HR Analytics: Employee Attrition

Abhi Arora
Shashannk Aggarwal
Sonu Bhandari

List of contents

1. Problem Statement and Objective
2. Data Preparation
3. Feature Extraction
4. EDA
5. Feature Selection
6. Normality Check
7. Model Building
8. Model Selection
9. Conclusion

Problem Statement

Our role is to identify the factors contributing to attrition of the employees and recommend possible solutions. The files contain complete staff utilization reports for all employees of the XYZ Corp. in two separate files for 2016-2017 (789, 115) and 2017-2018 (973, 115).

Another file contains all the attrition in the organization for the years 2015-18 with details such as reason of attrition along with other employee details. (293, 9)

Objectives:

- Identify factors influencing attrition
- Predict possible attritions
- Identify possible ways to retain high performers

Data preparation

- Replaced all the missing values (“-”) with NaN, for consistency
- Data-type changes:
  - Date columns to datetime
  - Employee Number (Termination) to object
  - Leave Hours to float
- Excluded all monthly utilization data
- Concatenated the utilization data of the two years along with termination data
  - Concatenated utilization datasets on “Employee No” and “Employee Name”
  - Then concatenated termination data on “Employee Name”
- Replaced all non-resigned statuses with Active
- For resigned employees, their last active profit centers were used to fill the missing profit centers after merging the two year’s data. Same was done for “Employee Location”, “People Group”, “Supervisor name” and “Employee Category”

Data preparation (contd.)

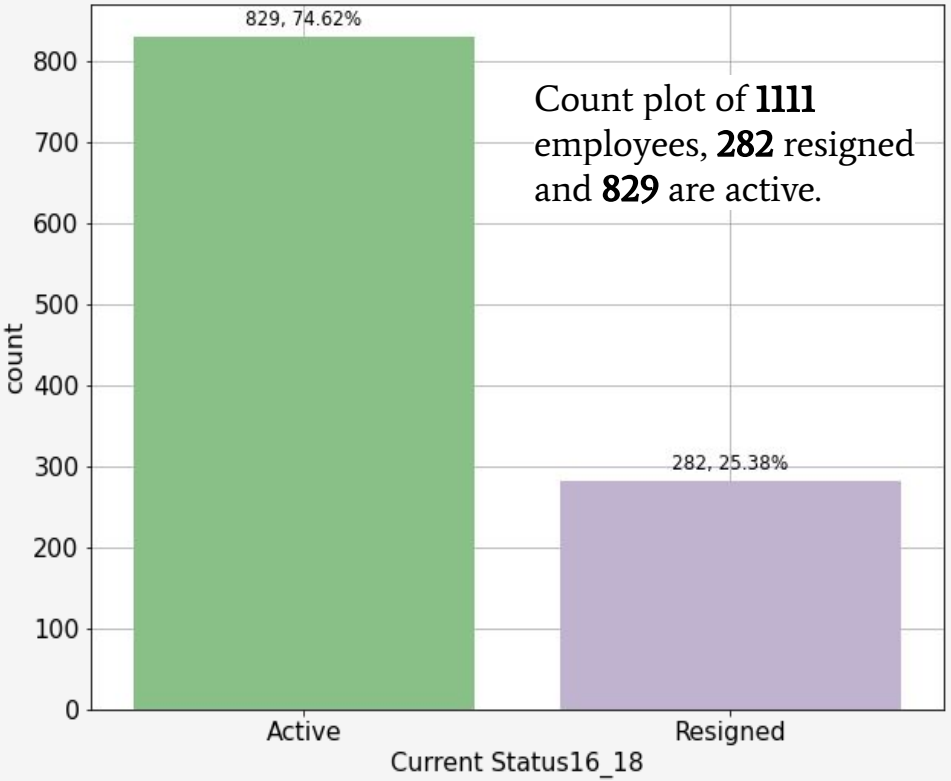
- Dropped “Emp Ref.” column from termination data since almost all values were missing (only 11 non-null values)
- The employees that were missing in the second year’s data were marked resigned based on their monthly utilization data of the previous year (Employee Nos: 96, 213, 248, 433)
- Any employee that is active or exists as a supervisor in the 17-18 dataset (Employee No: 5) is considered “Active” else “Resigned”

Feature Extraction

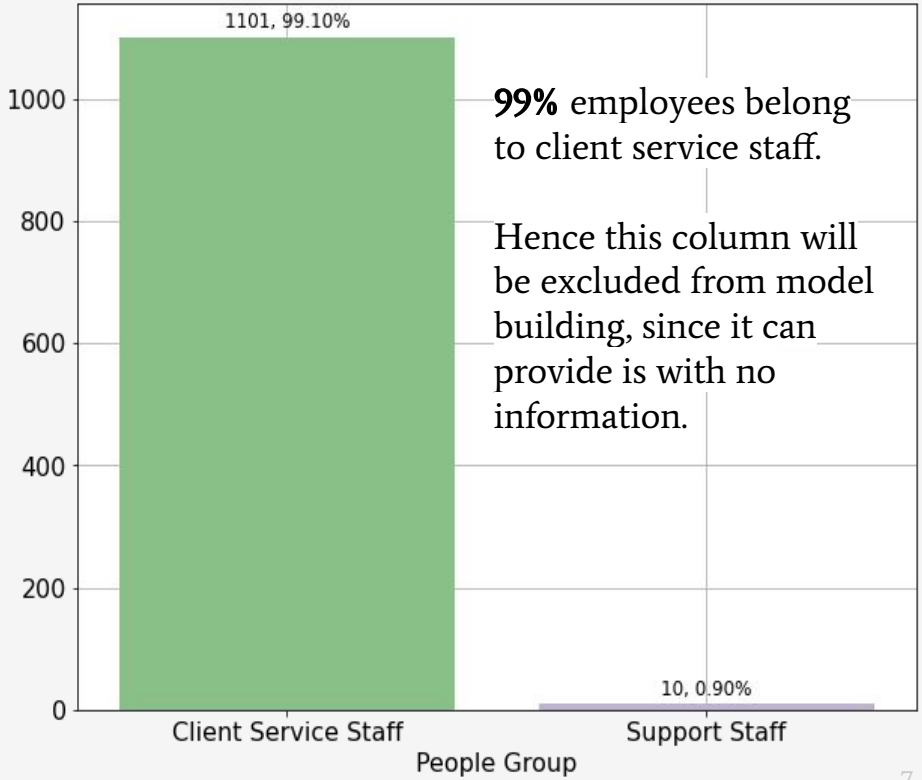
- Extracted year from date of termination to calculate the age and month to check whether the employees leave more on financial year end (April)
- Calculated “Tenure_till_18_in_years” using join date and fiscal year end date for active employees and termination date for resigned employees
- Created “Is Promoted” column using change in Employee Position:
  - Label encoded the levels based on hierarchy and compared them for both the years
  - If there was an increase in level (Level 1 > Level 2), then the value is “Yes” else “No”
- For overall hours columns, we added the hour values of both the fiscal years
- For overall utilization we have taken average of both the fiscal years

EDA

Distribution of employees (Target variable and People Group)



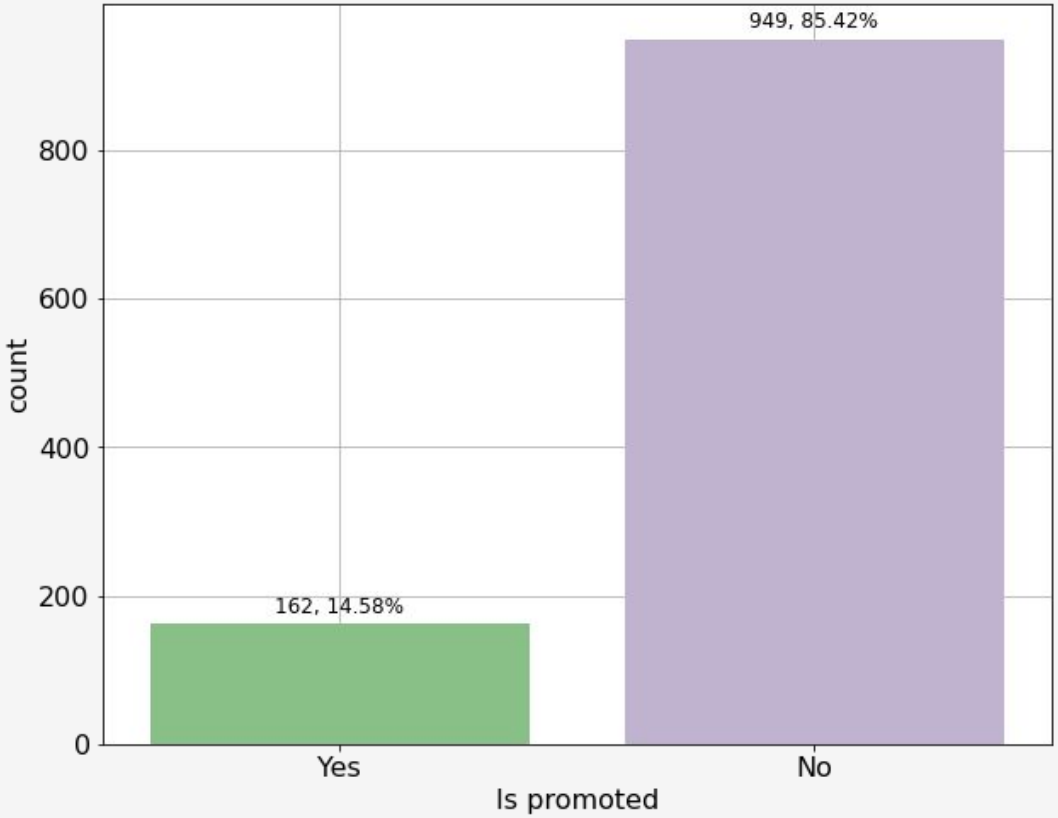
Count plot of **1111** employees, **282** resigned and **829** are active.



