1.0           1.0
recall        1.0    1.0       1.0        1.0           1.0
f1-score      1.0    1.0       1.0        1.0           1.0
support     863.0  166.0       1.0     1029.0        1029.0
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[356  14]
 [ 51  20]]
ACCURACY SCORE:
0.8526
CLASSIFICATION REPORT:
                   0          1  accuracy   macro avg  weighted avg
precision   0.874693   0.588235  0.852608    0.731464      0.828574
recall      0.962162   0.281690  0.852608    0.621926      0.852608
f1-score    0.916345   0.380952  0.852608    0.648649      0.830148
support   370.000000  71.000000  0.852608  441.000000    441.000000
```

Fig.3.4.4: Training and Testing results by using XGBoost Model.

5. LightGBM Model.

```
In [34]: lgb_clf = LGBMClassifier()
         lgb_clf.fit(X_train, y_train)

         evaluate(lgb_clf, X_train, X_test, y_train, y_test)
```

```
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Info] Number of positive: 166, number of negative: 863
[LightGBM] [Warning] Auto-choosing row-wise multi-threading, the overhead of testing was 0.000203 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 1176
[LightGBM] [Info] Number of data points in the train set: 1029, number of used features: 108
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.161322 -> initscore=-1.648427
[LightGBM] [Info] Start training from score -1.648427
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision     1.0    1.0       1.0        1.0           1.0
recall        1.0    1.0       1.0        1.0           1.0
f1-score      1.0    1.0       1.0        1.0           1.0
support     863.0  166.0       1.0     1029.0        1029.0
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[353  17]
 [ 54  17]]
ACCURACY SCORE:
0.8390
CLASSIFICATION REPORT:
                   0          1  accuracy   macro avg  weighted avg
precision   0.867322   0.500000  0.839002    0.683661      0.808184
recall      0.954054   0.239437  0.839002    0.596745      0.839002
f1-score    0.908623   0.323810  0.839002    0.616216      0.814469
support   370.000000  71.000000  0.839002  441.000000    441.000000
```

Fig.3.4.5: Training and Testing results by using LightGBM Model.

6. CatBoost Model.

```
In [37]: cb_clf = CatBoostClassifier()
         cb_clf.fit(X_train, y_train, verbose=0)

         evaluate(cb_clf, X_train, X_test, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [ 16 150]]
ACCURACY SCORE:
0.9845
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.981797    1.000000  0.984451     0.990899      0.984734
recall       1.000000    0.903614  0.984451     0.951807      0.984451
f1-score     0.990815    0.949367  0.984451     0.970091      0.984129
support    863.000000  166.000000  0.984451  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[363   7]
 [ 59  12]]
ACCURACY SCORE:
0.8503
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.860190   0.631579   0.85034    0.745884      0.823384
recall       0.981081   0.169014   0.85034    0.575048      0.850340
f1-score     0.916667   0.266667   0.85034    0.591667      0.812018
support    370.000000  71.000000  0.85034  441.000000    441.000000
```

Fig.3.4.6: Training and Testing results by using CatBoost Model.

7. AdaBoost Model.

```
In [40]: ab_clf = AdaBoostClassifier()
         ab_clf.fit(X_train, y_train)

