Project 1 – Employee Attrition Prediction

Milestone 3 : White Paper

****

College of Science and Technology, Bellevue University

DSC680-T301 Applied Data Science (2245-1)

Sashidhar Bezawada

March 30 , 2024

Topic : Employee Attrition Prediction

- ## Introduction

Organizations are realizing that employees are valuable assets. Employee Attrition is a major concern for organizations because the functioning of the organization entirely depends on the pool of employees. Attrition can impact productivity, team dynamics, and overall company performance. Attrition rate defines the organization's image. Higher the attrition rate, the organization has to face some incurred costs to recruit, induct, placement and train the employee.

- ## Business Problem

Develop a machine learning model to address the below:

- Analyze the reason for employee attrition.
- Predict employee attrition within an organization.

- ## Datasets

I am using the dataset is from Kaggle. This is a fictional dataset created by IBM data scientists, which can help uncover the factors that lead to employee attrition.

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv

Below Table lists the attributes & descriptions of the dataset used in this project. Total number of records in the dataset: 1470.

Sl no	Attribute Name	Attribute Description
1	**Age**	The age of the employee.
2	**Attrition**	Whether the employee has left the company or is still employed.
3	**BusinessTravel**	The frequency of the employee's business travel (e.g., "Travel Rarely," "Travel Frequently," "Non-Travel").
4	**DailyRate**	The daily salary rate of the employee.
5	**Department**	The department in which the employee works (e.g., "Sales," "Human Resources," "Research & Development").
6	**DistanceFromHome**	The distance between the employee's home and workplace.
7	**Education**	The level of education the employee has achieved.
8	**EducationField**	The field of study for the employee's education.
9	**EmployeeCount**	The number of employees with the same count (possibly constant for all records).
10	**EmployeeNumber**	A unique identifier for each employee.
11	**EnvironmentSatisfaction**	The employee's satisfaction with their working environment.
12	**Gender**	The gender of the employee.
13	**HourlyRate**	The hourly salary rate of the employee.

14	**JobInvolvement**	How engaged the employee is in their current job.
15	**JobLevel**	The level of the employee's job within the organization.
16	**JobRole**	The specific role or position the employee holds.
17	**JobSatisfaction**	The employee's satisfaction with their job.
18	**MaritalStatus**	The marital status of the employee.
19	**MonthlyIncome**	The monthly salary of the employee.
20	**MonthlyRate**	The monthly rate at which the employee is paid.
21	**NumCompaniesWorked**	The number of companies the employee has worked for previously.
22	**Over18**	Whether the employee is over 18 years old.
23	**OverTime**	Whether the employee works overtime.
24	**PercentSalaryHike**	The percentage increase in the employee's salary.
25	**PerformanceRating**	The employee's performance rating.
26	**RelationshipSatisfaction**	The employee's satisfaction with their relationships at work.
27	**StandardHours**	The standard number of working hours for the employee.
28	**StockOptionLevel**	The level of stock options the employee has.
29	**TotalWorkingYears**	The total number of years the employee has worked.
30	**TrainingTimesLastYear**	How many times the employee received training last year.
31	**WorkLifeBalance**	How well the employee perceives their work-life balance.
32	**YearsAtCompany**	The number of years the employee has been with the company.
33	**YearsInCurrentRole**	The number of years the employee has been in their current role.
34	**YearsSinceLastPromotion**	The number of years since the employee's last promotion.
35	**YearsWithCurrManager**	The number of years the employee has been managed by their current manager.

- ## Exploratory Data Analysis

The dataset doesn't have any duplicates or missing values.The category attributes were also converted to numerical types using one-hot encoding. Most of the columns seem to be poorly correlated with one another. Generally when making a predictive model, it would be preferable to train a model with features that are not too correlated with one another so that we do not need to deal with redundant features. All the variables were explored for numeric and non-numeric separately. Three Columns were dropped as they had same value across all observations. For eight Variables, Labelling of Categories in Numerical Features were performed.