99% employees belong to client service staff.

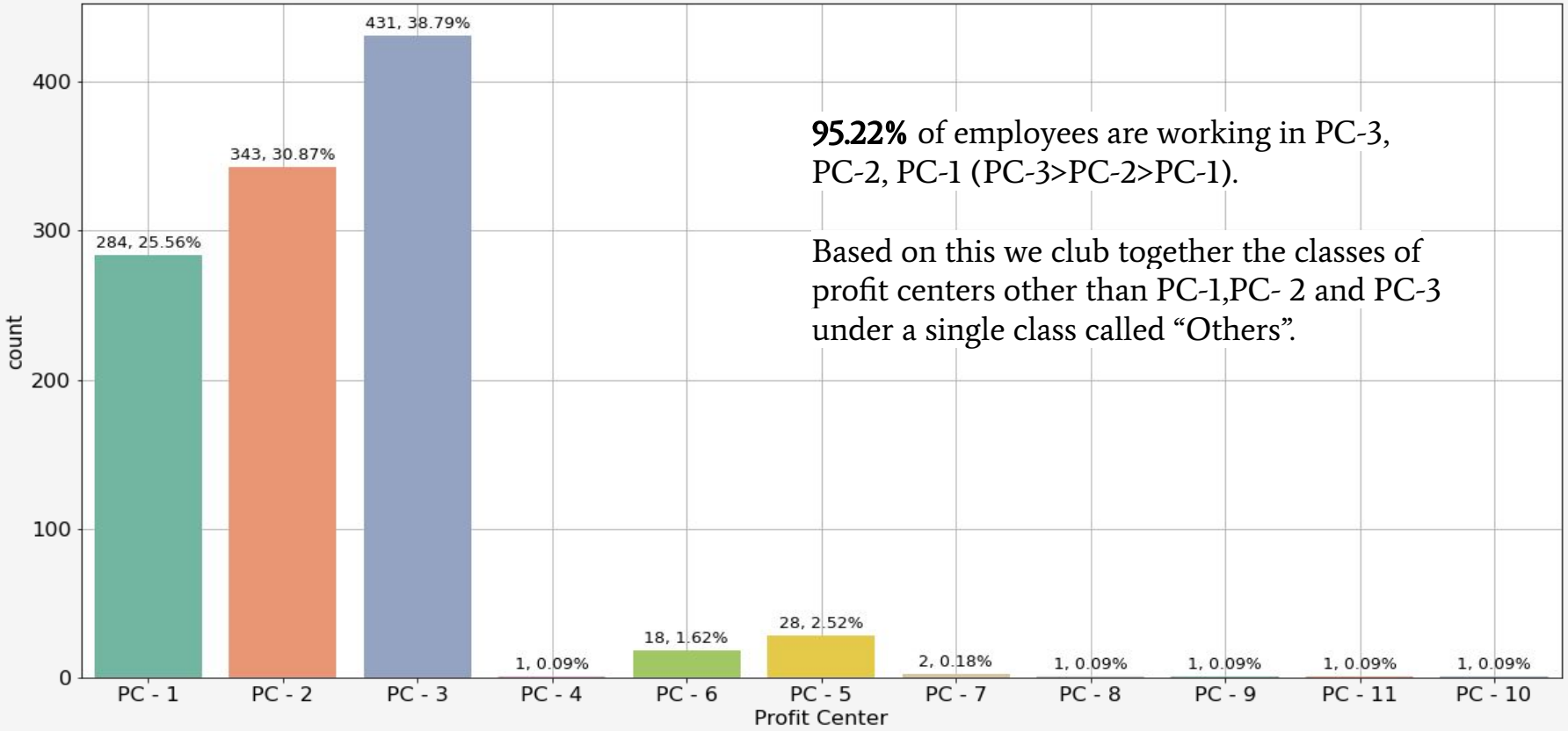
Hence this column will be excluded from model building, since it can provide is with no information.

Distribution of employees based on their promotion



14.58% of employees were promoted from 2016 to 2018

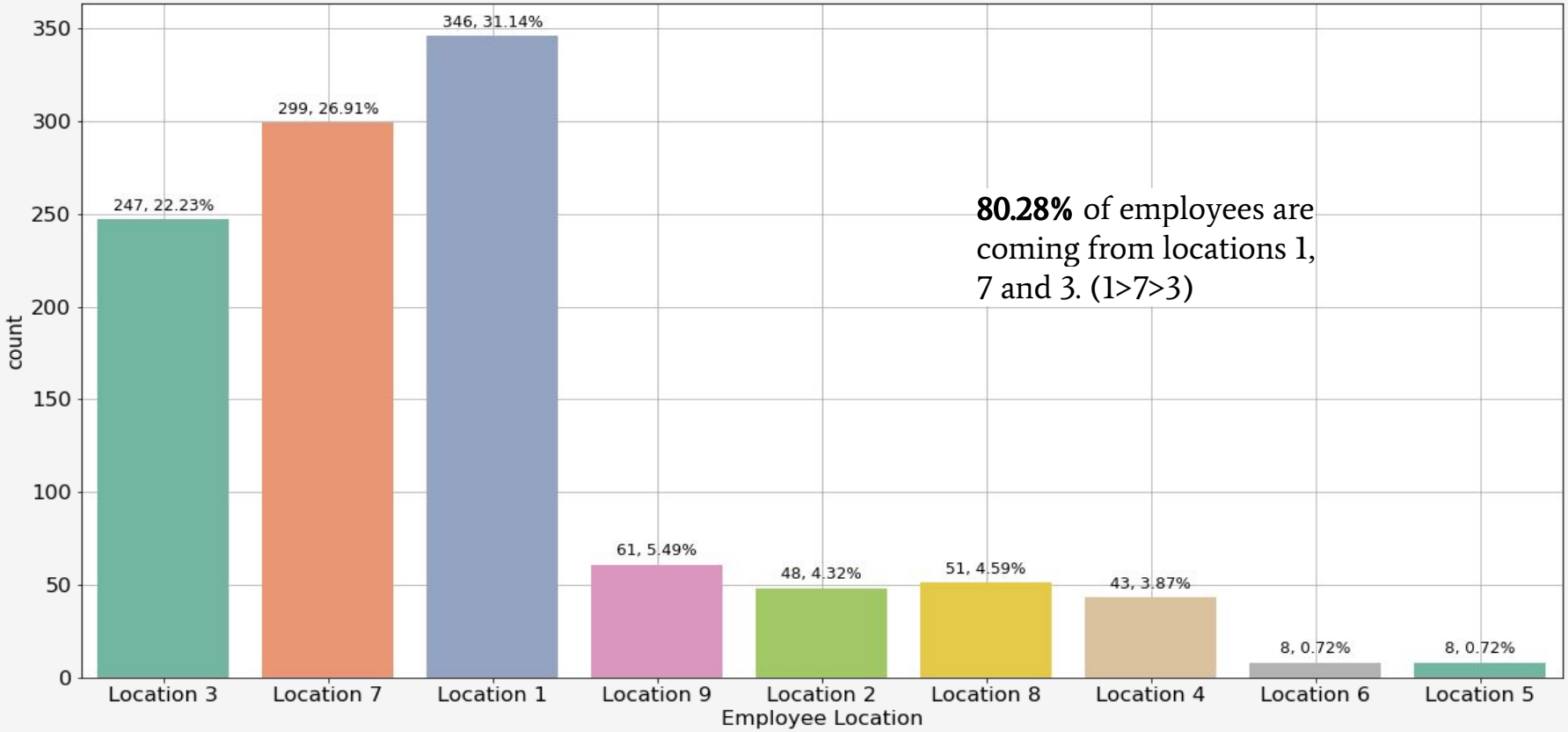
Distribution of employees in profit centers



95.22% of employees are working in PC-3, PC-2, PC-1 (PC-3>PC-2>PC-1).

