Description

Project Statement:

Portobello Tech is an app innovator that has devised an intelligent way of predicting employee turnover within the company. It periodically evaluates employees' work details including the number of projects they worked upon, average monthly working hours, time spent in the company, promotions in the last 5 years, and salary level.

Data from prior evaluations show the employee's satisfaction at the workplace. The data could be used to identify patterns in work style and their interest to continue to work in the company.

The HR Department owns the data and uses it to predict employee turnover. Employee turnover refers to the total number of workers who leave a company over a certain time period.

As the ML Developer assigned to the HR Department, you have been asked to create ML Programs to

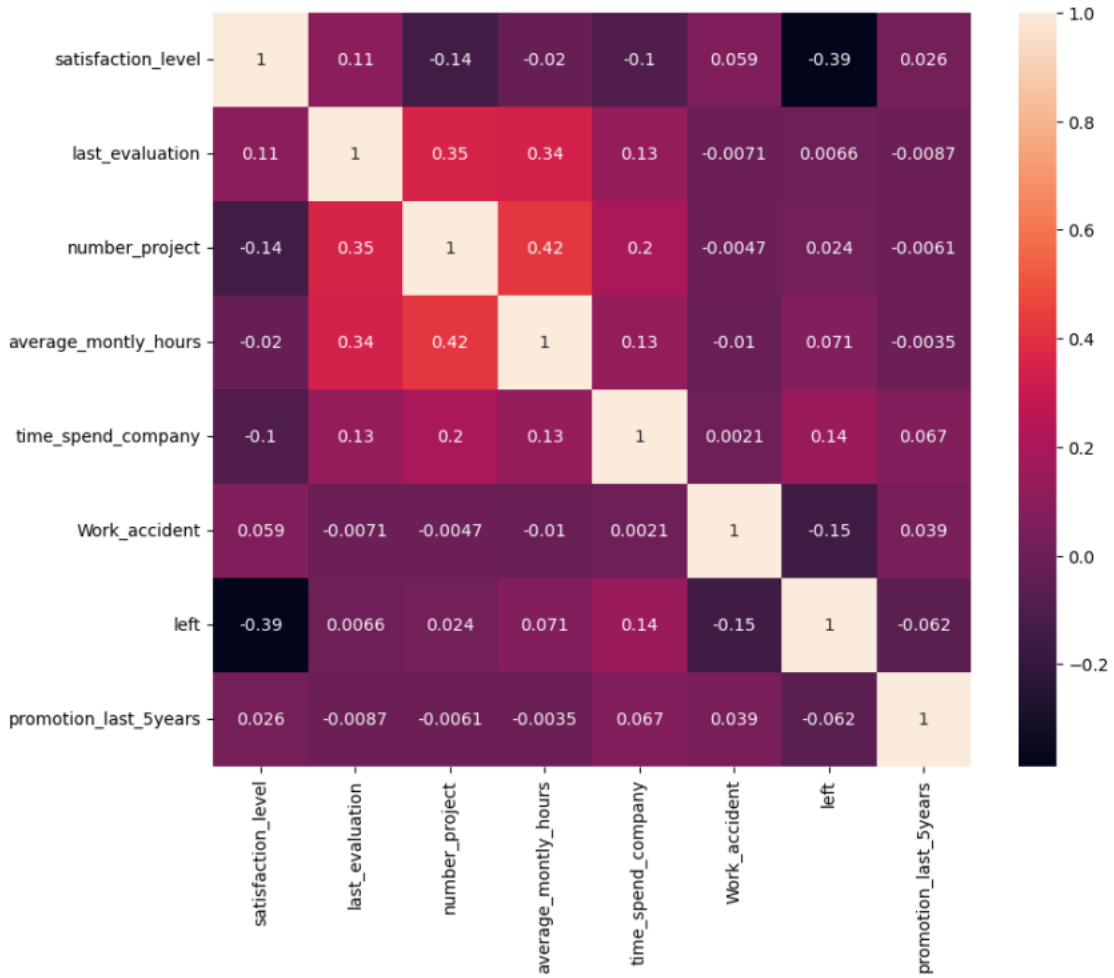
1. Perform data quality check by checking for missing values if any.