After the dataset was cleansed and transformed, different exploration were performed to gain a better understanding of the data especially in relation to the target variable ('Attrition'). In assessing the 'Attrition' variable's class distribution, it became apparent that the classes were highly imbalanced in their representation of the employees, as visualized in **Appendix**. The value 'No' had ten times as many records represented in the dataset compared to the value 'Yes', and since 'Yes' was our main value of interest, this imbalance represented a concern for the modeling portion of the project. Therefore, to handle this before modeling, I used a concept called SMOTE to generate new and synthetic data we used for training our model. The strategy helped to ensure

that bias was not introduced towards the majority class during the predictive modeling steps.

Besides investigating how to predict whether an employee will leave their organization, I also wanted to determine the factors which were important in the employees' decisions. As per the survey conducted by Mckinsey, employees felt the need to search for other jobs due to inadequate compensation , work–life balance , valued by manager, growth in the career, Job satisfaction. In the dataset, there are Variables which correlate with these reasons and I created charts to compare these factors with the McKinsey survey. These Charts are also shown in **Appendix.**

The variables where I noticed a significant difference between the two groupings was for job satisfaction and monthly income. For employees who did quit, their job satisfaction scores mostly leaned towards falling into the 1(low) and 3(high) levels. Whereas for the employees who did not quit, they scored their job satisfaction at the highest scores of 2(medium) and 4(Very High). This data showed that the employees in the Attrition 'Yes' group were either dissatisfied or relatively satisfied which are uneasy places to stand. For Monthly income, the distributions of the variable were left-skewed for the two groups of employees; however, for the employees who did resign, they were mainly represented at the lower end of the monthlt income scale and barely had any representation greater than $10,000 . The employees who did not quit had representation across the entire scale from $11,000 to $20,000 per month, and had much higher counts at the higher values than the other group. Pay is the most important factor to an employee's lifestyle, and if an employee is not being paid enough to take care of their means, then it would make sense that these lower-paying employees would leave to pursue better financial situations.

1. Splitting the dataset (SMOTE, Training and Testing) .
2. Regression Model Selection (KNN, Naïve Bayes, Decision tree, Random Forest, XGBoost, GradientBoost Classifier).

For model selection, 6 models which are appropriate for a classification model were used to evaluate performance accuracy. The models evaluated are listed below :

- KNN - Straight forward pattern recognition model which allows the testing of several k values and leaf sizes to determine the best performance
- Naïve Bayes - Calculates the possibility of whether a data point belongs within a certain category
- Decision tree - A decision tree is a supervised learning algorithm that performs strong in classification problems
- Random Forest - Expands beyond a decision tree by constructing multiple decision trees to remediate forcing a binary decision
- XGBoost - an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning model
- GradientBoost Classifier - Gradient boosting works by building simpler (weak) prediction models sequentially where each model tries to predict the error left over by the previous model.

Post-data exploration, the models were performed on all of the twenty-nine features in the dataset, . The model performance was then evaluated based on the models' accuracies and confusion matrices. Precision and recall can be derived from a confusion matrix; recall helps to determine the ability to find all relevant instances of a class in a data set, whereas precision is used to determine the proportion of data points that the model says exists within the relevant class were indeed relevant. The F1 score is calculated as the harmonic mean of precision and recall. The AUC measures the probability that the model will assign a randomly chosen positive instance a higher predicted probability compared to a randomly chosen negative instance. All accuracy, precision, recall, F1-Score, AUC results can be viewed in Appendix.

The Gradient Boosted Classifier model had the strongest performance with 81.86% accuracy; it also had the second-highest recall score for the 'No' class of Attrition at 35.52%. For the 'Yes' class of Attrition, it also had the highest precision score of 46.55% which was relatively higher than the scoring of the other models. This measure showed that 46.55% of the employees who were predicted as resigning were correctly predicted. Given the count imbalance in the target classes, I think these scores are relatively good and that the models were able to compensate for the lack of original data for the resigned employees.

- **Ethical Considerations**

Legal compliance: This dataset is a fictional dataset from Kaggle and doesn't have any legal issue.
Employee Privacy: This dataset doesn't contain any PII-related information (Personal Identifiable Information).This data research is not going to harm any privacy.
Unfair practices: The biggest ethical challenge is how the Organization is going to use the Attrition model and not treat employees unfairly based on some factors that contribute to attrition like - age ,salary, etc.

