Report: Employee Attrition Analysis

UNIFIED MENTOR

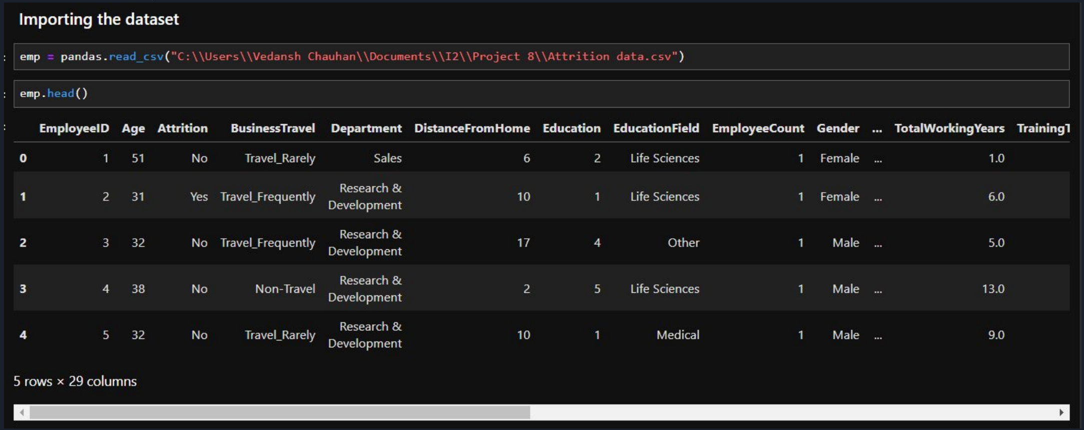
YOUR SKILL, SUCCESS & JOURNEY

Introduction

Employee attrition is the gradual reduction of a company's workforce over time. It may occur due to a number of factors such as voluntary resignations, retirements, and other forms of departure. It's different from employee turnover, which encompasses both voluntary and involuntary separations, including layoffs and terminations.

High attrition rates can indicate underlying issues within a company, such as poor management, lack of career growth opportunities, or inadequate compensation. Managing attrition effectively is crucial for any company, as it involves understanding its causes, addressing any systemic issues, and implementing strategies to retain valuable employees.

Overview of the Data



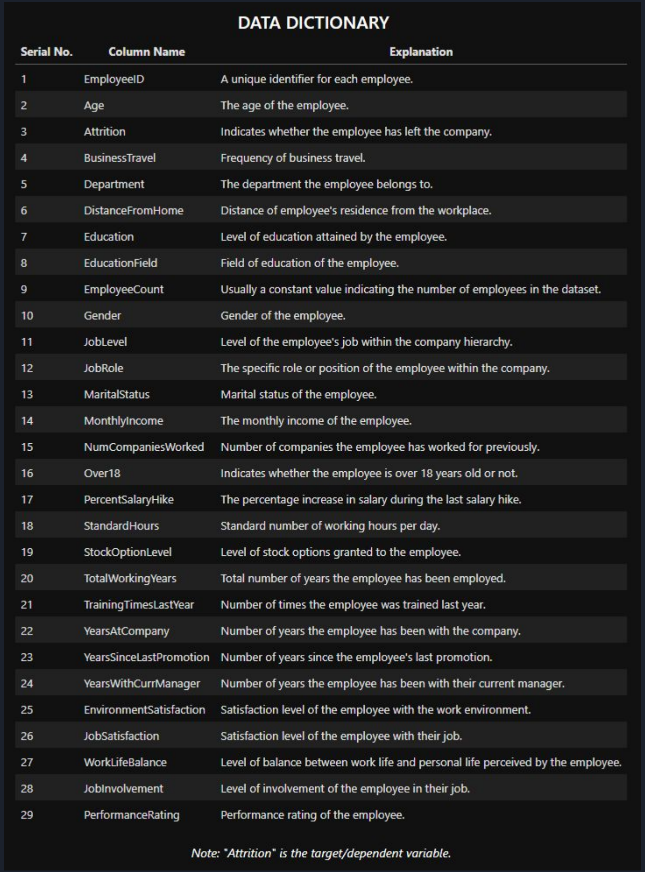
Importing the dataset

```
emp = pandas.read_csv("C:\\Users\\Vedansh Chauhan\\Documents\\I2\\Project 8\\Attrition data.csv")
```

```
emp.head()
```

	EmployeeID	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Gender	...	TotalWorkingYears	TrainingT
0	1	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	Female	...	1.0	
1	2	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	Female	...	6.0	
2	3	32	No	Travel_Frequently	Research & Development	17	4	Other	1	Male	...	5.0	
3	4	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	Male	...	13.0	
4	5	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	Male	...	9.0	