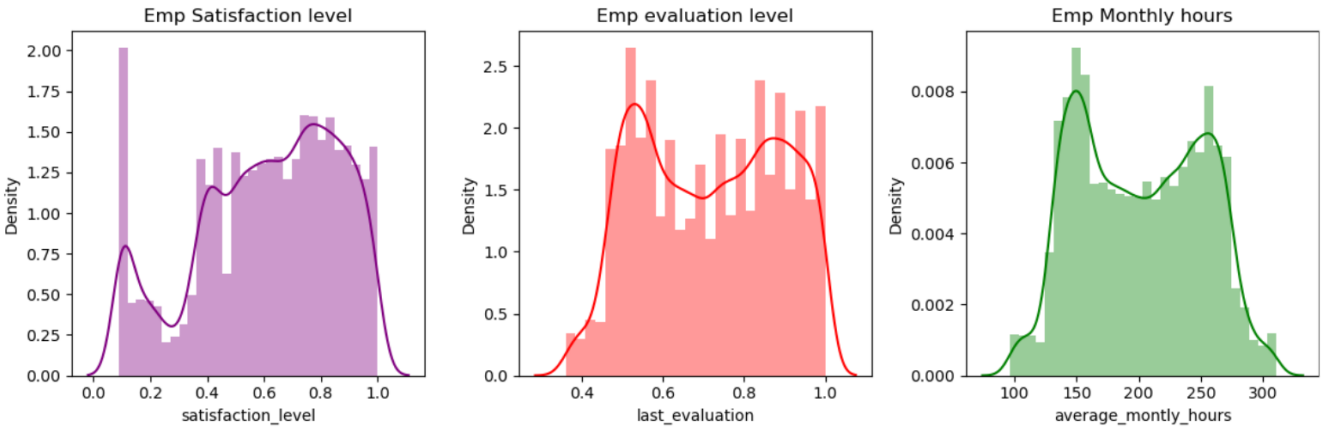
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   Work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   sales                  14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

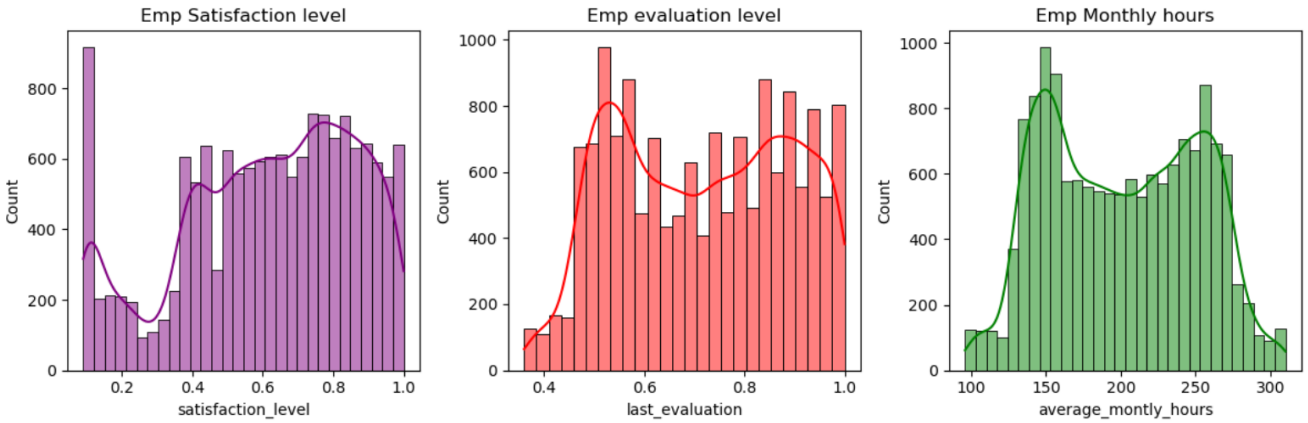
```
satisfaction_level       0
last_evaluation          0
number_project           0
average_montly_hours     0
time_spend_company       0
Work_accident            0
left                     0
promotion_last_5years    0
sales                    0
salary                   0
dtype: int64
```