- **Assumption**

I assumed that the employees recorded in the dataset were full-time employees who worked a standard of nine hours. Full-time employees are usually the main focus of an organization. With this assumption, I made choices in the project that would pertain to why an employee would leave an organization, which hopefully is what the dataset set out to accomplish as well.

- **Challenges/Issues**

This is a simulated dataset and has only 35 attributes. This might be an imbalanced dataset and a small volume. To improve the effectiveness, SMOTE has been used.

- **Conclusion**

  People are going to spend a majority of their lifetime as an employee, then it makes sense that they would want that time to be well-spent and somewhat-enjoyed. The dataset for this project conveyed that there are many features that factor into an employee's decision to leave their job; two of the important ones being job satisfaction level and income. Employees want to enjoy their jobs, and they also want to be paid enough to compensate for the time, energy and effort that they put into their work. If these needs are not being met, then it is highly likely that an employee will choose to leave their organization rather than stay and feel dissatisfied. With the use of the Gradient Boosted Classifier model, the features mentioned above (along with others) can even be utilized from a predictive perspective to identify at-risk employees ahead of time and potentially prevent them from resigning. The results from these predictions could help organizations save money on staff replacement and productivity loss, and also hopefully decrease turnover rates.

- **Future Uses**

  This information can be used as supporting material by Organization & HR department and take appropriate steps to retain the Employees.

- **Recommendations**

  In the current model, feature reduction is not explored to it's full extent. And that can be considered to make this model better.

- **Implementation Plan**

  As we already discussed in the methodology section about some of the implementation details. So, the language used in this project is Python programming. We're running python code in anaconda navigator's Jupyter notebook. Jupyter notebook is much faster than Python IDE tools likePyCharm or Visual studio for implementing ML algorithms. The advantage of Jupyter notebook is that while writing code, it's really helpful for Data visualization and plotting some graphs like histogram

- **References**

Dataset : https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

1.” Five Hidden Costs Of Employee Attrition” - Written by Lisa Wallace https://www.forbes.com/sites/forbeseq/2023/03/21/five-hidden-costs-of-employee-attrition/?sh=391d2a0c62f4

2. “A STUDY ON EMPLOYEE ATTRITION AND RETENTION WITH REFERENCE TO EVRON IMPEX” by Dr. B. Merceline Anitha , https://www.researchgate.net/publication/370471266_A_STUDY_ON_EMPLOYEE_ATTRITION_AND_RETENTION_WITH_REFERENCE_TO_EVRON_IMPEX

3. https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/chart-of-the-day/the-great-attrition-stems-from-a-great-disconnect

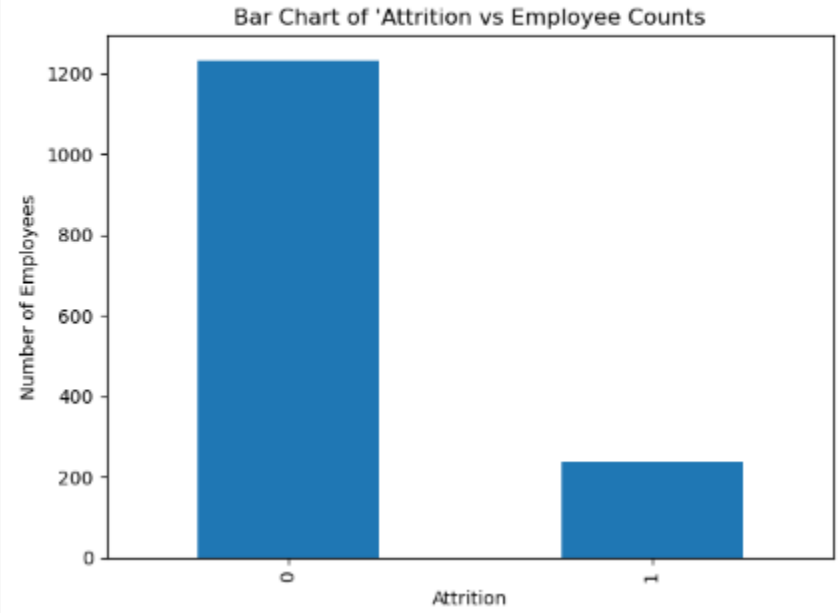
4. “Employee Attrition: Meaning, Impact & Attrition Rate Calculation” Written by Suzanne Lucas, https://www.aihr.com/blog/employee-attrition/