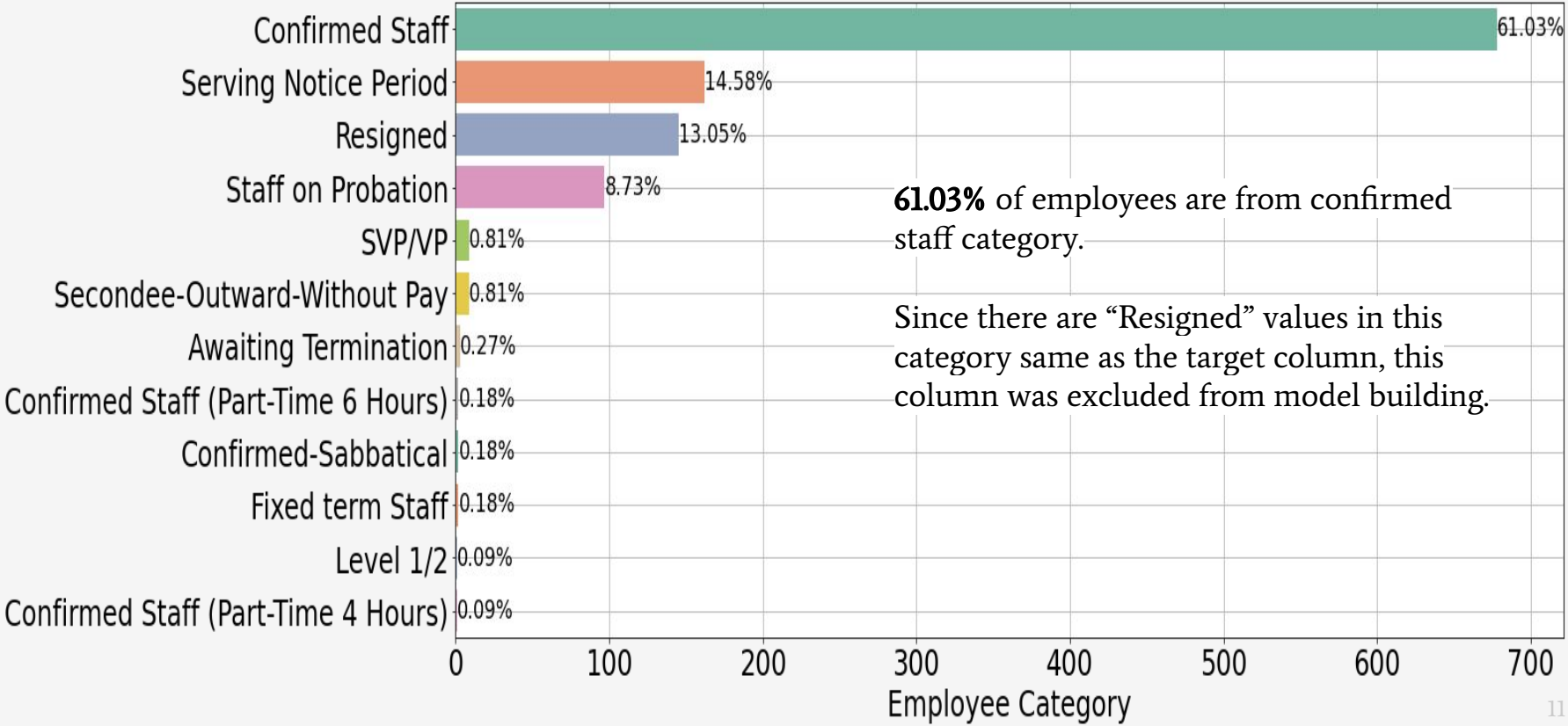
Based on this we club together the classes of profit centers other than PC-1,PC- 2 and PC-3 under a single class called “Others”.

Distribution of employees in all locations



80.28% of employees are coming from locations 1, 7 and 3. (1>7>3)

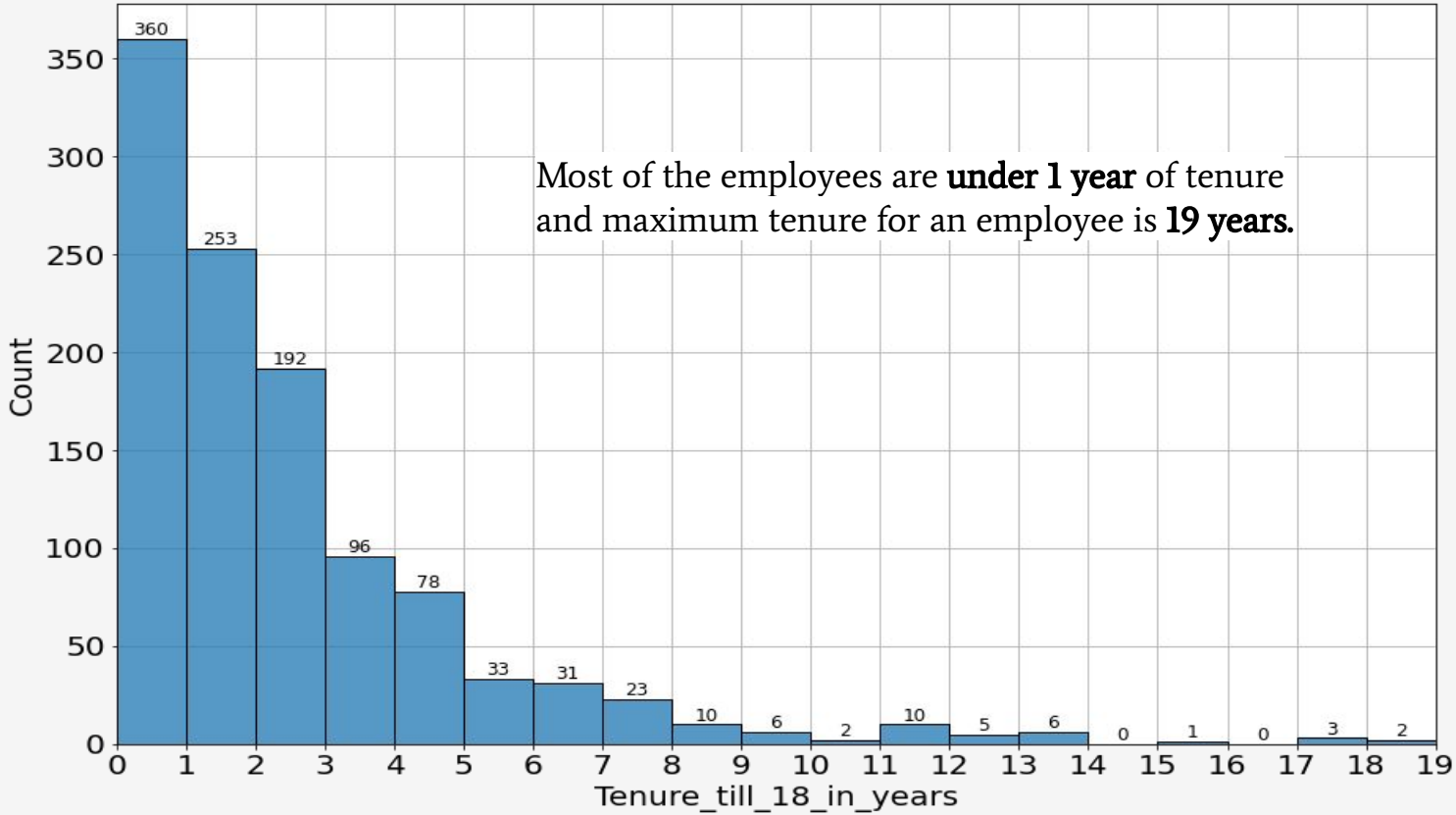
Distribution of employees based on their category