         evaluate(ab_clf, X_train, X_test, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[846  17]
 [ 78  88]]
ACCURACY SCORE:
0.9077
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.915584    0.838095  0.907677     0.876840      0.903084
recall       0.980301    0.530120  0.907677     0.755211      0.907677
f1-score     0.946838    0.649446  0.907677     0.798142      0.898863
support    863.000000  166.000000  0.907677  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[346  24]
 [ 50  21]]
ACCURACY SCORE:
0.8322
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.873737   0.466667    0.8322    0.670202      0.808200
recall       0.935135   0.295775    0.8322    0.615455      0.832200
f1-score     0.903394   0.362069    0.8322    0.632732      0.816242
support    370.000000  71.000000    0.8322  441.000000    441.000000
```

Fig.3.4.7: Training and Testing results by using AdaBoost Model.

ROC Curve: An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

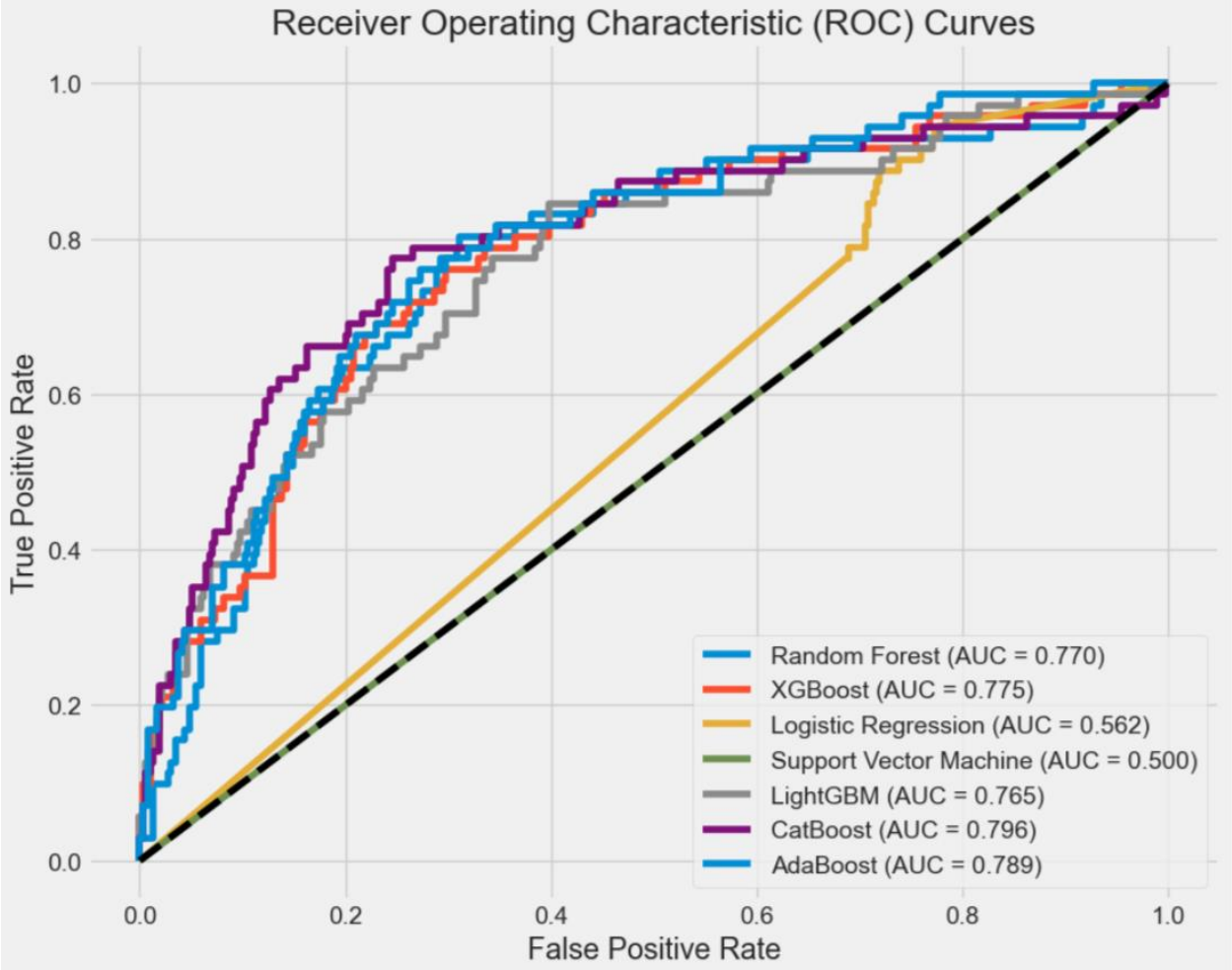

Fig.3.4.8: ROC Curve Diagram.

The graph is well-structured and displays multiple lines of varying colors, each representing a different machine learning model. The models include Random Forest, XGBoost, Logistic Regression, Support Vector Machine, LightGBM, CatBoost, and AdaBoost. Each model is also associated with an Area Under Curve (AUC) value, indicating their performance. The graph presents the True Positive Rate on the y-axis, ranging from 0.0 to 1.0, and the False Positive Rate on the x-axis, also ranging from 0.0 to 1.0. Here, Model like Random Forest, XGBoost, LightGBM, CatBoost, AdaBoost have better performance comparing with Support Vector Machine and Logistic Regression.

CHAPTER 4: PROJECT PLANNING

4.1 PROJECT PLANNING

Sr. No.	Tasks to be completed	Start Date	End Date
1.	Exploring the Domain	15 August	20 August
2.	Finalizing Topic and Domain	21 August	25 August
3.	Designing Methodology	26 August	30 August
4.	Data Exploration & Processing	01 September	10 September
5.	Data Visualization	11 September	30 September
6.	Statistical Analysis	01 October	10 October
7.	Data Modeling	11 October	25 October
8.	Report Making & PPT Making	26 October	20 November

Table 4.1 Project's Progress Planning

CHAPTER 5

5.1 CONCLUSION

In conclusion, we embarked on a comprehensive analysis of the IBM HR Analytics Attrition Dataset, from data loading to model evaluation. By implementing and evaluating various machine learning algorithms, we gained insights into which models are effective for predicting employee attrition. The results and visualizations generated throughout the process provide valuable information for decision-makers and HR professionals seeking to understand and mitigate employee attrition within the organization. This project showcases the power of data analysis and machine learning in addressing real-world business challenges.

5.2 FUTURE WORK