5. https://atlanticpayroll.us/business-basics/employee-attrition-vs-employee-turnover/

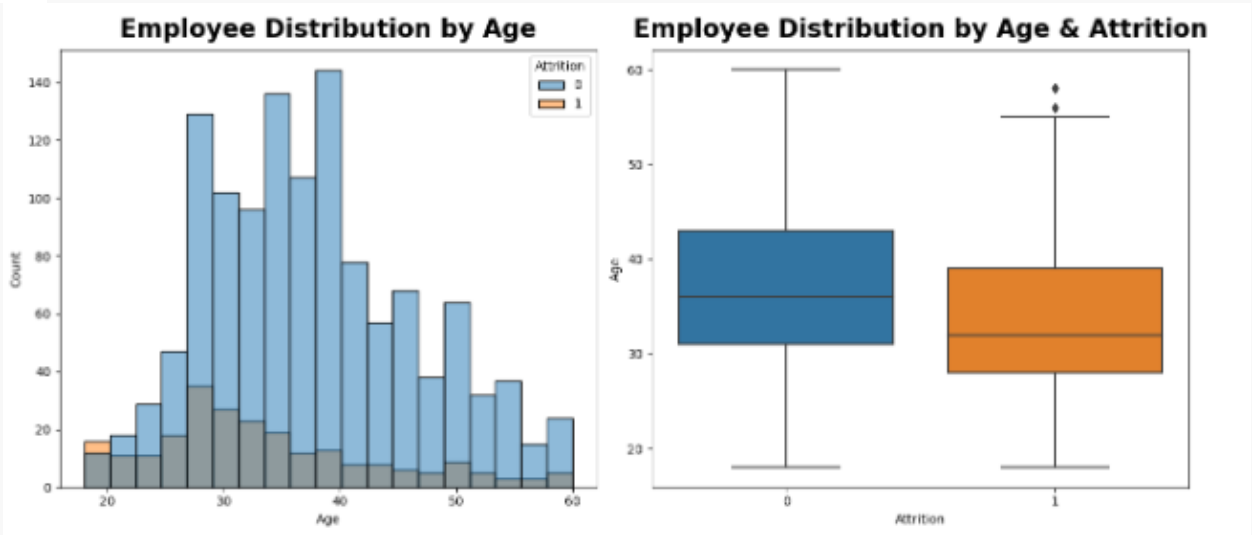
6. https://hbr.org/1973/07/why-employees-stay

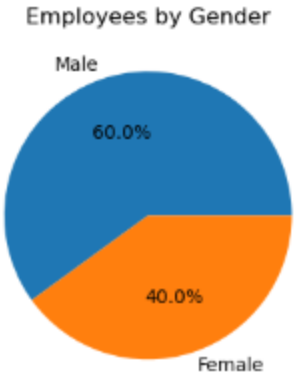
- **Appendix A - Target Class Distribution**

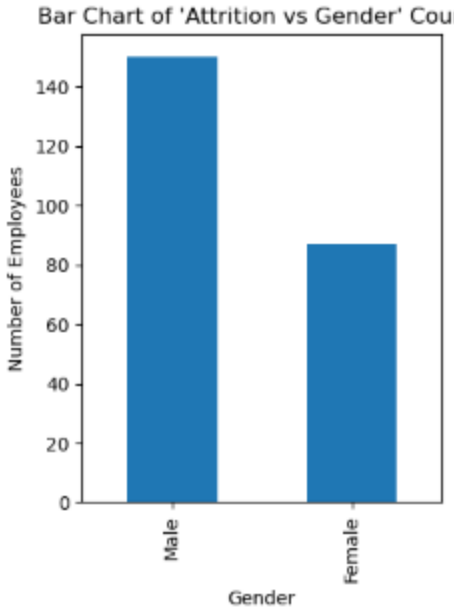
```
Attrition
0    1233
1     237
Name: count, dtype: int64
```