5 rows × 29 columns



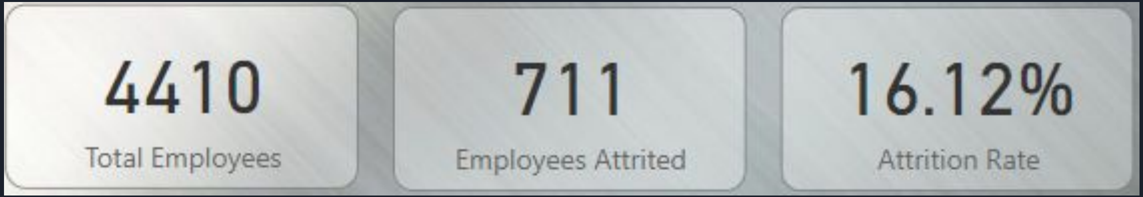
DATA DICTIONARY

Serial No.	Column Name	Explanation
1	EmployeeID	A unique identifier for each employee.
2	Age	The age of the employee.
3	Attrition	Indicates whether the employee has left the company.
4	BusinessTravel	Frequency of business travel.
5	Department	The department the employee belongs to.
6	DistanceFromHome	Distance of employee's residence from the workplace.
7	Education	Level of education attained by the employee.
8	EducationField	Field of education of the employee.
9	EmployeeCount	Usually a constant value indicating the number of employees in the dataset.
10	Gender	Gender of the employee.
11	JobLevel	Level of the employee's job within the company hierarchy.
12	JobRole	The specific role or position of the employee within the company.
13	MaritalStatus	Marital status of the employee.
14	MonthlyIncome	The monthly income of the employee.
15	NumCompaniesWorked	Number of companies the employee has worked for previously.
16	Over18	Indicates whether the employee is over 18 years old or not.
17	PercentSalaryHike	The percentage increase in salary during the last salary hike.
18	StandardHours	Standard number of working hours per day.
19	StockOptionLevel	Level of stock options granted to the employee.
20	TotalWorkingYears	Total number of years the employee has been employed.
21	TrainingTimesLastYear	Number of times the employee was trained last year.
22	YearsAtCompany	Number of years the employee has been with the company.
23	YearsSinceLastPromotion	Number of years since the employee's last promotion.
24	YearsWithCurrManager	Number of years the employee has been with their current manager.
25	EnvironmentSatisfaction	Satisfaction level of the employee with the work environment.
26	JobSatisfaction	Satisfaction level of the employee with their job.
27	WorkLifeBalance	Level of balance between work life and personal life perceived by the employee.
28	JobInvolvement	Level of involvement of the employee in their job.
29	PerformanceRating	Performance rating of the employee.

Note: "Attrition" is the target/dependent variable.

Supervised Learning selected for the data as the target column includes discrete value. The target column is “Attrition”.

General Statistics of the Data



4410
Total Employees

711
Employees Attrited

16.12%
Attrition Rate

Machine Learning Models Used

- Logistic Regression Classifier
- Random Forest Classifier
- eXtreme Gradient Boosting Classifier
- K-Nearest Neighbours (KNN) Classifier

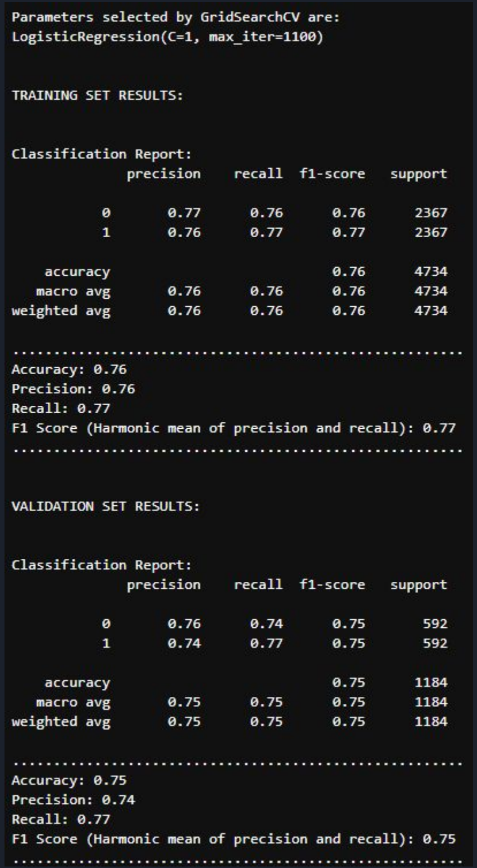
Prediction Results on the Training and Validation Sets.

Results for Logistic Regression Classifier

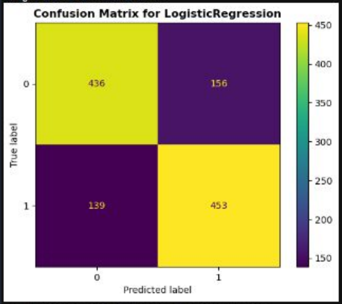
(Training and Validation Sets)

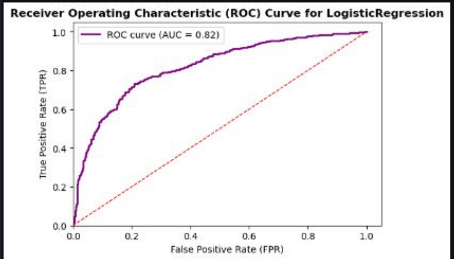
Logistic Regression Classifier did not predict any of the classes well.