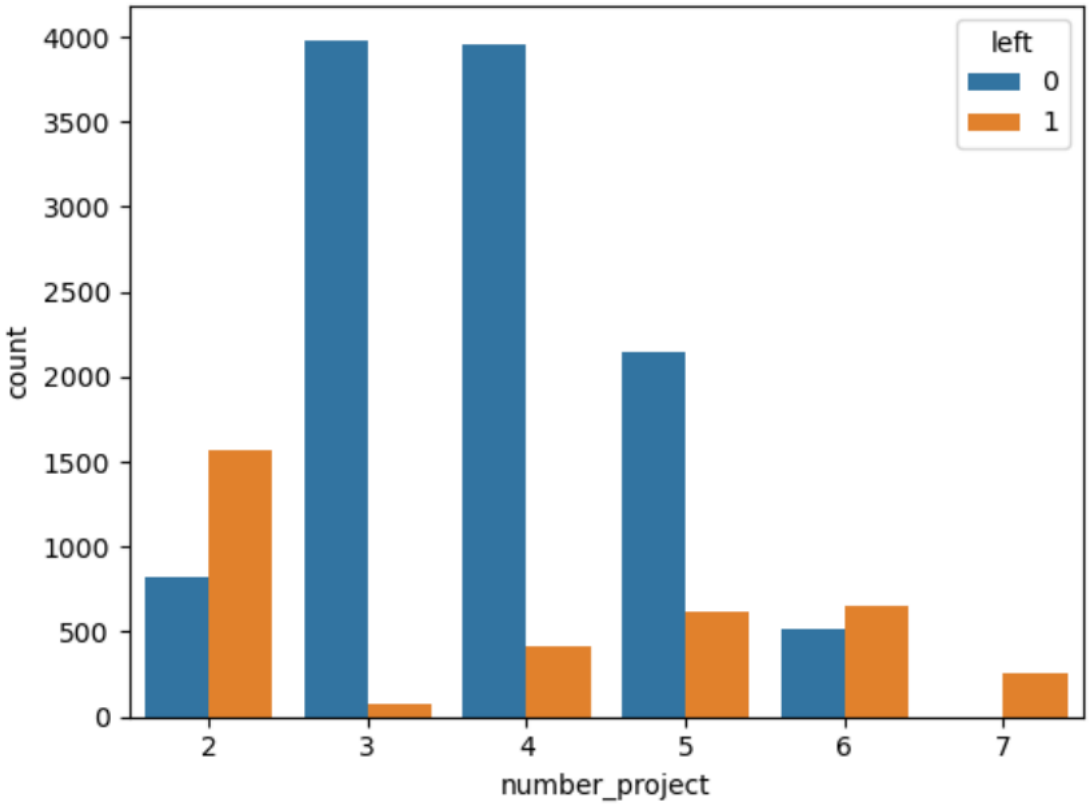
   - No missing values in the dataset

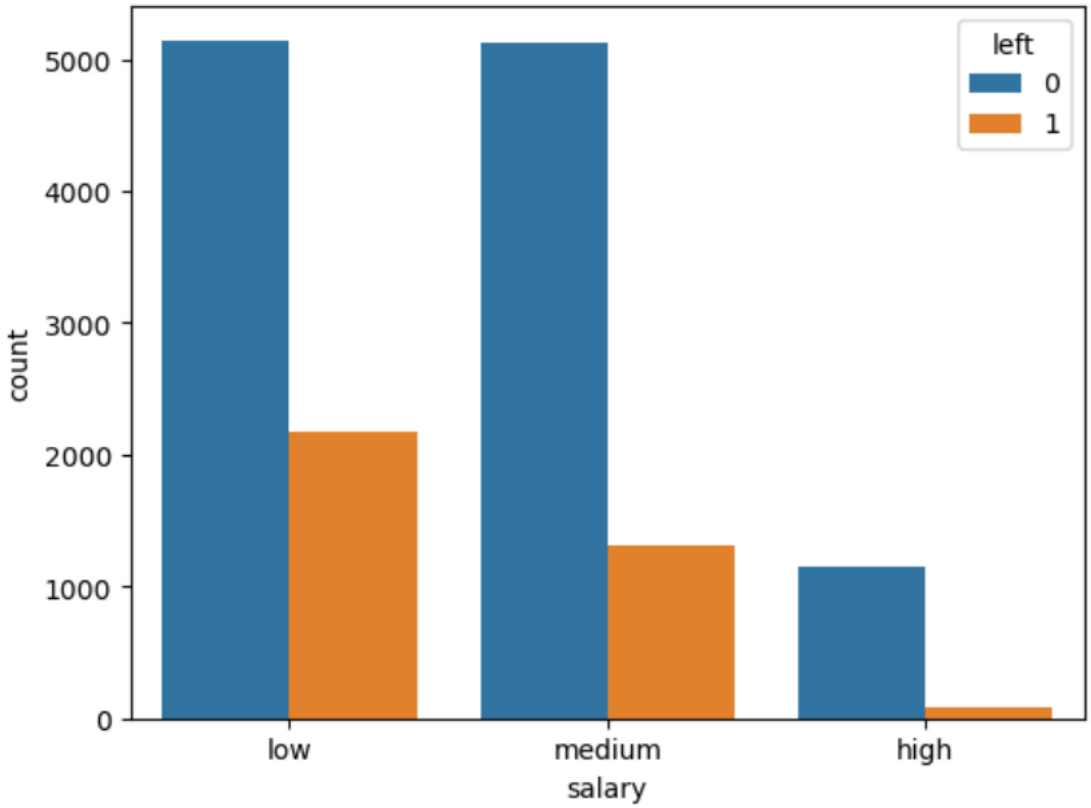
2. Understand what factors contributed most to employee turnover by EDA.