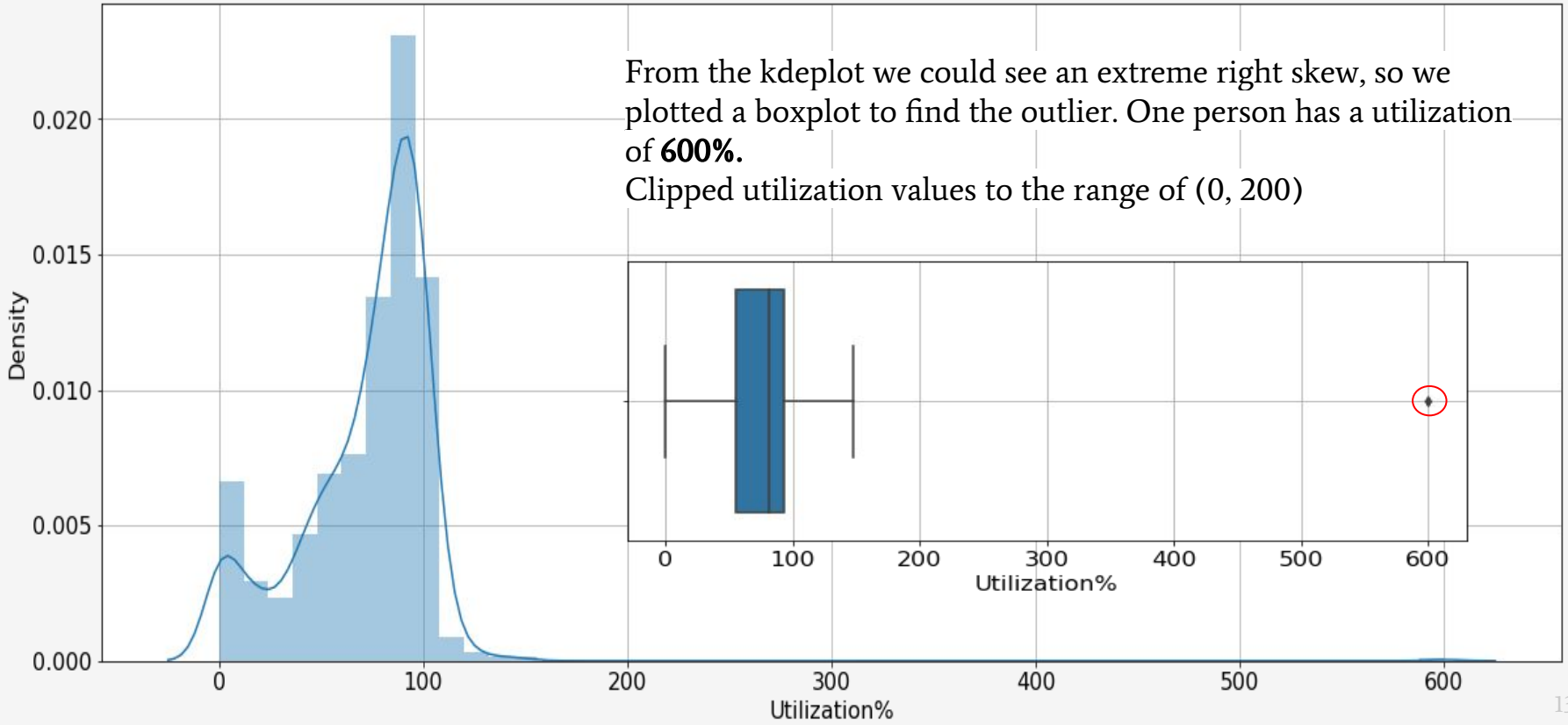
61.03% of employees are from confirmed staff category.

Since there are “Resigned” values in this category same as the target column, this column was excluded from model building.

Distribution of employees based on their tenure (in months)



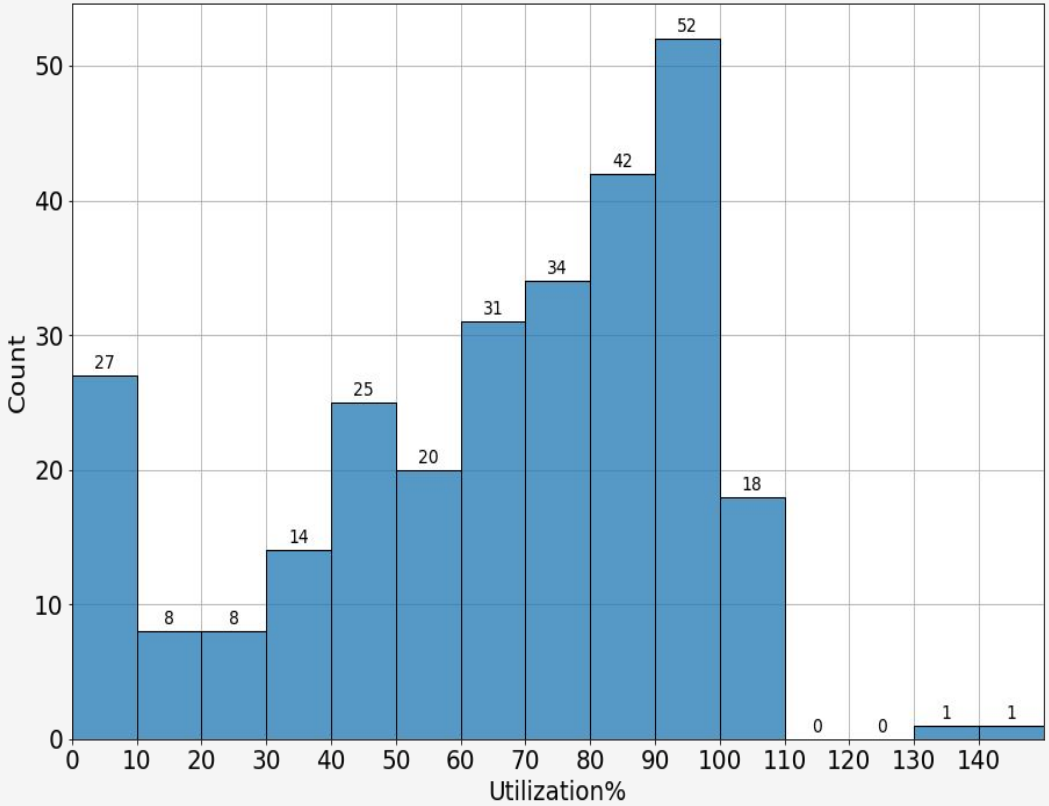
Distribution of employees based on their tenure (in months)



From the kdeplot we could see an extreme right skew, so we plotted a boxplot to find the outlier. One person has a utilization of **600%.**

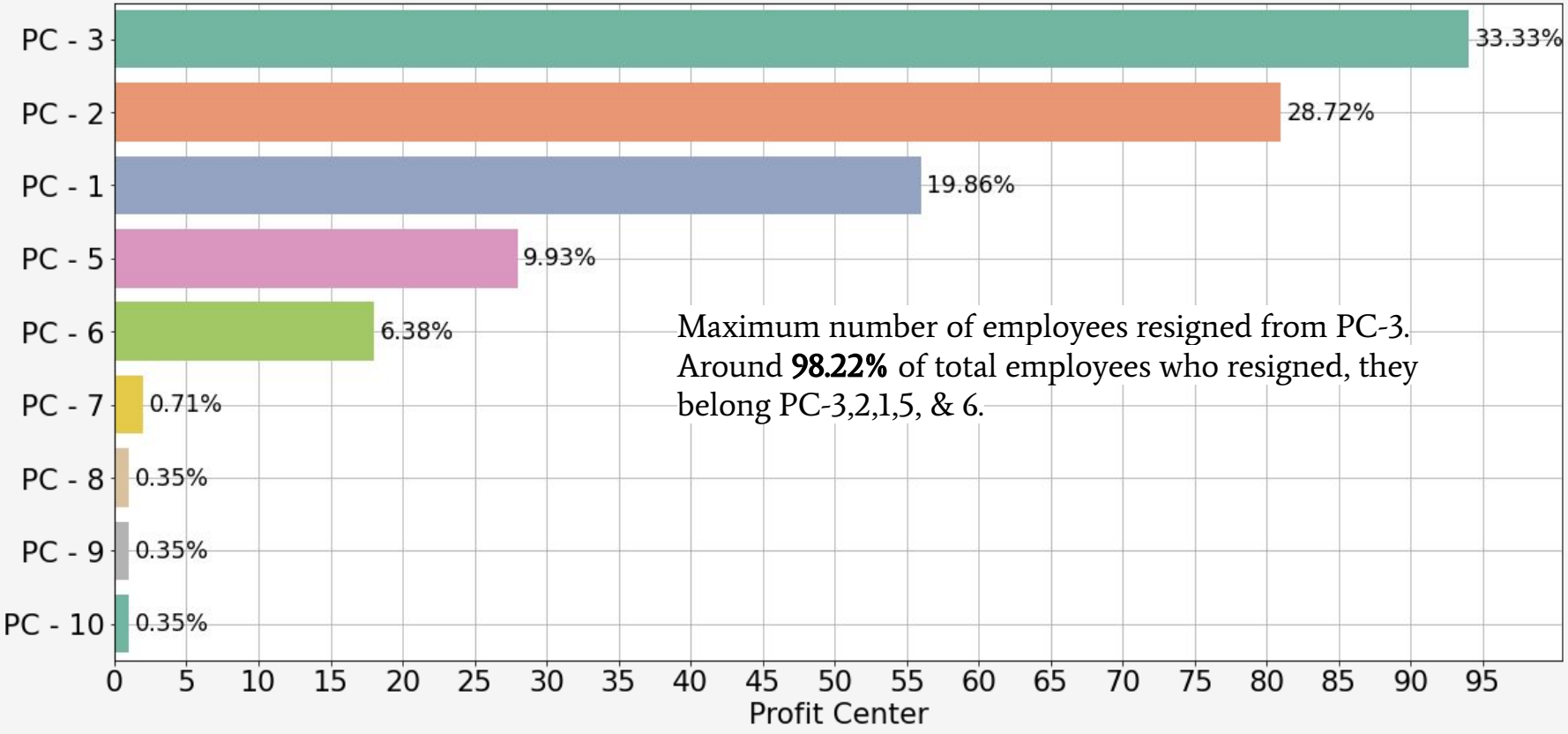
Clipped utilization values to the range of (0, 200)

