Report on Attrition Dataset

Analysis and prediction of Attrition

Shubham Koshti

Nature of Data set-

- **Objective**

1. To conduct an Exploratory Data Analysis and present the subsequent visualization via plots.
2. Uncover the factors that contribute to the attrition.
3. To implement classification algorithms for attrition prediction and present rationale supporting the output scores for each model.

- **Data Set Information**

Attrition data contains 1470 rows and 35 feature across All Department.

- **Library Used**

- Numpy
- Pandas
- Matplotlib
- Seaborn
- sklearn

Feature selection -

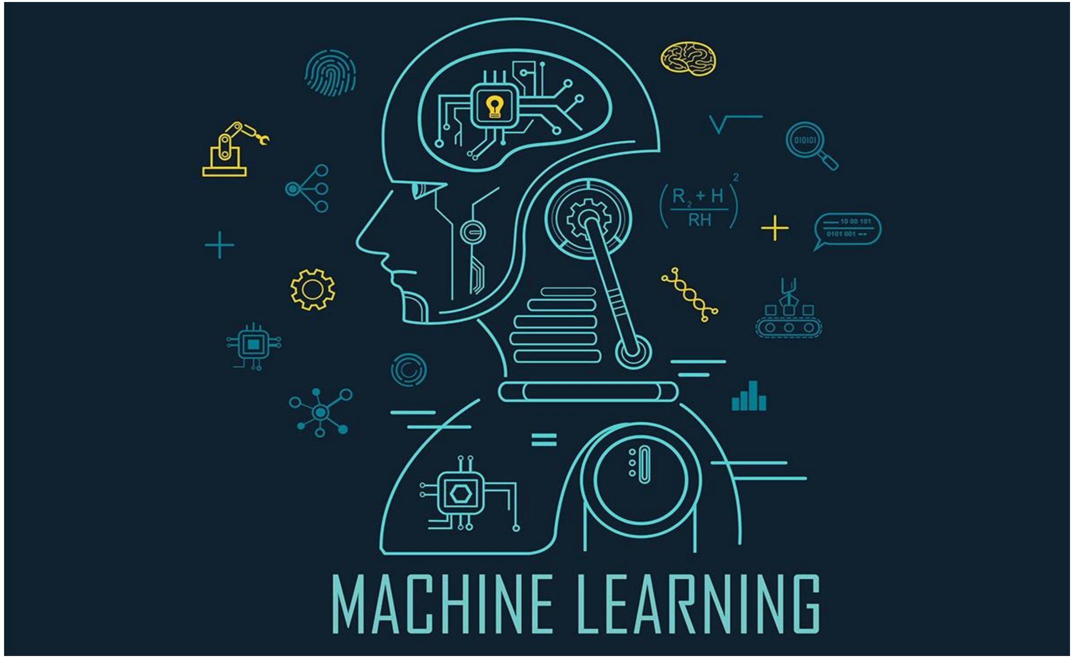
There are some attributes that are meaningless to the attrition prediction. ”Employee Count”, ”Over 18” and ”Standard Hours” have the same value for all the employees. Also, ”EmployeeNumber” according to its definition represents the employee”s ID. As these attributes don”t provide any information for prediction.

One hot encoding-

There are several categorical feature which we have to convert into numerical before modelling so i have used one hot encoding for converting them before modelling.

ML Tools-

1. Random Forest
2. Ada Boost
3. XG Boost



EDA-

Gender: More male employees.

Education: Most employees belongs to education 3 or 4.(Information about dataset is not given so unable to understand its meaning)

Marital Status: Mostly Married.

Business Travel: Most employees rarely travel.

Job Content & Department: Most employees studied in the life science or medical field, and the company have more employees in research, Laboratory and sales department.

Job Level: Most employees are in lower levels without stock option or only a few.

Workload: Most employees do not work overtime.

Satisfaction: Most employees have a high level of satisfaction, both with their work environment and workplace relationships, as well as a good work-life balance

.

EDA-

1.Employees who travel frequently have higher attrition rate.

2.Employees who have education in technical Degree and Human Resources have higher attrition rate.

3.Employees who are in sales and HR department have high attrition rate.

4.Employees who are in sales job role have high attrition rate.

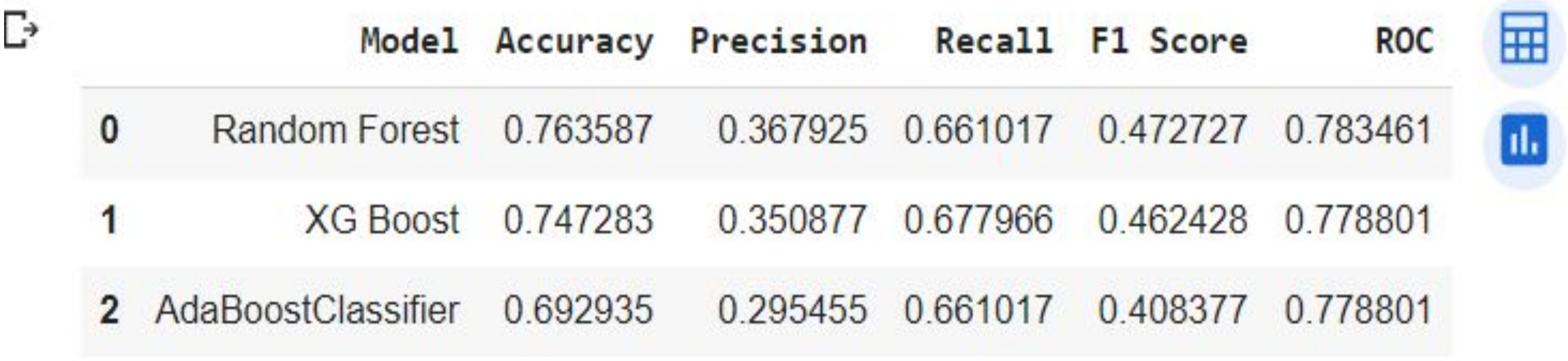
5.Employees whose Matiral Status is single have higher attrition rate.