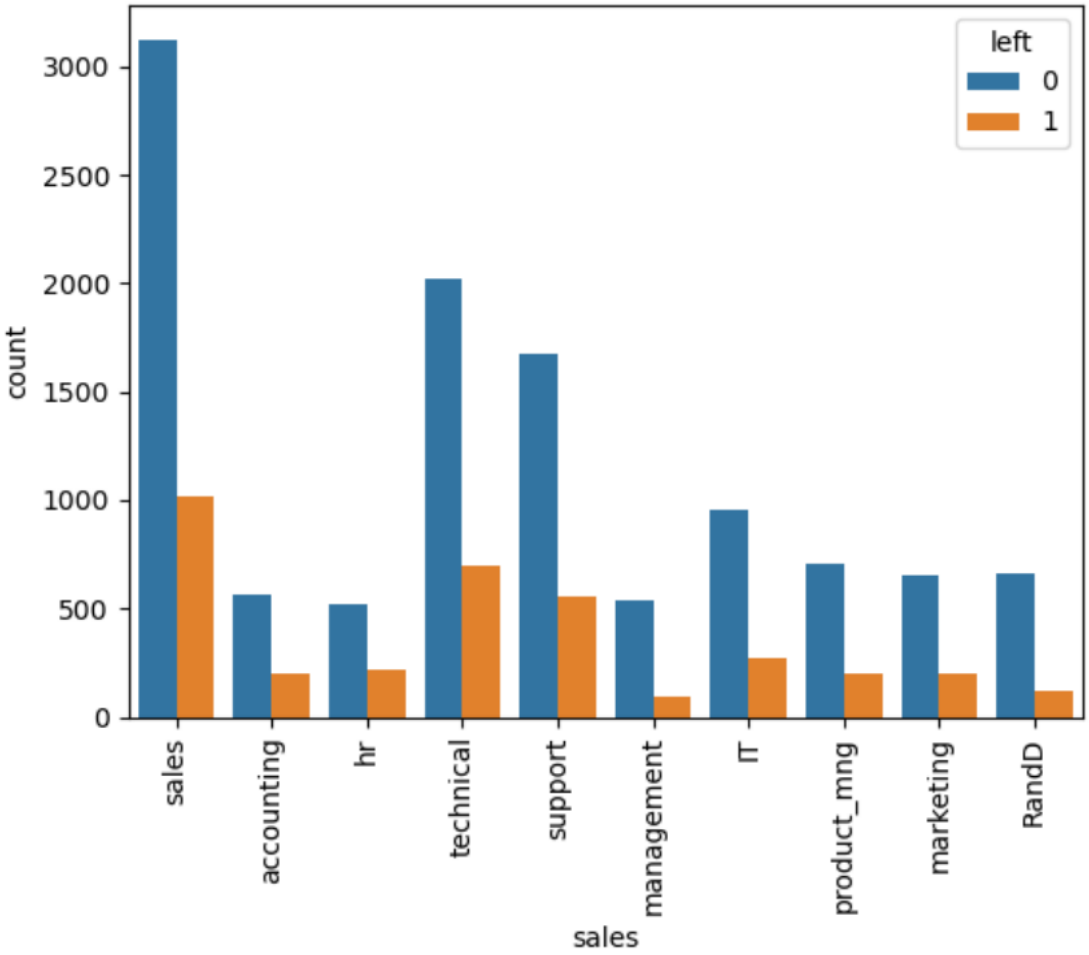

satisfaction_level
last_evaluation
number_project
average_montly_hours
time_spend_company
Work_accident
left
promotion_last_5years
1 0.11 -0.14 -0.02 -0.1 0.059 -0.39 0.026
0.11 1 0.35 0.34 0.13 -0.0071 0.0066 -0.0087
-0.14 0.35 1 0.42 0.2 -0.0047 0.024 -0.0061
-0.02 0.34 0.42 1 0.13 -0.01 0.071 -0.0035
-0.1 0.13 0.2 0.13 1 0.0021 0.14 0.067
0.059 -0.0071 -0.0047 -0.01 0.0021 1 -0.15 0.039
-0.39 0.0066 0.024 0.071 0.14 -0.15 1 -0.062
0.026 -0.0087 -0.0061 -0.0035 0.067 0.039 -0.062 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Emp Satisfaction level
Density
satisfaction_level
Emp evaluation level
Density
last_evaluation
Emp Monthly hours
Density
average_montly_hours