Employees resigned vs Utilization(%)



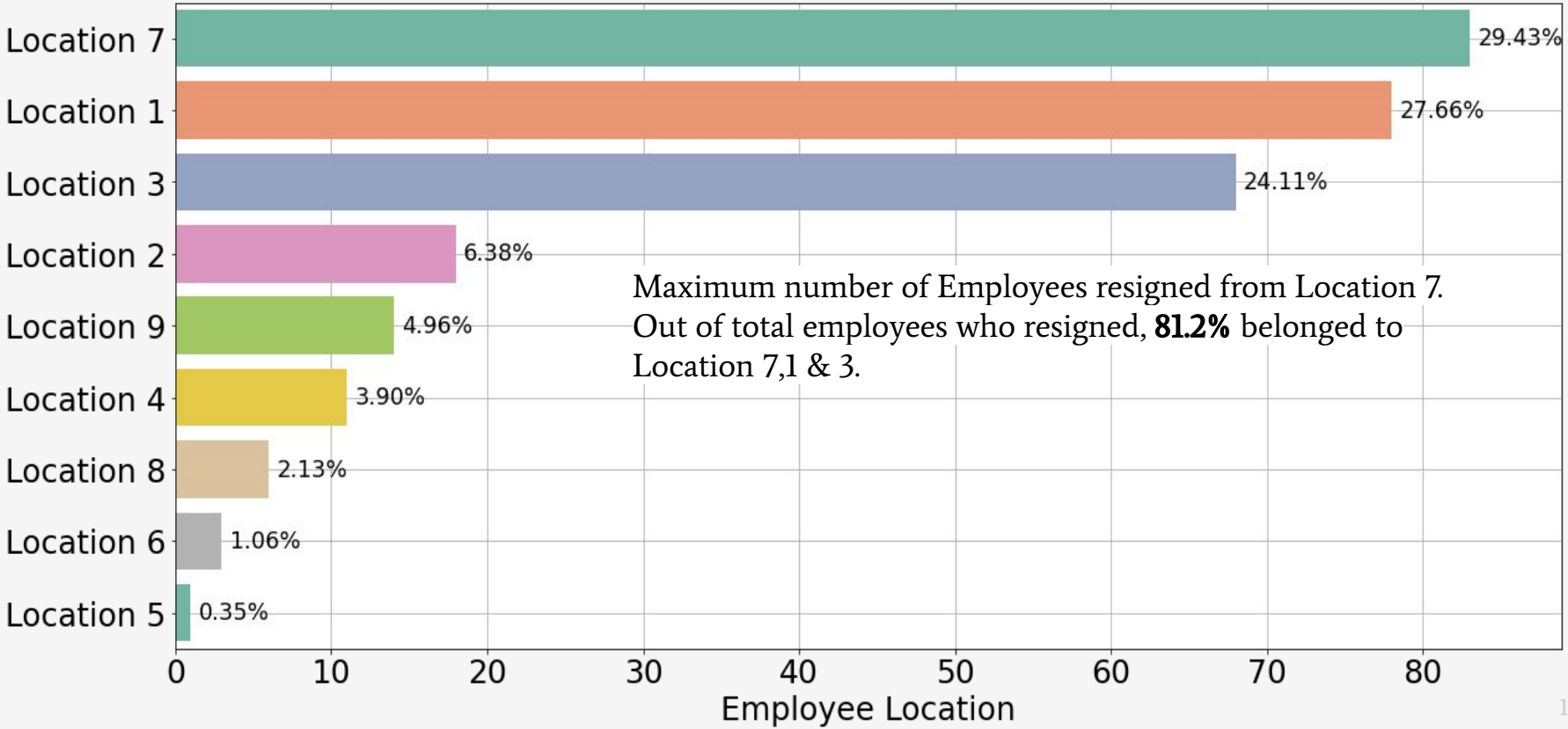
Most of the employees who resigned had utilization between **70%** and **100%.**

Employees resigned vs Profit Center



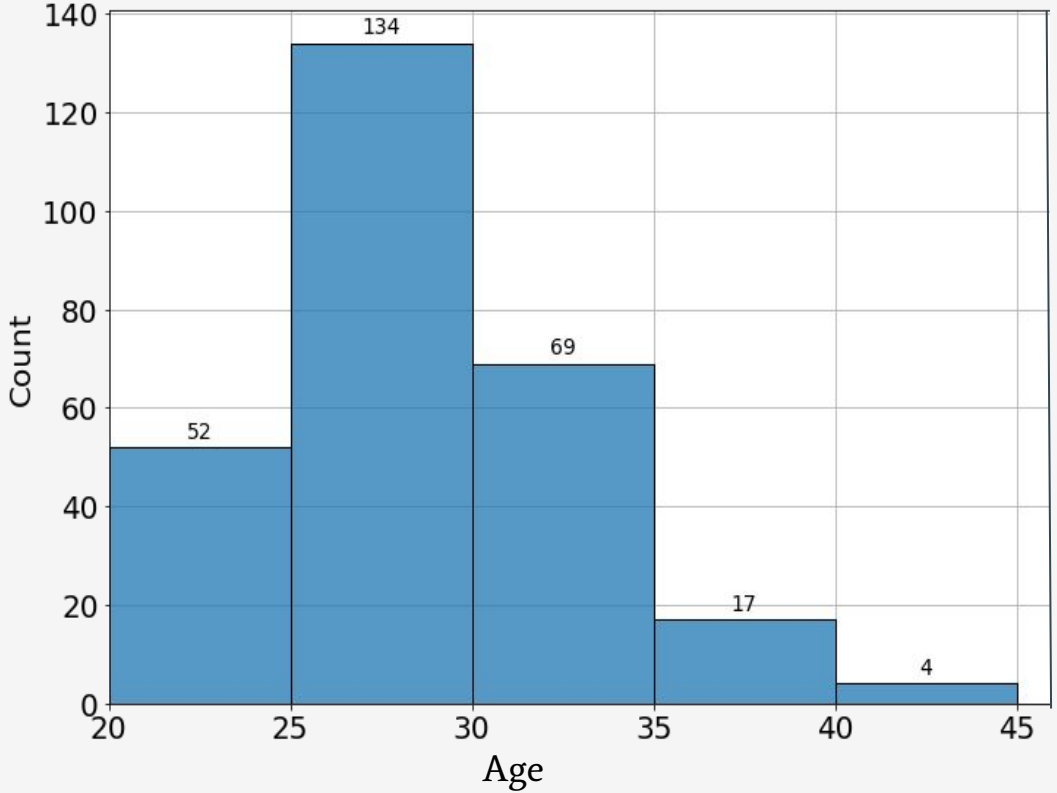
Maximum number of employees resigned from PC-3. Around **98.22%** of total employees who resigned, they belong PC-3,2,1,5, & 6.