6.Employees who often work overtime have higher attrition rate.

Methodology-

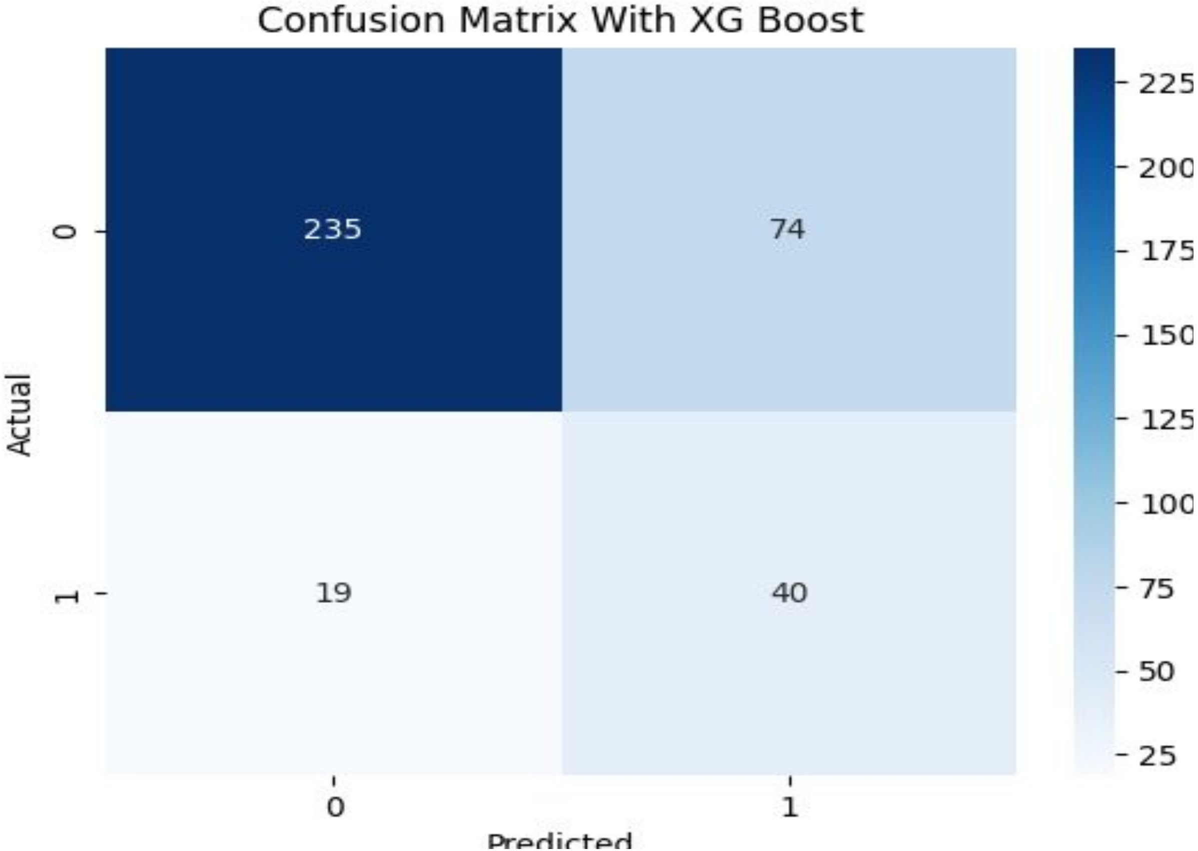
1. Check for multicollinearity, there is high multicollinearity so i am using tree based ensemble methods which will help us to deal with this.
2. To handel imbalance dataset I have used undersampling which was giving me best among all technique because data is so less that SMOTE can not work well.
3. Then I have Build the model after test train split and out of all model XG boost is giving me best results

Results



	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Random Forest	0.763587	0.367925	0.661017	0.472727	0.783461
1	XG Boost	0.747283	0.350877	0.677966	0.462428	0.778801
2	AdaBoostClassifier	0.692935	0.295455	0.661017	0.408377	0.778801

Confusion Matrix-



Comments-

1. Since I was handling with imbalance classes I tried multiple methods and find that undersampling was giving me best result.
2. I have map No attrition to 0 Yes attrition to 1.
3. Since our more focused on yes attrition thats why recall will be suitable metric for this.
4. Out of all model XG Boost is giving me best results accuracy- 75% and recall - 68 %

Thank You