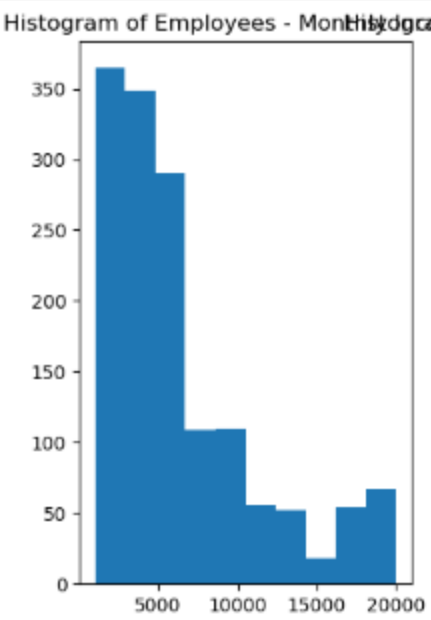


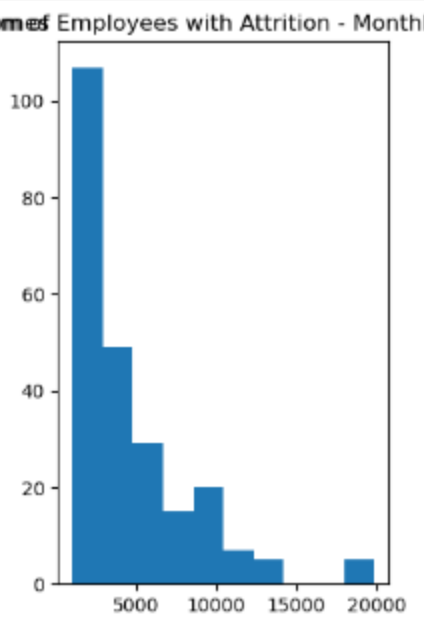
- **Appendix B - Feature Importance Visualizations**