It appears to have under fit the training data.

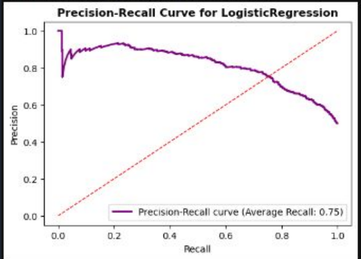


```
Parameters selected by GridSearchCV are:
LogisticRegression(C=1, max_iter=1100)


TRAINING SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.76      0.76      2367
           1       0.76      0.77      0.77      2367

    accuracy                           0.76      4734
   macro avg       0.76      0.76      0.76      4734
weighted avg       0.76      0.76      0.76      4734

.......................................................
Accuracy: 0.76
Precision: 0.76
Recall: 0.77
F1 Score (Harmonic mean of precision and recall): 0.77
.......................................................


VALIDATION SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       592
           1       0.74      0.77      0.75       592

    accuracy                           0.75      1184
   macro avg       0.75      0.75      0.75      1184
weighted avg       0.75      0.75      0.75      1184

.......................................................
Accuracy: 0.75
Precision: 0.74
Recall: 0.77
F1 Score (Harmonic mean of precision and recall): 0.75
.......................................................
```





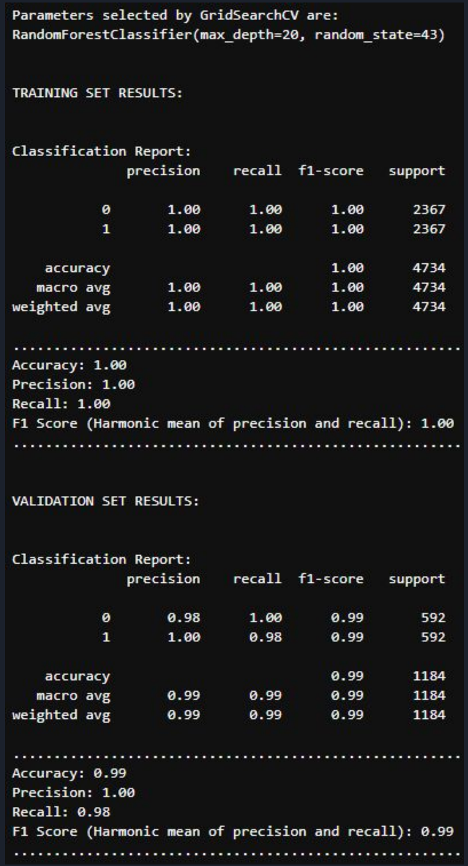
Note for AUC: A higher AUC indicates better performance.



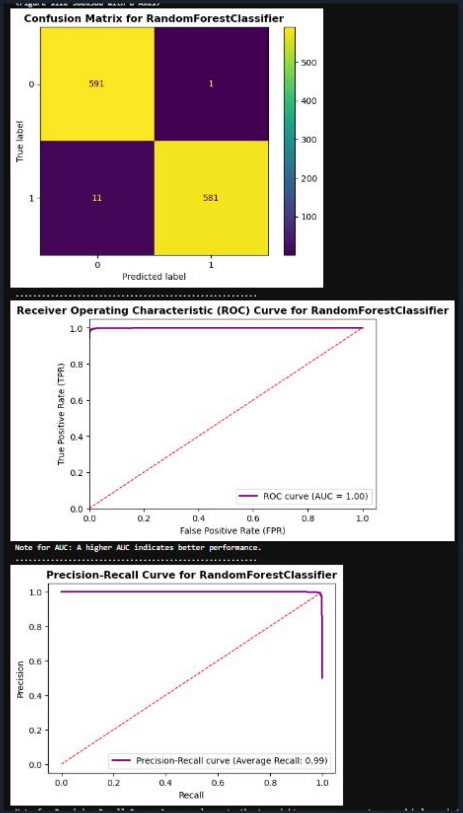
Results for Random Forest Classifier

(Training and Validation Sets)

Random Forest Classifier appears to have fit the data well and has predicted the Validation set class very well.



```
Parameters selected by GridSearchCV are:
RandomForestClassifier(max_depth=20, random_state=43)


TRAINING SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2367
           1       1.00      1.00      1.00      2367

    accuracy                           1.00      4734
   macro avg       1.00      1.00      1.00      4734
weighted avg       1.00      1.00      1.00      4734

.......................................................
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score (Harmonic mean of precision and recall): 1.00
.......................................................


VALIDATION SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       592
           1       1.00      0.98      0.99       592

    accuracy                           0.99      1184
   macro avg       0.99      0.99      0.99      1184
weighted avg       0.99      0.99      0.99      1184

.......................................................
Accuracy: 0.99
Precision: 1.00
Recall: 0.98
F1 Score (Harmonic mean of precision and recall): 0.99
.......................................................
```