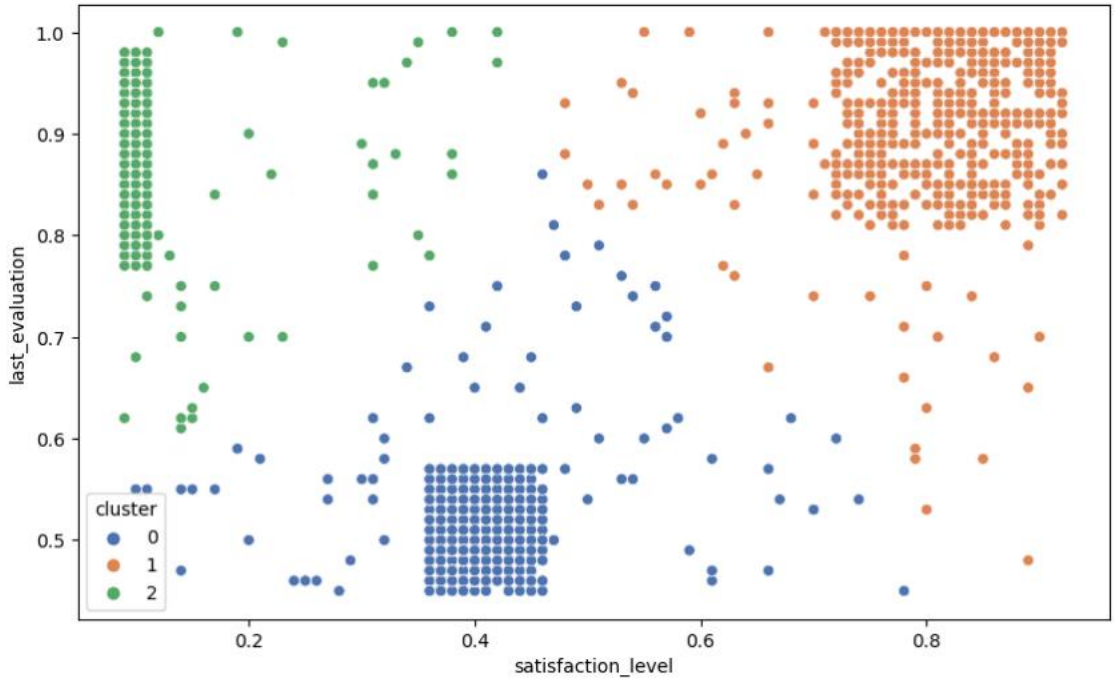

Emp Satisfaction level
Count
satisfaction_level
Emp evaluation level
Count
last_evaluation
Emp Monthly hours
Count
average_montly_hours


left
0
1
count
number_project
0
500
1000
1500
2000
2500
3000
3500
4000
2
3
4
5
6
7


left
0
1
count
salary
0
1000
2000
3000
4000
5000
low
medium
high



3. Perform clustering of Employees who left based on their satisfaction and evaluation.



- Employee left based on High Evaluation and Low Satisfaction level
- Employee left based on Low Evaluation and Average Satisfaction level
- Employee left based on High Evaluation and High Satisfaction level

```
0    1650
1     977
2     944
Name: cluster, dtype: int64
```

4. Handle the left Class Imbalance using SMOTE technique.

ales_IT	sales_RandD	sales_accounting	sales_hr	sales_management	sales_marketing	sales_product_mng	sales_sales	sales_support	sales_technical	salary_high
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0	0