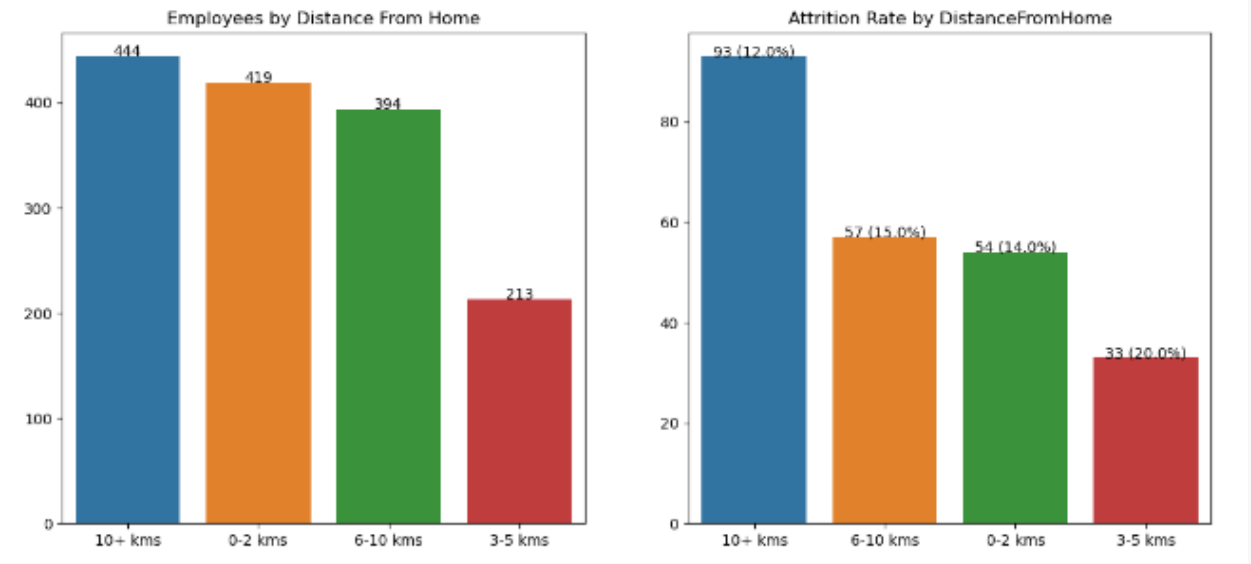


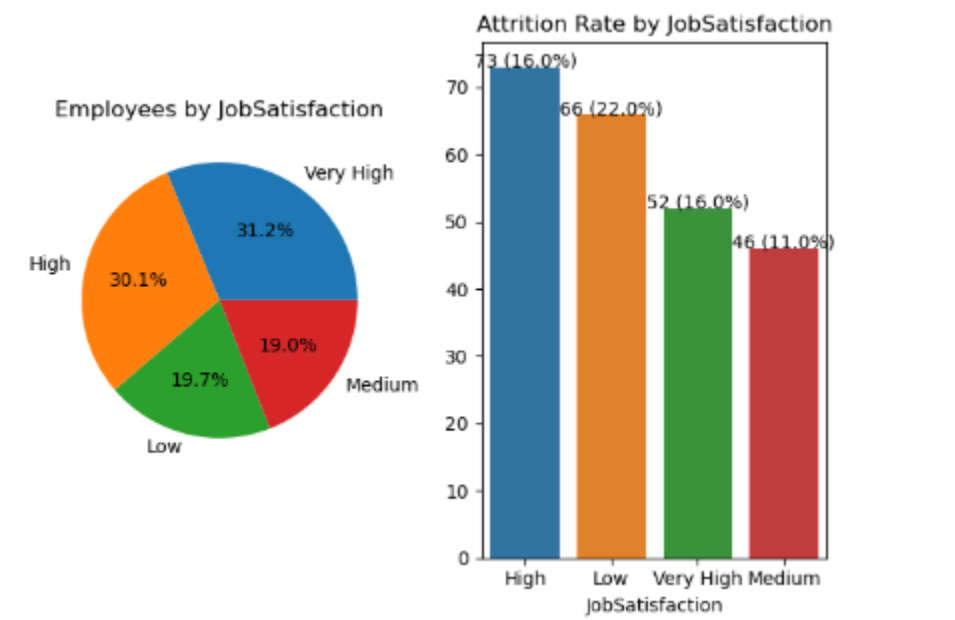

Employees by Gender
Male
60.0%
40.0%
Female

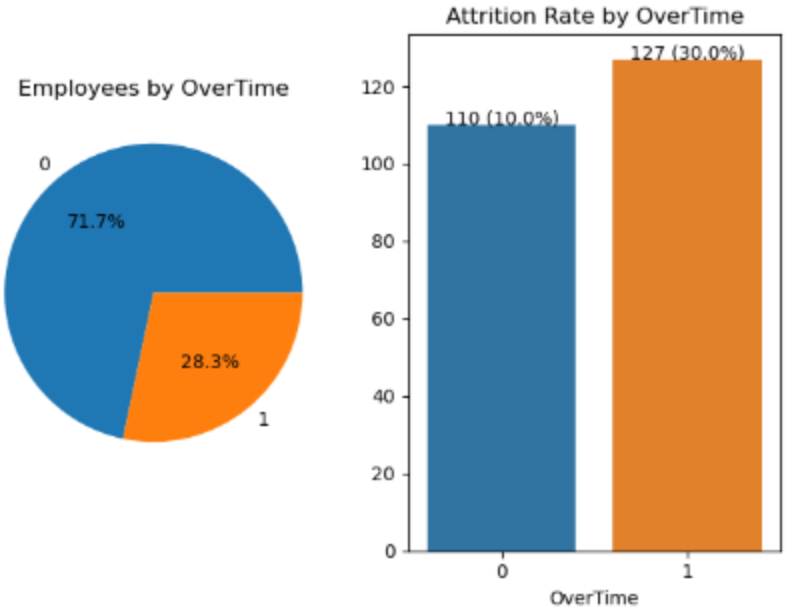

Bar Chart of 'Attrition vs Gender' Counts
Number of Employees
0
20
40
60
80
100
120
140
Male
Female
Gender