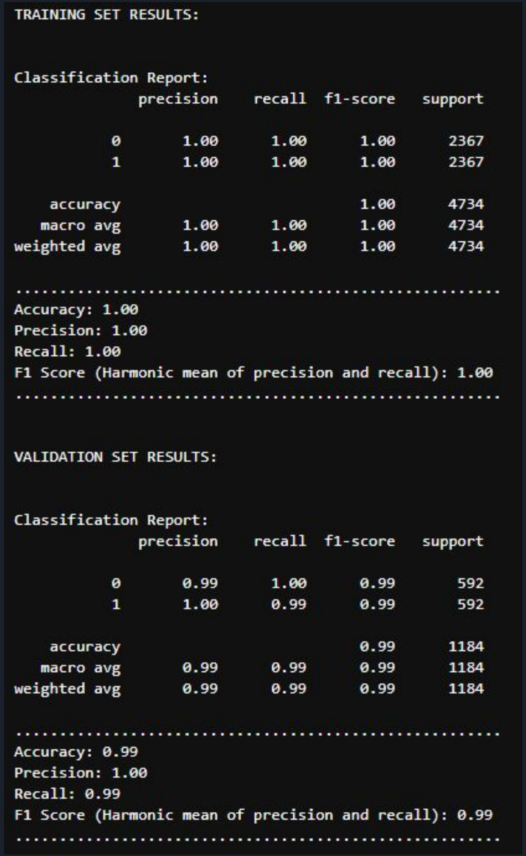


```
.......................................................
Note for AUC: A higher AUC indicates better performance.
.......................................................
```

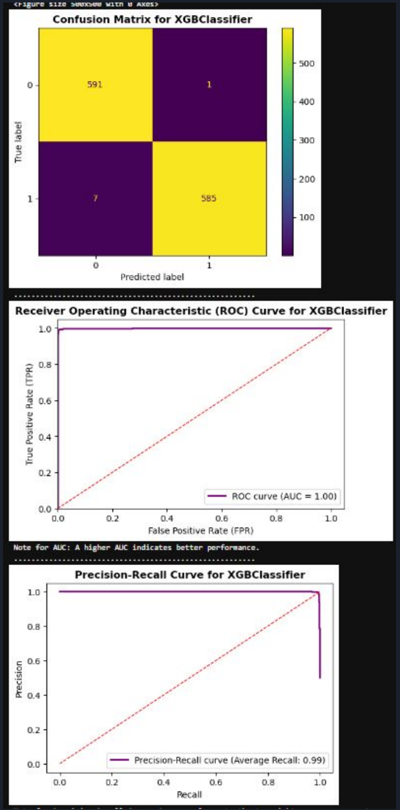
Results for eXtreme Gradient Boost Classifier

(Training and Validation Sets)

eXtreme Gradient Boost Classifier predicted the Validation set classes brilliantly.



```
TRAINING SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2367
           1       1.00      1.00      1.00      2367

    accuracy                           1.00      4734
   macro avg       1.00      1.00      1.00      4734
weighted avg       1.00      1.00      1.00      4734

.......................................................
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score (Harmonic mean of precision and recall): 1.00
.......................................................


VALIDATION SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       592
           1       1.00      0.99      0.99       592

    accuracy                           0.99      1184
   macro avg       0.99      0.99      0.99      1184
weighted avg       0.99      0.99      0.99      1184

.......................................................
Accuracy: 0.99
Precision: 1.00
Recall: 0.99
F1 Score (Harmonic mean of precision and recall): 0.99
.......................................................
```



```
<Figure size 500x500 with 0 Axes>
```

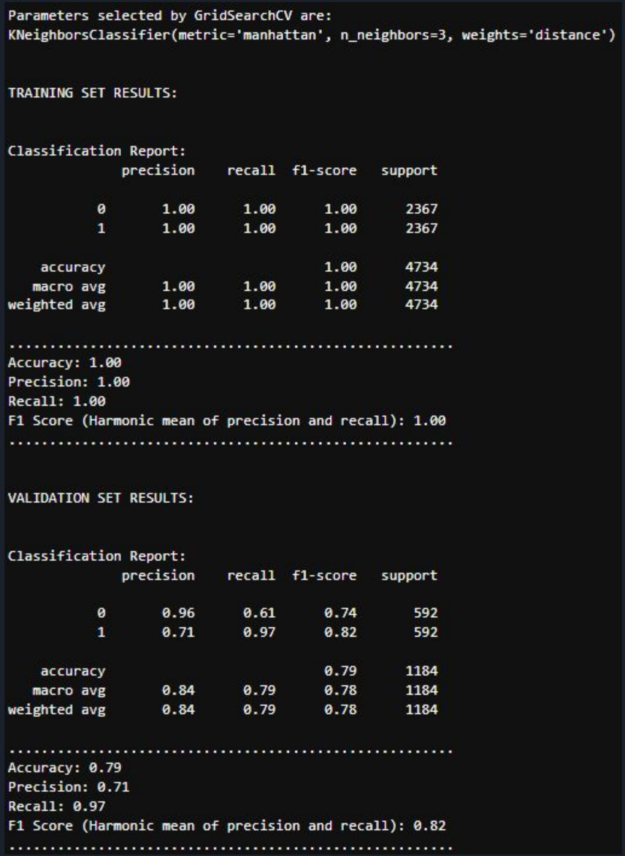
```
Note for AUC: A higher AUC indicates better performance.
```