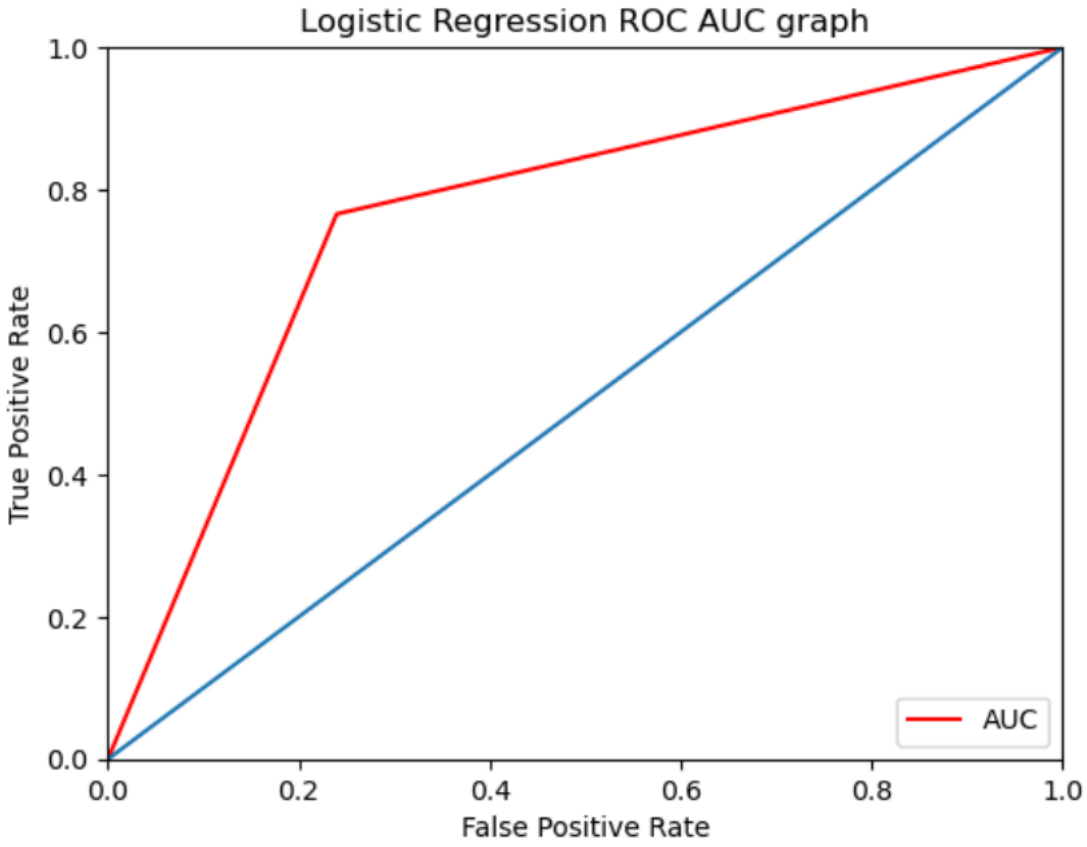
```
0    9137
1    9137
Name: left, dtype: int64
```

5. Perform k-fold cross-validation model training and evaluate performance.
6. Identify the best model and justify the evaluation metrics used.

```
Accuracy score:  0.7616666666666667
```