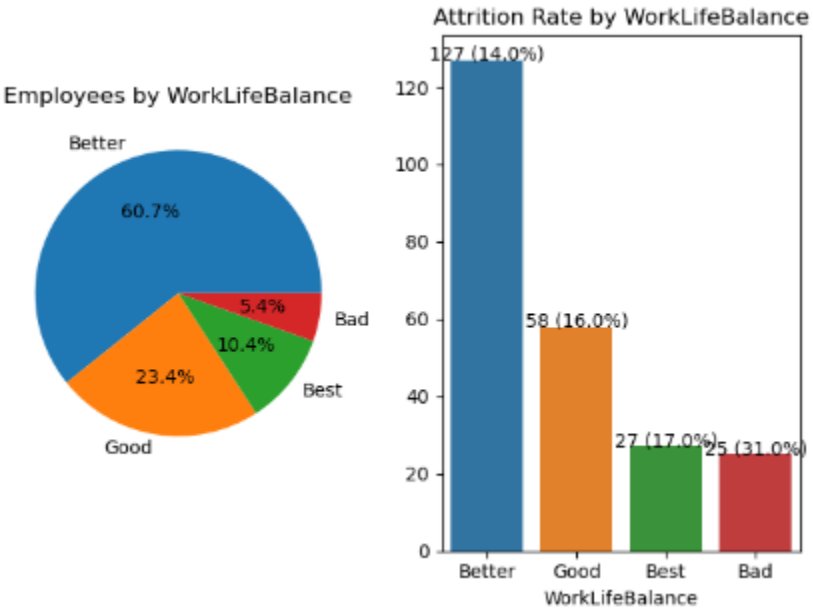

Histogram of Employees - Monthly Incomes
0
50
100
150
200
250
300
350
5000
10000
15000
20000

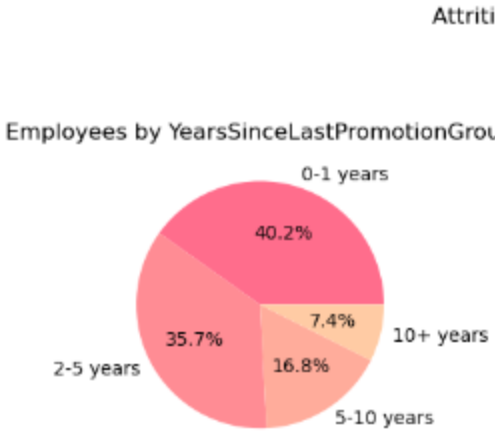

Histogram of Employees with Attrition - Monthly Incomes
0
20
40
60
80
100
5000
10000
15000
20000

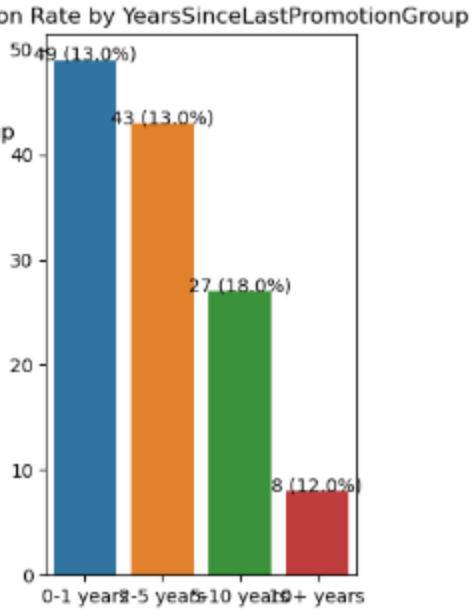

Employees by Distance From Home
444
419
394
213
0
100
200
300
400
10+ kms
0-2 kms
6-10 kms
3-5 kms
Attrition Rate by DistanceFromHome
93 (12.0%)
57 (15.0%)
54 (14.0%)
33 (20.0%)
0
20
40
60
80
10+ kms
6-10 kms
0-2 kms
3-5 kms

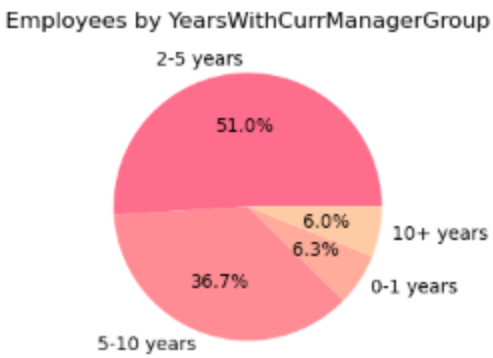

Employees by JobSatisfaction
Very High
31.2%
High
30.1%
19.0%
Medium
19.7%
Low
Attrition Rate by JobSatisfaction
73 (16.0%)
66 (22.0%)
52 (16.0%)
46 (11.0%)
0
10
20
30
40
50
60
70
High
Low
Very High
Medium
JobSatisfaction