Results for K-Nearest Neighbours (KNN) Classifier

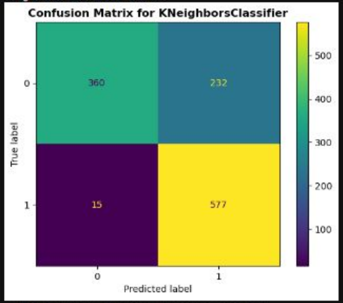
(Training and Validation Sets)

K-Nearest Neighbours (KNN) Classifier performed flawlessly on the training set however, it struggled with unseen data of the Validation Set.

It seems to have overfit the training data.

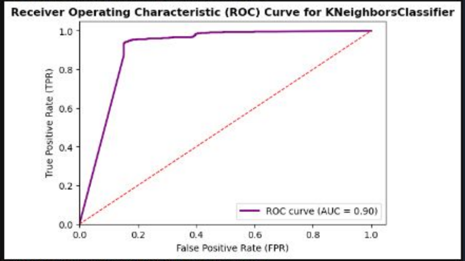


```
Parameters selected by GridSearchCV are:
KNeighborsClassifier(metric='manhattan', n_neighbors=3, weights='distance')


TRAINING SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2367
           1       1.00      1.00      1.00      2367

    accuracy                           1.00      4734
   macro avg       1.00      1.00      1.00      4734
weighted avg       1.00      1.00      1.00      4734

.......................................................
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score (Harmonic mean of precision and recall): 1.00
.......................................................


VALIDATION SET RESULTS:


Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.61      0.74       592
           1       0.71      0.97      0.82       592

    accuracy                           0.79      1184
   macro avg       0.84      0.79      0.78      1184
weighted avg       0.84      0.79      0.78      1184

.......................................................
Accuracy: 0.79
Precision: 0.71
Recall: 0.97
F1 Score (Harmonic mean of precision and recall): 0.82
.......................................................
```



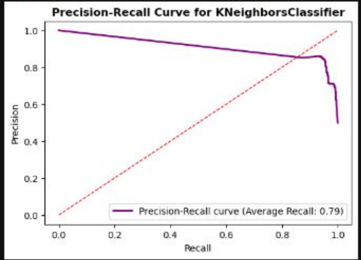
Confusion Matrix for KNeighborsClassifier

```
.......................................................
```



Receiver Operating Characteristic (ROC) Curve for KNeighborsClassifier

```
Note for AUC: A higher AUC indicates better performance.
.......................................................
```



Precision-Recall Curve for KNeighborsClassifier

Prediction Results on the Test Sets

(Results for Random Forest Classifier and eXtreme Gradient Boost Classifier as their predictions were quite accurate).

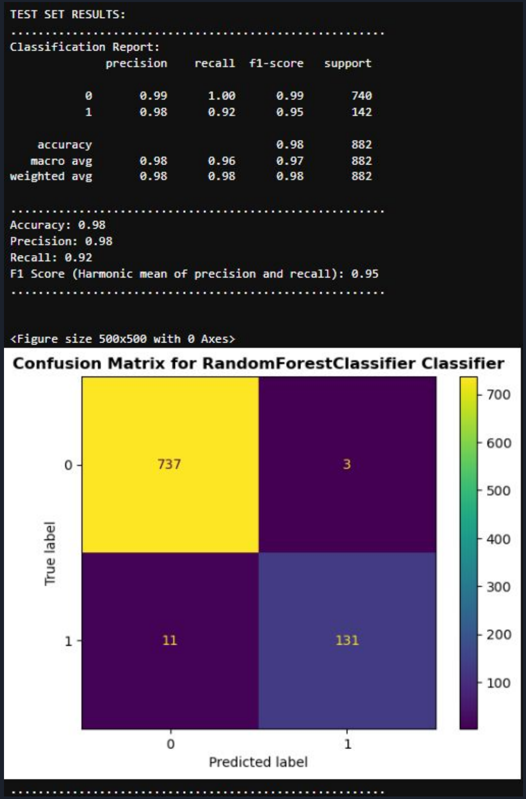
Results for Random Forest Classifier and eXtreme Gradient Boost Classifier

(Test Set)

Both Random Forest Classifier as well as eXtreme Gradient Boost Classifier have performed very well in predicting the target classes well.

However, eXtreme Gradient Boost Classifier appears to scored higher in the evaluation metrics. Hence, it became my choice of model of predicting attrition.