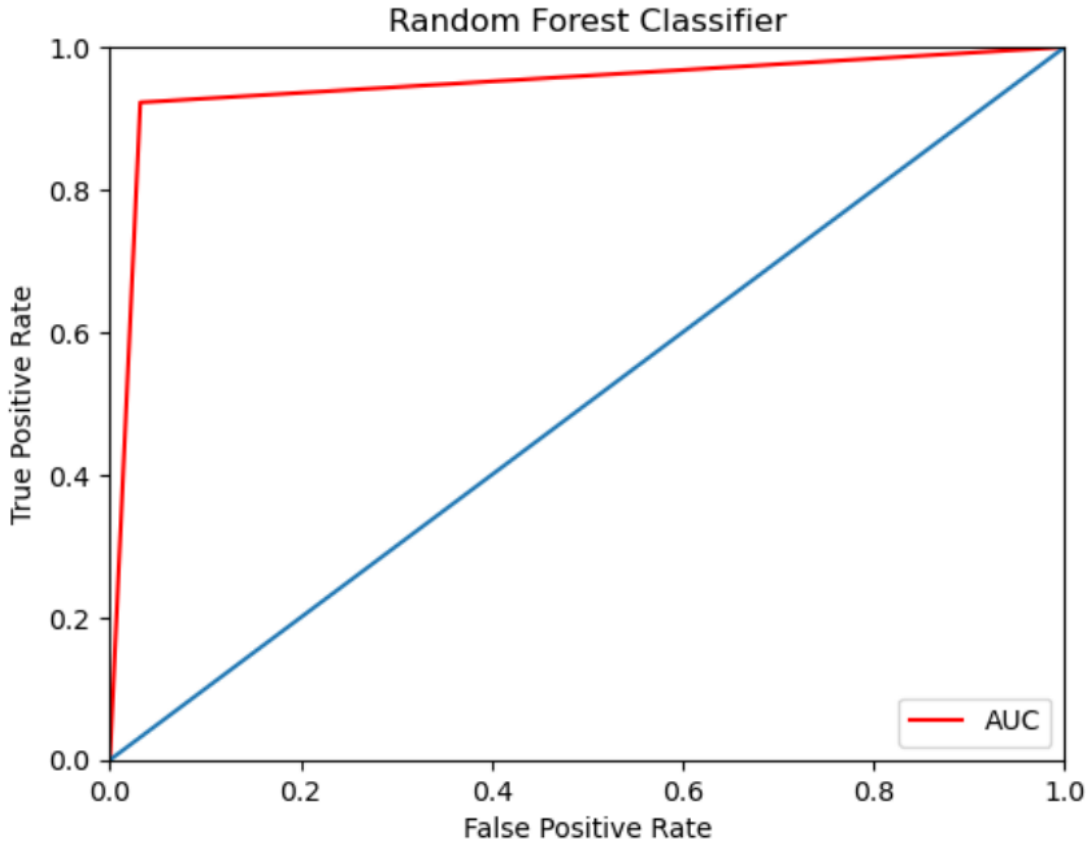
- Here, accuracy score is 76.16%

	precision	recall	f1-score	support
0	0.91	0.76	0.83	2291
1	0.50	0.77	0.60	709
accuracy			0.76	3000
macro avg	0.71	0.76	0.72	3000
weighted avg	0.81	0.76	0.78	3000



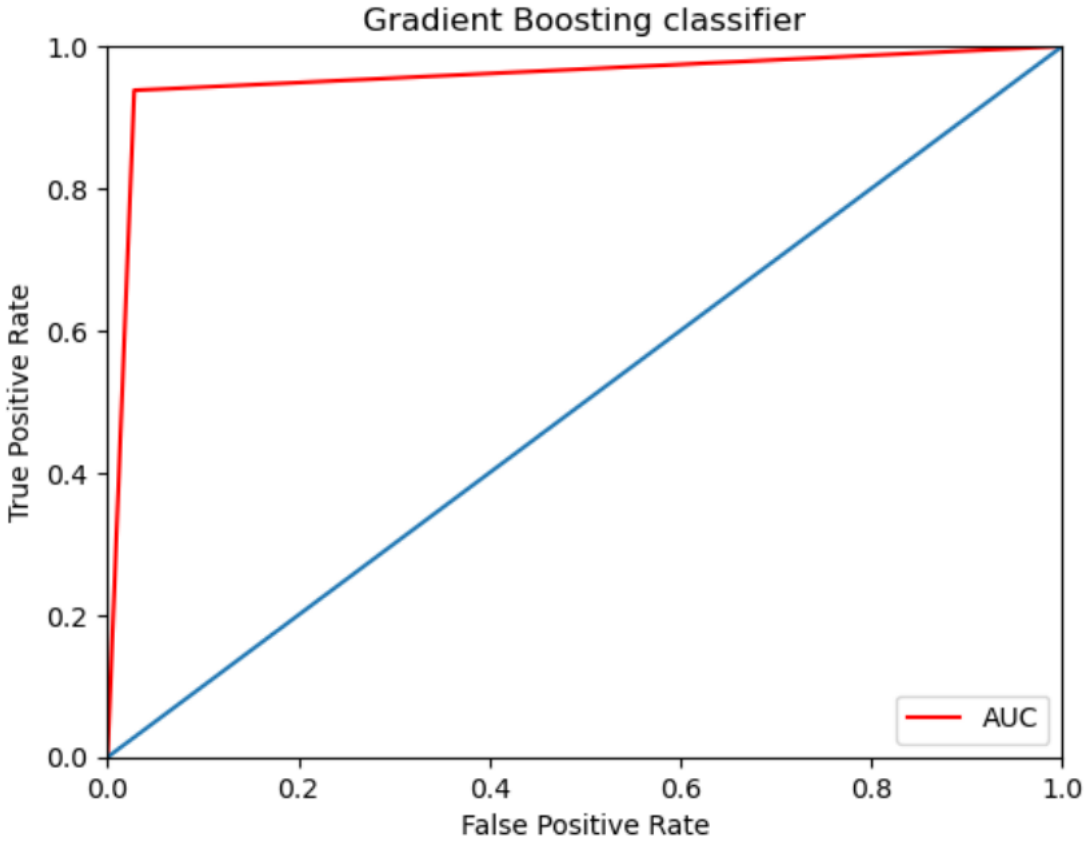
```
Accuracy score 0.9573333333333334
```

	precision	recall	f1-score	support
0	0.98	0.97	0.97	2291
1	0.90	0.92	0.91	709
accuracy			0.96	3000
macro avg	0.94	0.95	0.94	3000
weighted avg	0.96	0.96	0.96	3000