In the context of the previous HR analytics project on employee attrition, future work in sentiment analysis involves implementing sentiment analysis on employee feedback data to gain insights, monitoring sentiment in real-time, categorizing sentiments by topics, and analyzing historical sentiment trends. In terms of dashboard development, there's a need to create interactive, predictive, and benchmarking-enabled dashboards with custom alerts, engagement metrics, and mobile accessibility. Additionally, user training and support, data privacy, feedback integration, and performance monitoring are crucial aspects to ensure the dashboard's effectiveness in facilitating data-driven HR decisions and actions while adhering to privacy regulations.

REFERENCES

[1] Mishra S N, Lama D R and Pal Y 2016 Human Resource Predictive Analytics (HRPA) for HR Management in Organizations International Journal Of Scientific & Technology Research 5(5) 33-35

[2] Hoffman M and Tadelis S 2018 People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis National Bureau of Economic Research

[3] Frye A, Boomhower C, Smith M, Vitovsky L and Fabricant S 2018 Employee Attrition: What Makes an Employee Quit? MU Data Science Review 1(1)

[4] S. Rabiyathul Basariya, Ramyar Rzgar Ahmed, A STUDY ON ATTRITION - TURNOVER INTENTIONS OF EMPLOYEES, International Journal of Civil Engineering and Technology (IJCIET), 2019, 10(1), PP2594-2601

[5] Halkos, George & Bousinakis, Dimitrios, 2017. "The effect of stress and dissatisfaction on employees during crisis," Economic Analysis and Policy, Elsevier, vol. 55(C), pages 25-34.

[6] Glavas, A., & Willness, C. (2020). Employee (dis)engagement in corporate social responsibility. In D. Haski-Leventhal, L. Roza, & S. Brammer (Eds.), Employee engagement in corporate social responsibility (pp. 10–27). Sage Publications Ltd. https://doi.org/10.4135/9781529739176.n2

[7] S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 349-354, doi: 10.1109/IADCC.2018.8692137.

[8] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2018, pp. 113-120, doi: 10.1109/SYSMART.2018.8746940.

[9] lduayj, Sarah & Rajpoot, Kashif. (2018). Predicting Employee Attrition using Machine Learning. 93-98. 10.1109/INNOVATIONS.2018.8605976.

[10] Setiawan, Irwan & Suprihanto, Suprihanto & Nugraha, Ade & Hutahaean, Jonner. (2020). HR analytics: Employee attrition analysis using logistic regression. IOP Conference Series: Materials Science and Engineering. 830. 032001. 10.1088/1757-899X/830/3/032001.

[11] Yadav, Sandeep & Jain, Aman & Singh, Deepti. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. 349-354. 10.1109/IADCC.2018.8692137.

[12] I. Ballal, S. Kavathekar, S. Janwe, P. Shete, and N. Bhirud, "People Leaving the Job-An Approach for Prediction Using Machine Learning," Int. J. Res. Anal. Rev., vol. 7, no. 1, pp. 8– 10, 2020, [Online]. Available: www.ijrar.org

[13] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods," Int. J. Mach. Learn. Comput., vol. 11, no. 2, pp. 110–114, 2021, doi: 10.18178/ijmlc.2021.11.2.1022.

[14] N. Mansor, N. S. Sani, and M. Aliff, "Machine Learning for Predicting Employee Attrition," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 435–445, 2021, doi: 10.14569/ IJACSA.2021.0121149.

[15] D. Saisanthiya, V. M. Gayathri, and P. Supraja, "Employee Attrition Prediction Using Machine Learning and Sentiment Analysis," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 5, pp. 7550–7557, 2020, doi: 10.30534/ijatcse/2020/91952020.