Employees resigned vs Location



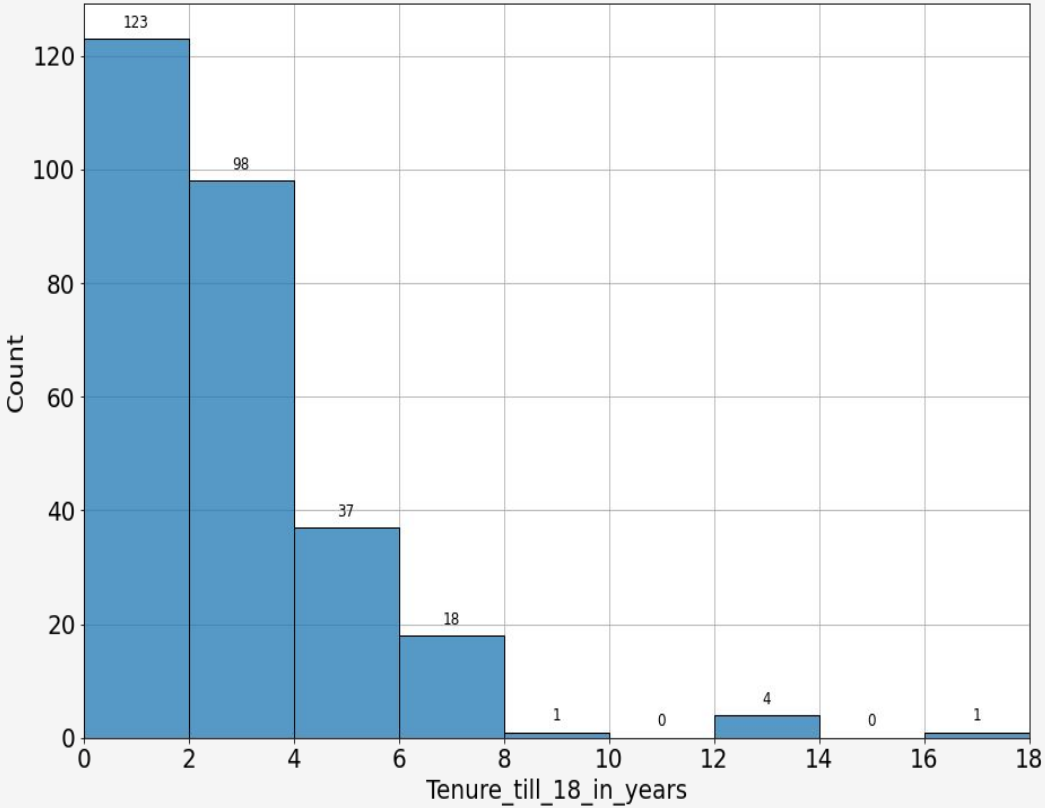
Maximum number of Employees resigned from Location 7.
Out of total employees who resigned, **81.2%** belonged to Location 7,1 & 3.

Employees resigned vs Age (in years)



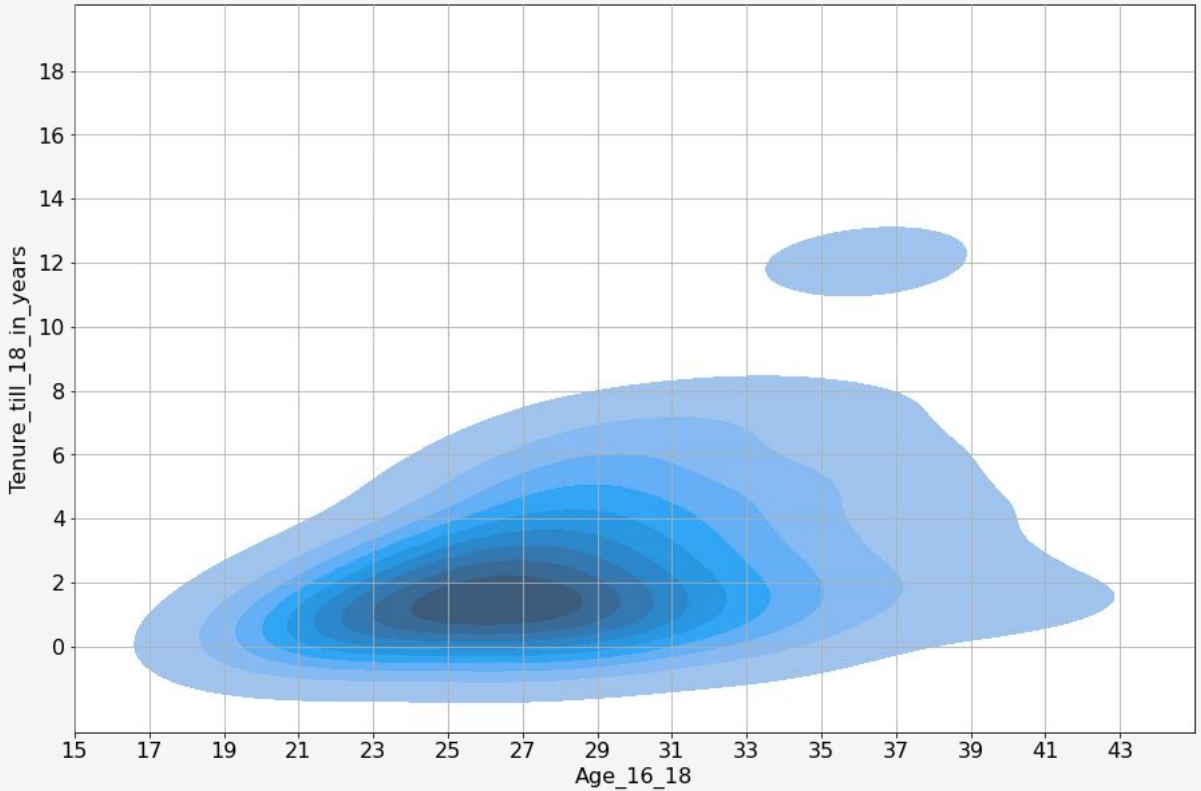
Most of the employees who resigned belong to the **age group (25 - 30)** . So, most of the people who are resigning belong younger generation.

Employees resigned vs Tenure (in years)



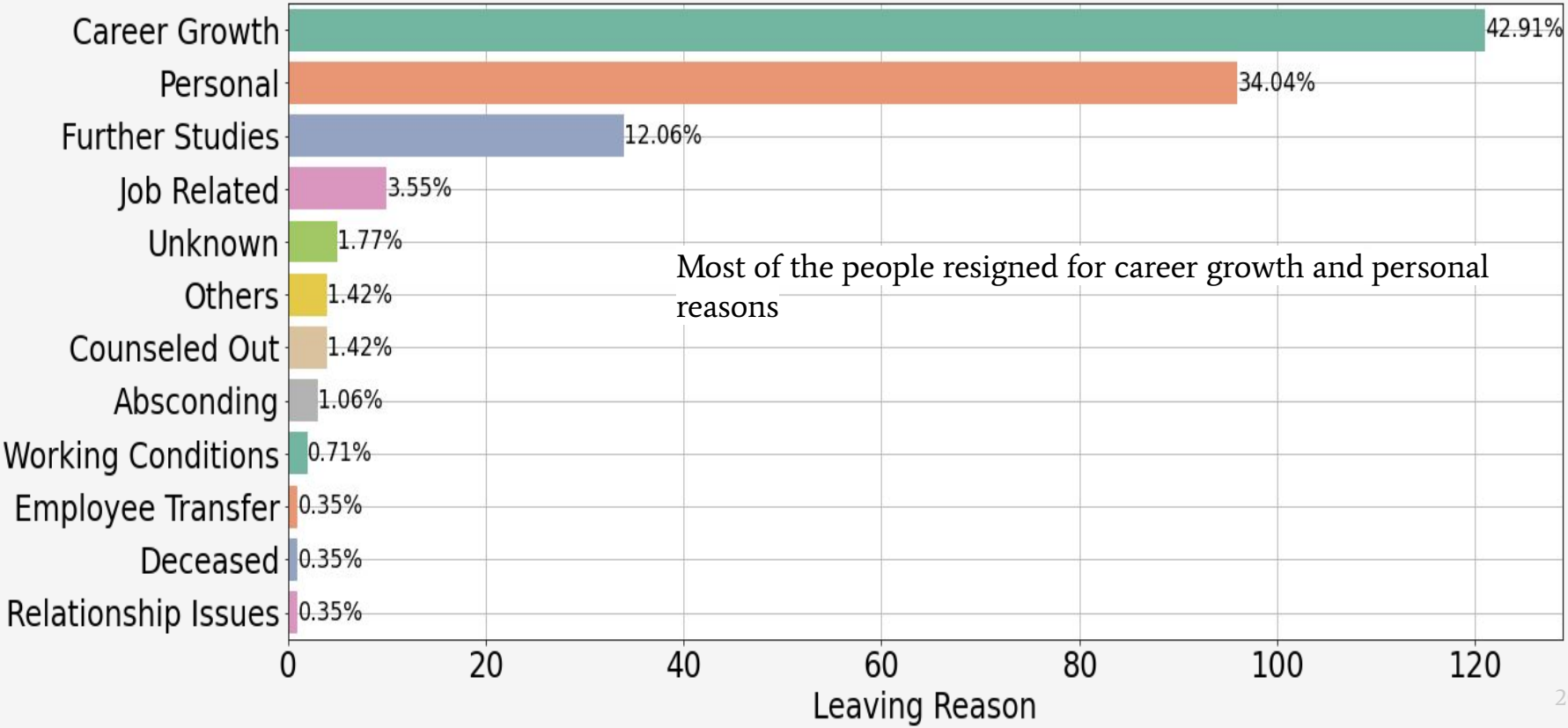
Most of the employees who resigned belong to the tenure of **0-2** & **2-4** years and beyond that the rate of resignation significantly decreases.