Random Forest Classifier:

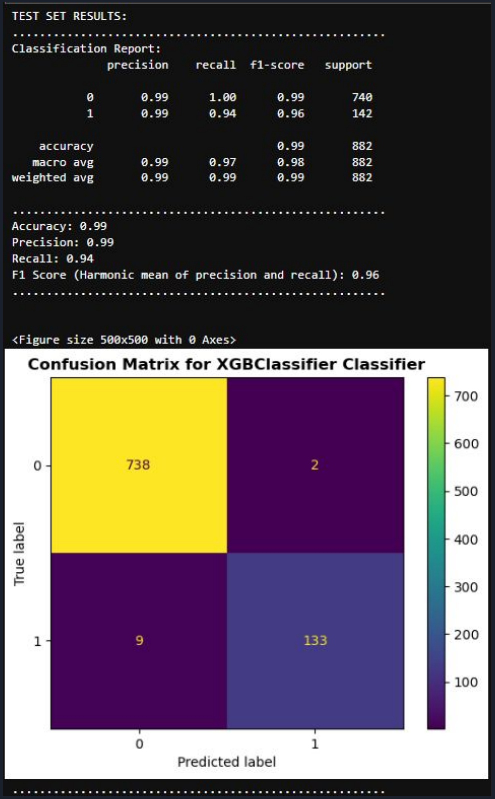


```
TEST SET RESULTS:
.......................................................
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       740
           1       0.98      0.92      0.95       142

    accuracy                           0.98       882
   macro avg       0.98      0.96      0.97       882
weighted avg       0.98      0.98      0.98       882

.......................................................
Accuracy: 0.98
Precision: 0.98
Recall: 0.92
F1 Score (Harmonic mean of precision and recall): 0.95
.......................................................

<Figure size 500x500 with 0 Axes>
```

Confusion Matrix for RandomForestClassifier Classifier

```
.......................................................
```

eXtreme Gradient Boost Classifier:

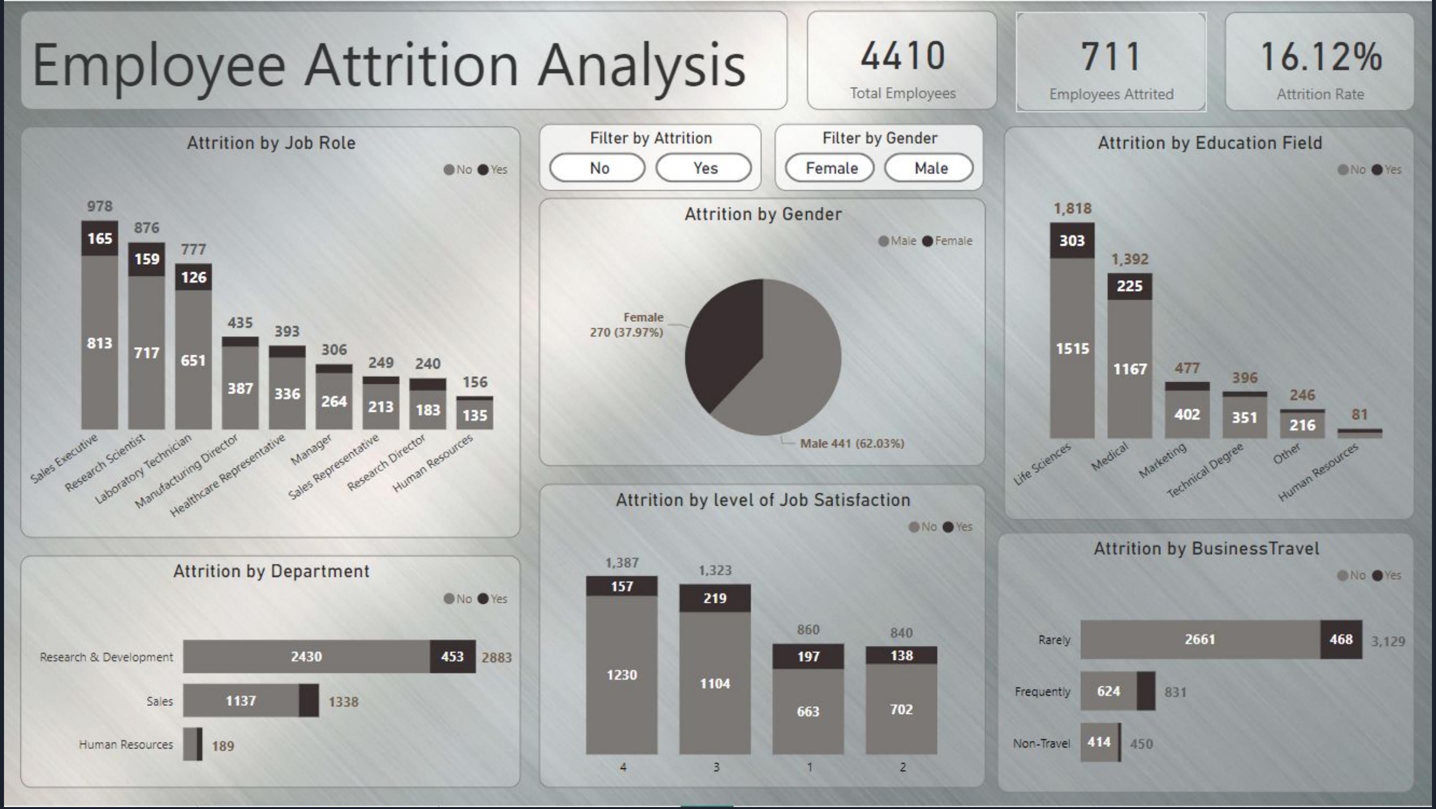


```
TEST SET RESULTS:
.......................................................
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       740
           1       0.99      0.94      0.96       142

    accuracy                           0.99       882
   macro avg       0.99      0.97      0.98       882
weighted avg       0.99      0.99      0.99       882

.......................................................
Accuracy: 0.99
Precision: 0.99
Recall: 0.94
F1 Score (Harmonic mean of precision and recall): 0.96
.......................................................

<Figure size 500x500 with 0 Axes>
```