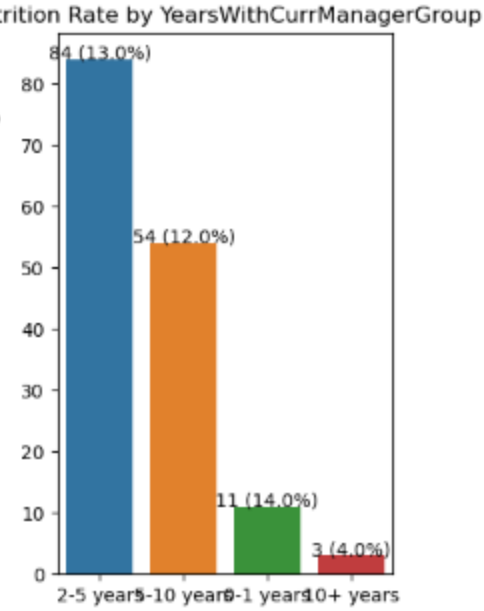

Employees by OverTime
0
71.7%
28.3%
1
Attrition Rate by OverTime
127 (30.0%)
110 (10.0%)
120
100
80
60
40
20
0
0
1
OverTime


Employees by WorkLifeBalance
Better
60.7%
5.4%
Bad
10.4%
23.4%
Best
Good
Attrition Rate by WorkLifeBalance
127 (14.0%)
120
100
80
60
58 (16.0%)
40
27 (17.0%)
25 (31.0%)
20
0
Better
Good
Best
Bad
WorkLifeBalance


Employees by YearsSinceLastPromotionGroup
0-1 years
40.2%
7.4%
10+ years
35.7%
2-5 years
16.8%
5-10 years


Attrition Rate by YearsSinceLastPromotionGroup
49 (13.0%)
43 (13.0%)
27 (18.0%)
8 (12.0%)
0
10
20
30
40
50
0-1 years
2-5 years
5-10 years
10+ years


Employees by YearsWithCurrManagerGroup
2-5 years
51.0%
6.0%
10+ years
6.3%
0-1 years
36.7%
5-10 years


Attrition Rate by YearsWithCurrManagerGroup
84 (13.0%)
54 (12.0%)
11 (14.0%)
3 (4.0%)
0
10
20
30
40
50
60
70
80
2-5 years
5-10 years
0-1 years
10+ years

- **Appendix C - Model Evaluation Results**

	Model	K-Nearest Neighbor(SMOTE)	Random Forest (SMOTE)	Naive Bayes(SMOTE)	Decision Tree (SMOTE)	XGBoost (SMOTE)	Gradient Boosted Classifier (SMOTE)
0	Test_accuracy :	0.73469	0.73583	0.62925	0.71088	0.81859	0.81859
1	Test_Precision_Score :	0.23026	0.32159	0.26027	0.25359	0.46491	0.46551
2	Test_F1_Score :	0.23026	0.38522	0.36750	0.29363	0.39850	0.40298
3	Test_Recall_Score :	0.23026	0.48026	0.62500	0.34868	0.34868	0.35526
4	Test_AUC_Score :	0.53499	0.63465	0.62757	0.56749	0.63256	0.63516
5	Test_Balanced_Accuracy_Score:	0.53499	0.63465	0.62757	0.56749	0.63256	0.63516

Results interpretation :

SMOTE uses a nearest neighbors algorithm to generate new and synthetic data we used for training our model.

Naive Bayes - Model accuracy is the lowest among the 6 Models.

K-Nearest Neighbor - Has good Accuracy Score, but has the lowest F1 & Recall scores. The False positive & False negative cases are higher than all models.

Random Forest - Accuracy score is good, however the model is not predicting the Attrition correctly (precision is low) and it identifies False negatives better than all (Recall Score is high).

Decision Tree - Accuracy score is low, however the model is not predicting the Attrition correctly (precision is low)

XGBoost - Accuracy is high as well as it is identifying the Attrition better. Also, it has high Recall & F1 Scores.

Gradient Boosted Classifier - Accuracy is highest of all . And it is identifying the Attrition with the highest Precision Score. It also has the highest Recall, F1 & AUC Scores.