Tenure (in years) vs Age (in years) for resigned employees



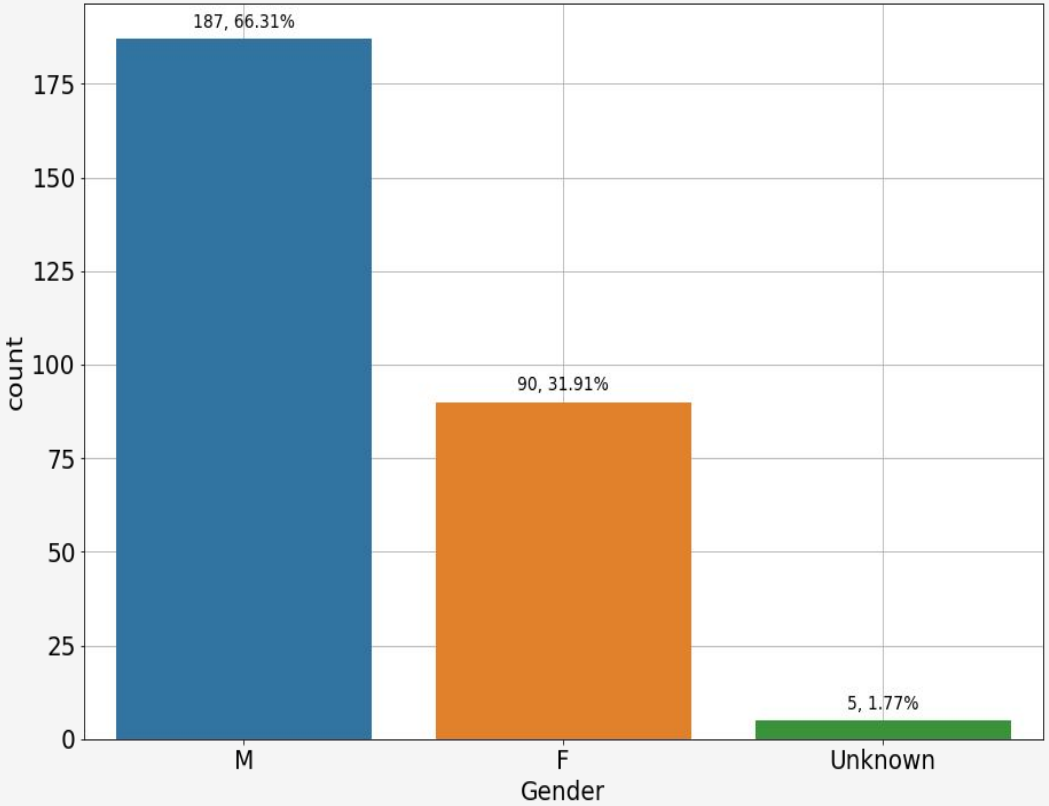
Most of the employees who resigned belong to the tenure of **0-2.5** years with age range of **23 - 30** years

EDA
Based on Termination data: Leaving Reason



Most of the people resigned for career growth and personal reasons

Employees resigned vs Gender



Out of all the employees resigned, **66.31%** employees are male which is more than double of the female, that are **31.91%**..

Feature Selection

We have selected all the columns here that have p-value (significance level) less than 0.05 (5%). The rest were excluded.

	ColumnName	P-value	IsSignificant
0	Is promoted	2.114717e-15	True
1	Employee Category	1.464466e-197	True
2	Employee Location	1.053689e-01	False
3	People Group	4.485451e-01	False
4	Profit Center	5.498416e-29	True

	ColumnName	P-value	IsSignificant
0	Total Hours	0.00000	True
1	Total Available Hours	0.00000	True
2	Tenure_till_18_in_months	0.44056	False
3	Tenure_till_18	0.44056	False
4	Work Hours	0.00000	True
5	Leave Hours	0.00000	True
6	Training Hours	0.00000	True
7	NC Hours	0.00000	True
8	Utilization%	0.01718	True

Final dataset

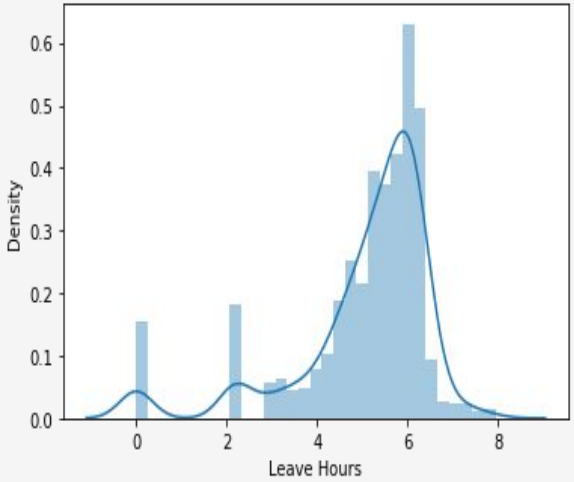
	Profit Center	Current Status16_18	Is promoted	Total Hours	Total Available Hours	Work Hours	Leave Hours	Training Hours	NC Hours	Utilization%
0	PC - 1	Active	Yes	4168.0	3666.5	1084.0	404.0	97.5	74.0	29.661004
1	PC - 2	Active	No	4168.0	3750.0	1277.5	364.0	54.0	1330.5	34.083718
2	PC - 2	Active	No	4168.0	3618.5	2177.5	482.0	67.5	907.0	60.553933
3	PC - 3	Resigned	No	3928.0	3491.5	546.0	404.0	32.5	279.0	15.695026
4	PC - 4	Active	No	1312.0	1205.0	149.0	104.0	3.0	818.0	12.365145

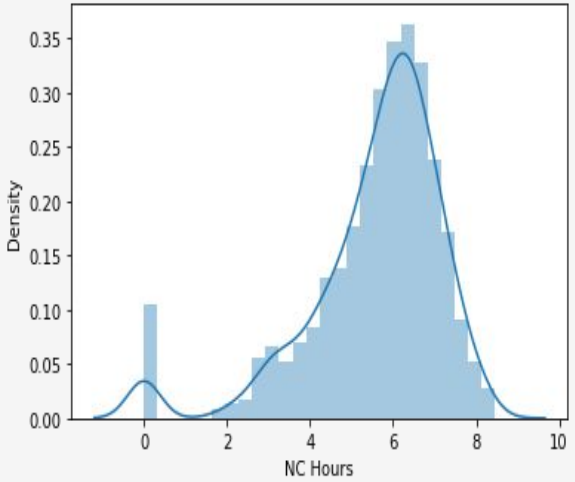
Categorical Columns: “Is promoted”, “Profit Center”
Continuous Columns: "Total Hours", “Total Available Hours”, “Tenure_till_18_in_months”, “Tenure_till_18”, “Work Hours”, “Leave Hours”, “Training Hours”, “NC Hours”, “Utilization%”

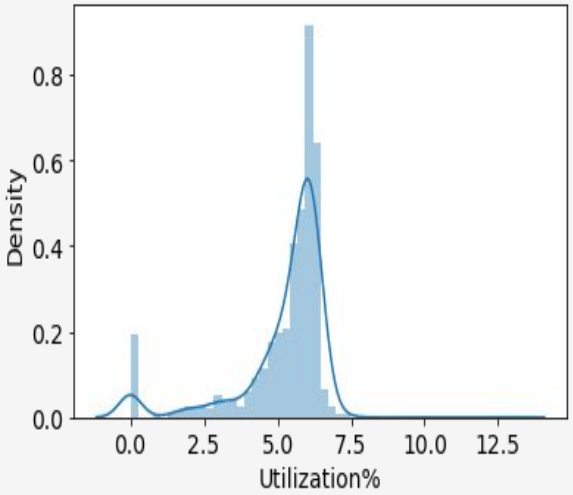
Target Column: “Current Status16_18”

Normality Check

- “Leave Hours” and “NC Hours” had a skew which was dealt with using log transformation
- Transformed Utilization column as it was right-skewed







Model Building

- The final data set was divided into train and test set using 80/20 ratio
- Grid search (cv=5) was done for two models: Decision Tree and Random Forest
- In order to improve the results, SMOTE oversampling was tried before training the models
- Finally, since there wasn't a significant improvement upon using SMOTE a model trained on the original dataset was chosen