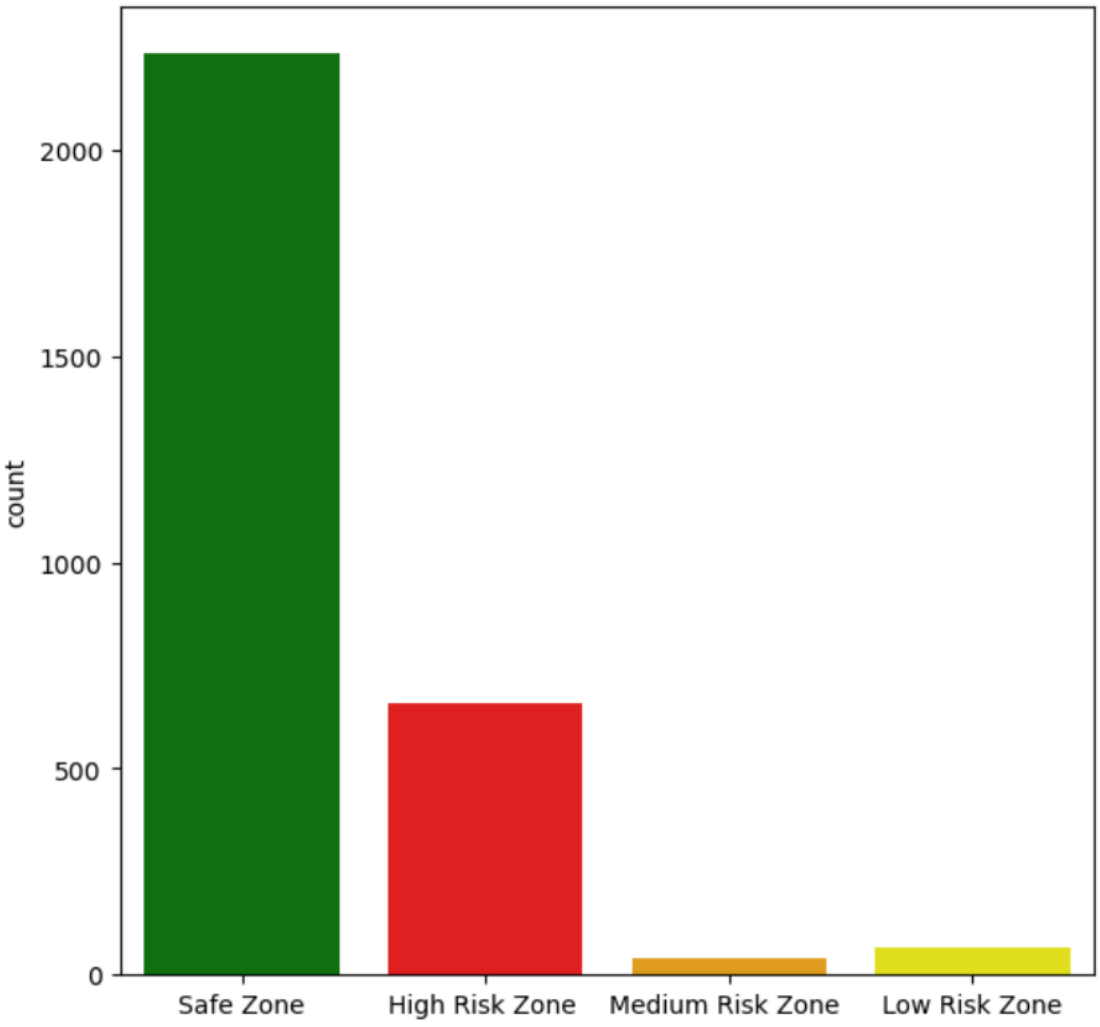
Accuracy score 0.964

	precision	recall	f1-score	support
0	0.98	0.97	0.98	2291
1	0.91	0.94	0.92	709
accuracy			0.96	3000
macro avg	0.95	0.96	0.95	3000
weighted avg	0.96	0.96	0.96	3000



- Here, the best fit models were Random Forest and Gradient Boosting classifiers

7. Suggest various retention strategies for targeted employees.


count
2000
1500
1000
500
0
Safe Zone
High Risk Zone
Medium Risk Zone
Low Risk Zone