Confusion Matrix for XGBClassifier Classifier

```
.......................................................
```

PowerBI DashBoard (Page 1)



Employee Attrition Analysis

4410 Total Employees

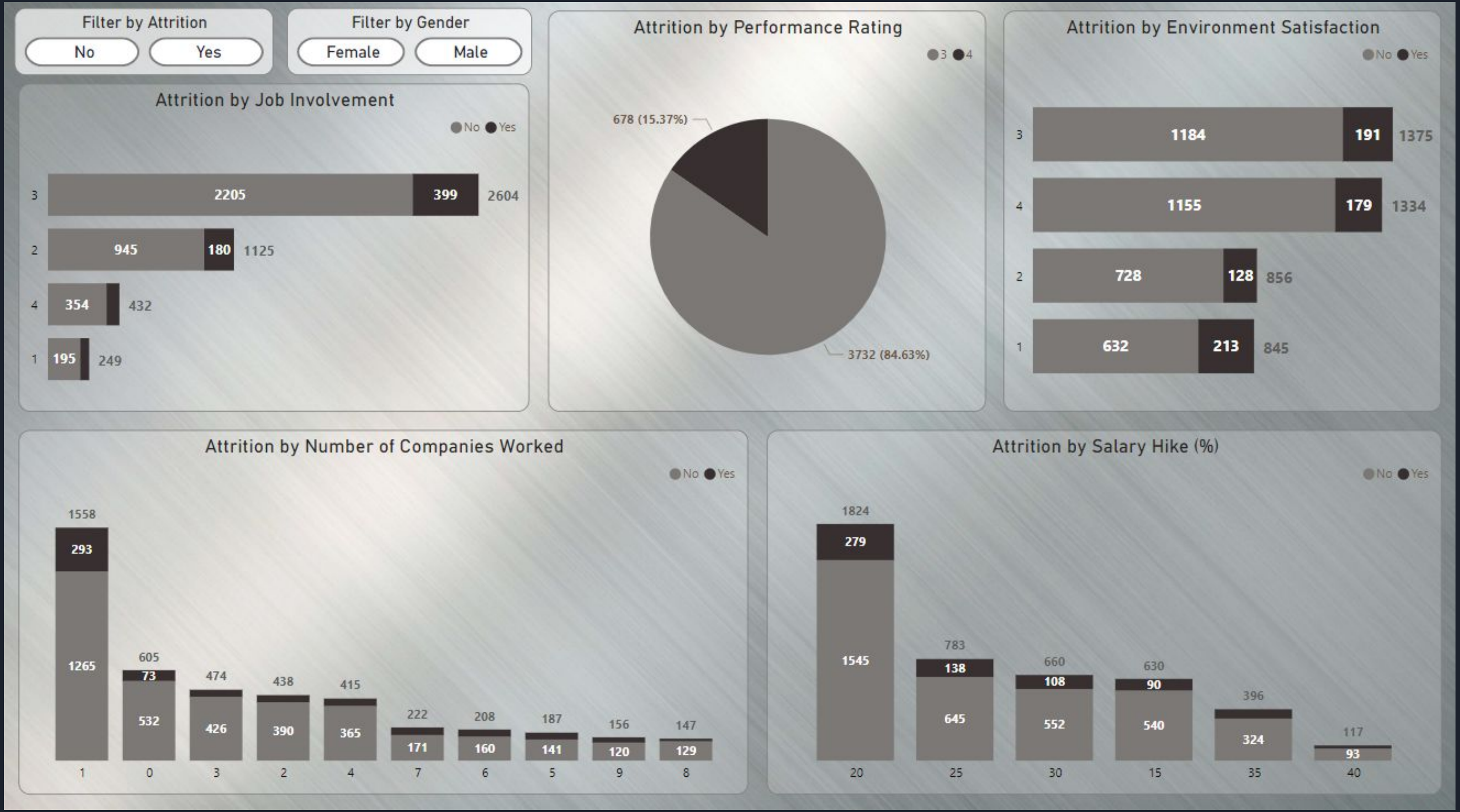
711 Employees Attrited

16.12% Attrition Rate

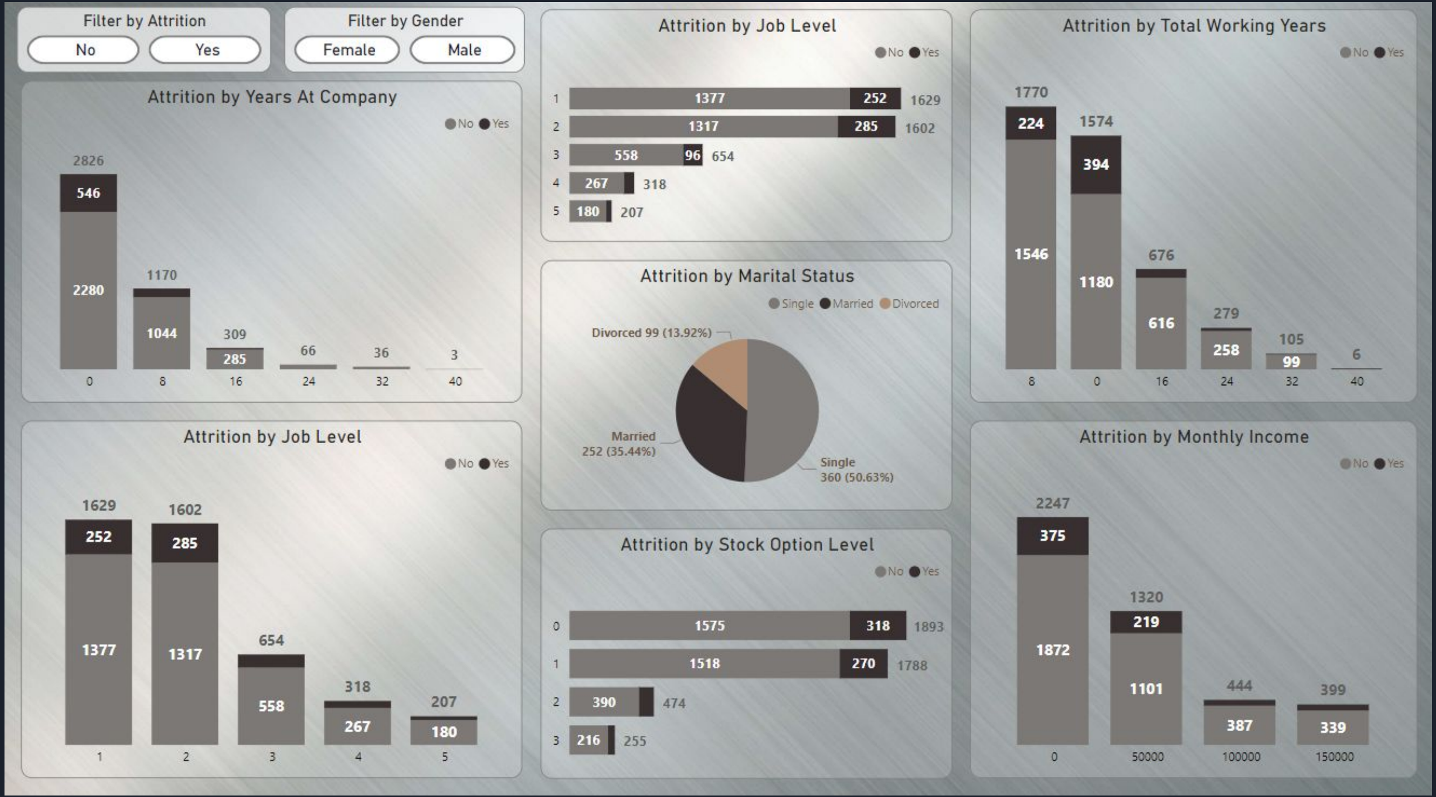
Filter by Attrition: No | Yes

Filter by Gender: Female | Male

PowerBI DashBoard (Page 2)



PowerBI DashBoard (Page 3)



Conclusion and Insights

- The eXtreme Gradient Boosting (XGBoost) Classifier made the most accurate predictions over the test set, followed by Random Forest Classifier. Therefore, I select the same as the final model for future attrition predictions.

Based on the PowerBI Dashboard, the attrition was the highest among the following categories:

- Department: Research & Development
- Job Role: Sales Executive
- Gender: Male
- Education Field: Life Sciences
- Marital Status: Single

Other trends for most attrition among employees:

- The total number of employees who attrited was 771 out of a total of 4410 employees.
- The attrition rate was 16.12%.
- Most attrited employees left in less than a year and were earning relatively low salaries.
- Employees who rarely experienced business travel opportunities.
- Employees who had only worked at a single company.
- Employees who gave an Environment Satisfaction rating of 1.

Recommended Solutions:

- Given the high attrition of new employees, managers should be asked to explain the reasons for this and even be replaced if necessary.
- The work environment needs to be analyzed for potential discouraging factors.
- Employees in the roles of Sales Executive, Research Scientist, along with other roles, should be asked for their feedback.
- More employees should be given business travel opportunities or short vacations if feasible.
- Information related to future salary hikes and promotions should be disclosed early on to improve employee satisfaction.
- Other relevant steps should be taken to avoid overburdening employees, and a level of mutual respect must be maintained.

GitHub Link for the Jupyter Notebook and PowerBI Dashboards

GitHub Link

THANK YOU!

Prepared by Vedansh Chauhan