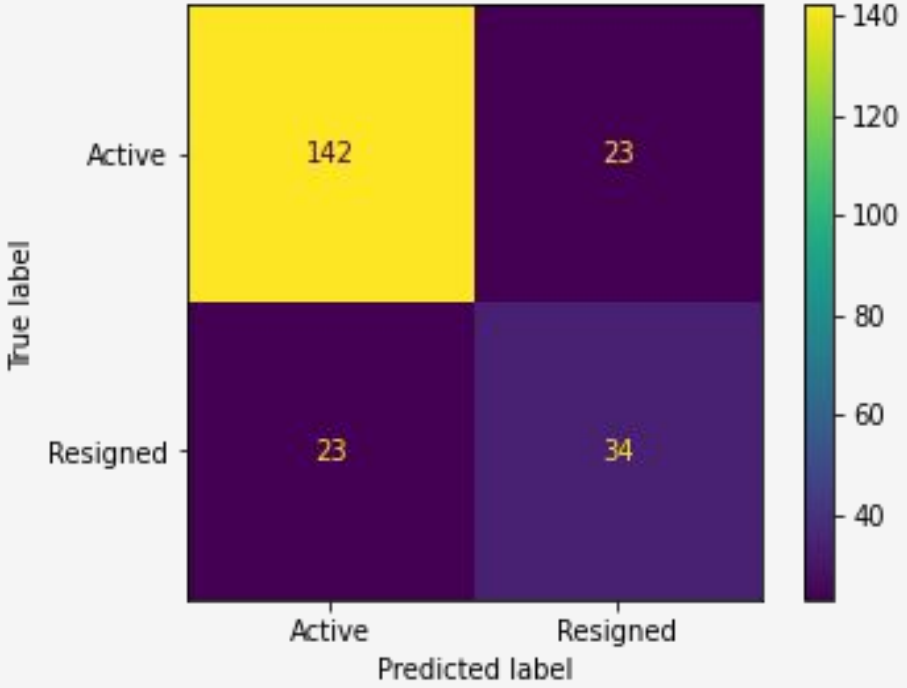
Decision Tree - Results on Unseen Data

ROC : 0.7285

Recall : 0.5965

Precision : 0.5965

Accuracy : 0.7928



Random Forest - Results on Unseen Data

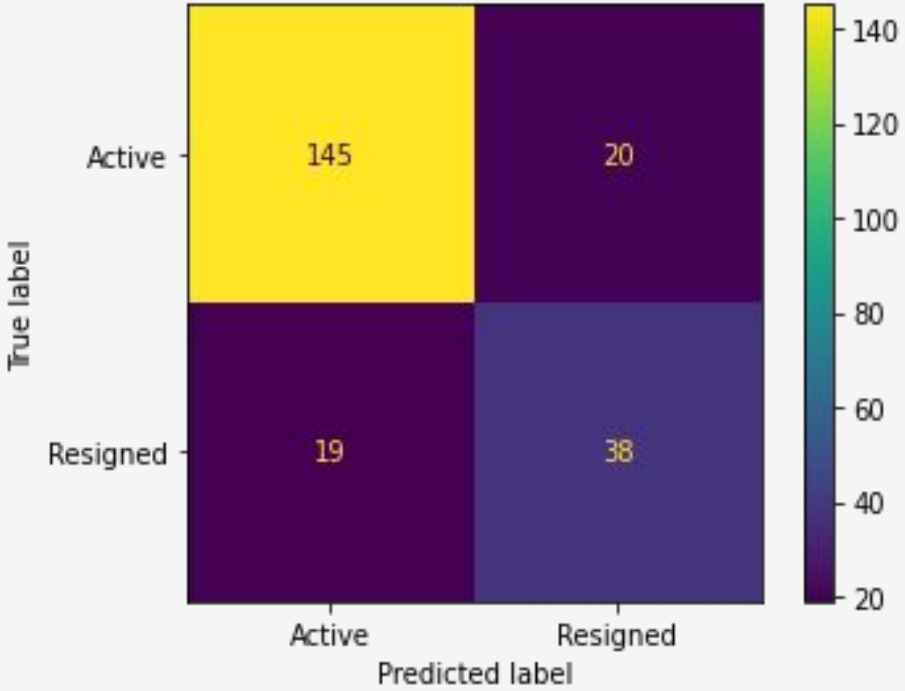
ROC : 0.7727

Recall : 0.6667

Precision : 0.6552

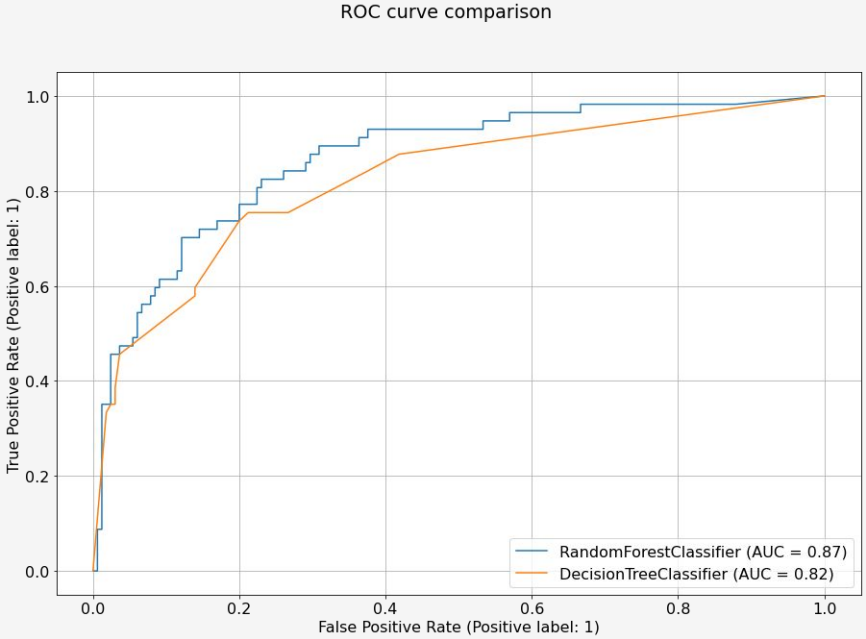
Accuracy : 0.8243

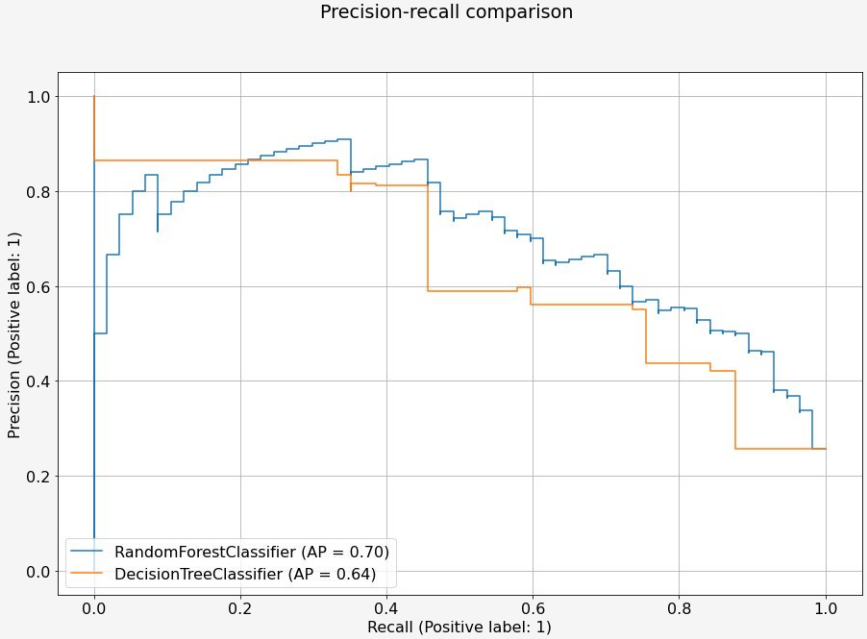
Using threshold as 0.6243



Model Selection

Based on these graphs (test data performance), we have selected the Random Forest model.





Final predictions on 2nd fiscal year data

1. No promoted employee is predicted to leave at the end of the 2nd fiscal year
2. Names of a few employees predicted to resign are shown to the right

	Employee No	Prediction	Is promoted	Employee Name
14	19	1	No	Cordey Sofia
38	45	1	No	Jenny Kasey
43	50	1	No	Leta Evangeline
54	66	1	No	Leela Jsandye
62	75	1	No	Norene Ethelyn

Conclusion

Identify factors influencing attrition	1. Training Hours is negatively related to attrition 2. Promotion prevents attrition
Predict possible attritions	Shown in last slide
Identify possible ways to retain high performers	As no promoted employee has left, we can say that promotion is a potential way of retention. But since promotion cannot always be an option, creating rewards for best performers in the form of perks, bonuses and awards can be introduced. Since most of the employees are leaving for career growth, providing better training (in duration and in quality) opportunities will allow them to upskill